International Journal on Advances in Security

2009 vol. 2 nr. 2&3

The *International Journal on Advances in Security* is Published by IARIA.
ISSN: 1942-2636
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Security
Volume 2, Numbers 2&3, 2009

- Aljosa Pasic, ATOS Origin, Spain
- Vladimir Stantchev, Berlin Institute of Technology, Germany
- Michiaki Tatsubori, IBM Research - Tokyo Research Laboratory, Japan
- Ian Troxel, SEAKR Engineering, Inc., USA
- Hans P. Zima, Jet Propulsion Laboratory/California Institute of Technology - Pasadena, USA // University of Vienna, Austria

Security in Internet

- Evangelos Kranakis, Carleton University, Canada
- Clement Leung, Victoria University - Melbourne, Australia
- Sjouke Mauw, University of Luxembourg, Luxembourg
- Yong Man Ro, Information and Communication University - Daejon, South Korea

International Journal on Advances in Security
Volume 2, Numbers 2&3, 2009

CONTENTS

Post-Payment Copyright System versus Online Music Shop: Business Model and Privacy

Heikki Kokkinen
Nokia Research Center
Helsinki, Finland
heikki.kokkinen@nokia.com

Mikko V. J. Heikkinen
TKK Helsinki University of Technology
Helsinki, Finland
mikko.heikkinen@tkk.fi

Markus Miettinen
Nokia Research Center
Helsinki, Finland
markus.miettinen@nokia.com

Abstract—**A post-payment copyright system is used to legalize copyrighted music files which a user has obtained illegally. We compare a post-payment copyright system to an online music shop by analyzing three scenarios using both qualitative business and quantitative techno-economic modeling. We analyze the privacy challenges and solutions related to the post-payment copyright system. According to our quantitative analysis, the post-payment copyright system is potentially a more profitable business than an online shop when no media replacement is required. Our qualitative analysis suggests benefits in bundling post-payment copyright system with online music shop and customer sensitivity to the marketing message. The privacy threat analysis and the list of suggested solutions show that privacy is a key factor in the system success, but it is possible to develop adequate protection for the user privacy. Our research is a continuation to the trend of studies suggesting peer-to-peer networks as a part of a viable business model for media distribution.**

Keywords-business model; copyright; post-payment system; privacy; risk analysis; security; techno-economic modeling

I. INTRODUCTION

This paper analyzes a post-payment copyright system with three methods: (1) qualitative business modeling; (2) quantitative techno-economic modeling; and (3) the attack tree method for security analysis of user data privacy. It extends and refines our previous work on quantitative techno-economic modeling of such a system [1] with extensive qualitative analysis on business model and privacy issues. In a post-payment system the users are able to legalize the unauthorized music files on their hard disks and memory cards. In order to understand the service, let us consider a user of peer-to-peer (P2P) networks. The user has downloaded music files on his computer from a peer-to-peer network. With the post-payment copyright system, the user can pay the required fees to copyright agencies and avoid potential litigation, resulting in both personal and commercial security. In this paper we study the cost efficiency of such a system in comparison to two related systems: a conventional online music shop and a post-payment copyright system where the illegal file is replaced with a legal file.

We study the following research questions: what are the main differences in the business models of post-payment copyright systems and conventional online music shops; what are the differences in profit, risk and cost distribution between post-payment copyright systems and conventional online music shops; and what is the role of privacy in the post-payment copyright system.

For our qualitative analysis we use the STOF (Service-Technology-Organization-Finance) business model analysis framework established by Bouwman et al. [2]-[5]. In our quantitative analysis, we conduct techno-economic analysis for digital music sales complemented with risk analysis using Monte Carlo simulations.

The term business model has been defined in several ways in the academic literature. Timmers [6] concentrated on technology elements, whereas Amit and Zott [7] emphasized revenue generation aspects, and Chesbrough and Rosenbloom [8] design aspects. Based on the previous research, Bouwman et al. [3] proposed a unified definition, which acts as a basis for the STOF business model analysis framework.

The other bases for the STOF framework are several componentizations of business models. Alt and Zimmerman [9] recognized mission, structure, process, revenues, legal issues, and technology as the main elements of business models. Osterwalder et al. [10] proposed product, customer interface, infrastructure management, and financial aspects as the basic elements of business models. Shafer et al. [11] identified strategic choices, value creation, value capturing, and value network as the main components in 12 componentization publications. Bouwman et al. [3] decided to focus on four components in their STOF framework: service, technology, organization, and finance.

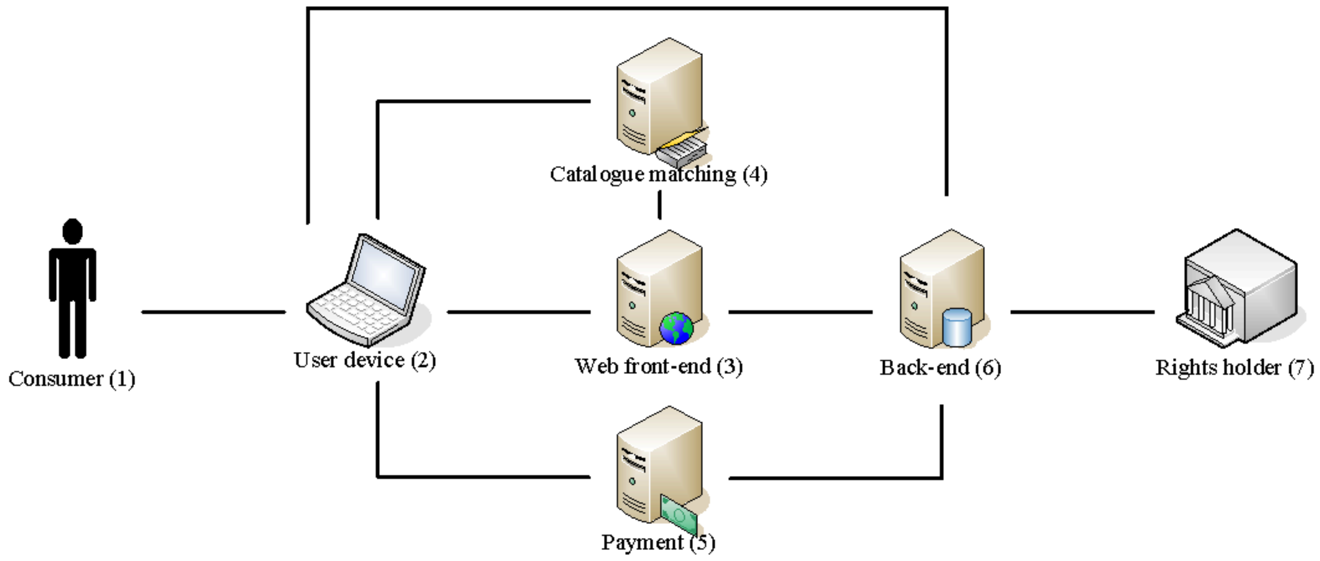


Figure 1. Post-payment copyright system architecture

The post-payment copyright system was introduced by Kokkinen et al. [12]. The legal framework for the service was discussed in [13], and the method for illegal vs. legal classification of MP3 files in [14]. In an online survey Hietanen et al. [15] identified post-payment copyright system as the second most attractive new P2P related business model for consumers, following a monthly paid file sharing service with unlimited access to music and video.

Consumers, copyright authorities and legislators have varying views about P2P. Cohn and Vaccaro [16] apply neutralization theory to the ethics of P2P music file sharing. The P2P file sharing divides opinions about its impact on music business. Peitz and Waelbroeck [17] show that P2P music file sharing has a positive impact on music sales through wider sampling of music by consumers. Bhattacharjee et al. [18] simulate online music sales with different strategies in the presence of P2P-based piracy.

Techno-economic modeling can be considered as a quantitative extension to qualitative business modeling. It analyzes profitability of a new technology or service. Techno-economic modeling and its application to broadband access networks is introduced in [19] where a quantitative framework for conducting techno-economic analysis on broadband networks and several case studies based on it are depicted. Elnegaard and Stordahl [20] demonstrate the use of Monte Carlo simulations as a method for risk analysis in quantitative techno-economic models. Both qualitative and quantitative techno-economic modeling have been used in several studies related to telecommunications: Monath et al. [21] on fixed broadband access network strategies, Jerman-Blažič [22] on network backbone upgrade investments, Kumar and Kueh [23] on international mobile roaming, Smura et al. [24] on virtual operators, Kivisaari et al. [25] on mobile broadcast, Rokkas et al. [26] on fixed-mobile convergence, and Heikkinen and Luukkainen [27] on mobile P2P communications.

Security is essential in deploying the post-payment copyright system. Schäfer et al. [28] studied security in P2P networks from a general perspective, Suomalainen et al. [29] from a mobile perspective, and Merz et al. [30] from a grid perspective. Koshutanski et al. [31] analyzed security in digital ecosystems. Without underestimating the concerns in the web generally, the post-payment copyright service raises even more serious concerns than an average web or e-commerce site, because the post-payment copyright service requests user to give information about past illegal activities. Such information is very sensitive to the user, and in normal circumstances past illegal activities are not disclosed to anyone. Storing so sensitive information also requires special measures.

Privacy in the general context is understood as a “right to be left alone”. In the digital information world, privacy can be interpreted as a right of the user to control what personal information is disclosed to whom, and under which circumstances [32]. In the post-payment copyright service, privacy means among other things that the user of the service can control the information that is submitted to the service and that the exposure can be limited to only such data that will not have negative consequences for the user.

Already in the early days of online commerce, consumer privacy concerns were prevailing. However, during the last decade, consumers have learnt to trade private information for personalized services. Wang et al. [33] classify privacy concerns in Internet marketing as follows: data is acquired improperly including access, collection, and monitoring; data is used improperly including analysis and transfer; privacy is invaded as unwanted solicitation; and data is stored improperly. Kobsa [34] has found the following privacy principles in the European legislation: personalized services based on traffic or location data require the anonymization of such data or user's consent; users must be able to withdraw their consent to the processing of traffic and location data at any time; the personalized service provider must inform the user of the type of data that will be processed, of the purposes and duration of the processing, and whether the data will be transmitted to a third party prior to obtaining the user's consent; personal data obtained for different purposes may not be grouped; and usage data must be erased immediately after each session. Lu et al. [35] classified the elements of privacy in a peer-to-peer system to be identity of peers, content, and interests.

The major privacy concerns of the user arise due to the fact that in the course of using the post-payment copyright service, the users are submitting indirect evidence about their past illegal activities. Users want to be absolutely sure that there is no considerable risk of their data being used in any other way than what it is necessary for fulfilling the purpose of the service, i.e., legalizing the user's content.

The nature of the service itself can feel for the user like being accused as a criminal. Such feelings raise easily negative reactions. An example of this was the blog discussion triggered by a questionnaire study investigating internet users' perceptions on peer-to-peer file sharing. The study was published in [15]. The blog writers used very strong language when they expressed their outrage. They felt that the survey questions blamed all peer-to-peer file sharing content to be illegal and that the creators of the questionnaire allegedly accused the respondents as criminals.

Many web and e-commerce sites use methods which decrease privacy concerns and build trust in users. Kobsa [34] has found the following aspects to decrease the privacy concerns on web sites: positive past experiences, design and operation of the site, reputation of the site operator, presence of a privacy statement, presence of a privacy seal, privacy laws, pseudonymous users and user models, client-side personalization, and privacy enhancing techniques for collaborative filtering. Hoffman et al. [36] discussed how to build trust online by anonymity or pseudonymity, cooperative interaction between site owner and consumers, and privacy policies. Palmer et al. [37] showed how trusted third parties and privacy statements increase the trust on an e-commerce site. Freenet [38] protects the peer-to-peer users from privacy infringements and uses anonymity to decrease privacy concerns.

The post-payment copyright service provider should use the known methods of creating trust on web sites and e-commerce sites. In Table 1, we list the web site privacy enhancing techniques, and how they could be applied in the post-payment system. In addition to the generic challenges of a typical Internet site or peer-to-peer network, the post-payment system has its own privacy and trust challenges. The web and online shop-related privacy challenges in the post-payment copyright system can be solved by utilizing the methods suggested above.

TABLE I. ONLINE SHOP AND WEB PRIVACY SOLUTIONS AND THEIR APPLICATION IN POST-PAYMENT SYSTEM

Web site privacy solutions	**Application in post-payment copyright system**
Positive past experiences	Objective to meet the customer expectations
Design and operation of the site	Simple and reliable workflow on the site
Reputation of the site operator	Established site operator brands visible on the site
Presence of privacy statement	Use of privacy statement
Presence of a privacy seal	Not applied
Privacy laws	Finnish and EU privacy legislation
Pseudonymous users and user models	Possibility to use the service without giving personal information before actually paying for the content.
Client-side personalization	Not used in the beginning
Privacy-enhancing techniques for collaborative filtering	Web analytics tools, which have privacy enhancing techniques

In this paper, we study the digital music business from the perspective of copyright owners. According to our analysis, the post-payment copyright system is potentially a more profitable business than the online shop when no media replacement is required.

Our paper is structured in a following way: Section 2 portrays the post-payment copyright system; Section 3 describes our qualitative business modeling, quantitative techno-economic modeling, and privacy threat analysis methods; Section 4 presents our scenarios; Section 5 contains our results; and Section 6 discusses our findings.

II. POST-PAYMENT COPYRIGHT SYSTEM

The post-payment copyright system allows the user to pay the copyright fees of illegally copied files. The system was introduced by Kokkinen et al. [12], and it is described here in more detail.

In the system, the payment process and distributing the copyright fees to the rights holders are similar to the respective functions of an online music shop. After creating the shopping basket, the user is directed to a payment page. Depending on the payment method, the user interface for typing in the payment details may belong to an online shop or to a financial institute. Typically, credit card information is given through the online shop user interface and bank account information through the financial institute interface. The details of the shopping basket are not visible to the financial institution and the online shop does not know the specifics of the payment arrangement. A transaction identification code ties together the operations in these two systems.

The payments for the rights holders are regulated by the legislation and signed contracts between the online shop operator and the rights holders. In most jurisdictions a value added tax is reported and paid to the tax authorities. Artists, composers, writers, technicians, and other people involved typically get their share through record labels and copyright organizations. The contracts between the online shop and these organizations define the method and amount of payments for the rights holders.

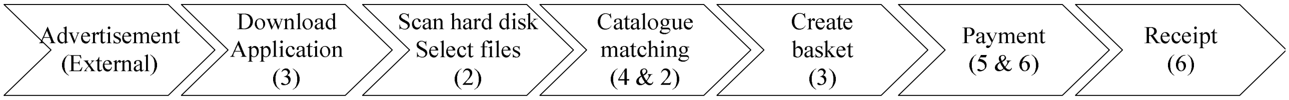


Figure 2. Post-payment copyright process

The architecture of an online music shop is similar to that of a post-payment copyright system; see Fig. 1 for component numbering. Rights holder (7), back-end (6), payment (5) and consumer (1) are identical. The web front-end (3) and user device (2) are present in both architectures, but their functionality differs from each other. The catalogue matching server (4) is unique to the post-payment copyright system.

We use corresponding numbering to describe the components involved in the post-payment copyright process, see Fig. 2. Compared to online music shops, the post-payment copyright specific part takes place prior forming the shopping basket.

Consumers are made aware of the post-payment copyright site through advertising. A consumer enters the post-payment copyright site and downloads an application, which is used to scan user's hard disk. The application helps the user in selecting the files for legalizing. This phase consists of classifying the illegal and legal files, creating a shopping basket, and matching the selected files to the titles in the music catalogue of the system provider. The music catalogue is a list of tracks and albums, which are available in the service.

With the user device and the user application, the consumer can access the storage where the user has content files. The user application scans the device for illegal files and stores the information locally. The user application helps the user to select the relevant files. With the user client it is possible to access the services on the web front-end and on the payment system, although also a web browser interface is needed as a part of the process. The client communicates with the catalogue matching server and allows the consumer to operate the system through the user interface.

The web front-end provides the user with information about the service and the capability to download the required client. The web-front end manages the transfer of the content catalogue from the back-end to the catalogue matching server. The transfer requirement is due to an organizational setup. The web front-end manages the contents of the shopping baskets.

For an overview of the privacy related issues of the system components, see Table 2. For using the web front-end, the user does not need to provide an authentication, i.e., the user can interact with the service in a pseudonymous fashion. The only personal information that the user has to provide to the service is an email address as contact information. The contact information is cryptographically embedded in the purchase receipt, and its purpose is to discourage users from copying and redistributing fake licenses to other users. The user address in the receipt can not be forged without making the forgery detectable and invalidating the receipt.

TABLE II. PRIVACY RISKS OF THE SYSTEM COMPONENTS

System component	Required private information	User identifi-cation	Re-identification risks
Front-end and back-end	User contact address and list of content in shopping basket	Pseudonym	User's contact information is available as long as it has not been deleted
Catalogue matching server	Illegal and legal content in user's possession	Anonymous	User's IP address can be tracked
Payment system	Payment information, possibly including user identification	Payment information	Linking of payment information to actual user identification

The catalogue matching server matches the user file information with the music catalogue items. When there is a mismatch between the user file information and the catalogue, the catalogue server returns one or more closest matching items from the catalogue. Respectively, when there is more than one match in the catalogue, the matching server returns all matches. Catalogue matching server collects only statistical data about the matching requests. Individual catalog matching requests are anonymous and no private information is stored on the catalogue matching server. The requester can only be identified based on the source IP address of the device. However, the catalogue matching server does not track source IP addresses. It only collects statistical information about the incoming requests.

The payment server has a commercial online payment security level. The user can pay with major credit cards, online bank, and a selection of micro-payment systems. The payment system collects all data that are required for the payment and stores them as long as legislation requires. The payment system does not store information about the contents of the shopping basket but only the transaction identification codes. Linking of payments to shopping carts happens via transaction identifiers and only the back-end server has access to the actual shopping cart information. The payment system has the most specific information about the user in the whole system, since the user has to provide payment information in order to fulfill the payment process. The payment information typically consists of credit card information. The payment information may include information that identifies the user.

The back-end system maintains the catalogue by communicating with the rights holders. It sets the prices of the catalogue items based on the rights owner price, value added tax, payment method, and the target margin of the service. The back-end system maintains records of the payments. It may have a username database, and it stores the information about the shopping basket, payment, and the username. The back-end does not contain any information about illegal files. The purchase information and the related personal information are kept as long as the legislation requires. Currently, the retention period in the book-keeping legislation is six years in Finland. However, no direct link to the user identity exists.

Before issuing a receipt, the back-end communicates with the payment service to make sure that the payment has been completed. Each paid basket is assigned a transaction identification code, which is used to link basket information to payments in the payment system. The only personal information required by the service is the email address of the user. It is used to send the payment receipt to the user by email. The receipts issued by the system also include the email address of the user in order to reduce the temptation to copy and resell the receipt to other users, i.e., to gain financial benefit by selling fake authorizations. The user email addresses are deleted in the back-end daily.

In the post-payment copyright system model, the rights holders are customers of the system. The music right holders include record labels, and copyright organizations representing the rights of the artists. Tax authorities belong to the rights holder category from the system point of view. The rights holders receive periodically a share of the user payments. The content owners cannot track personal payments in the system without consulting the post-payment service provider.

III. Methods

A. Qualitative Business Modeling

The STOF framework for business model evaluation by Bouwman et al. [2] consists of four domains: Service, Technology, Organization, and Finance. A business model outline based on the four domains is evaluated based on Critical Design Issues (CDIs) and Critical Success Factors (CSFs). Finally, after internal and external issues are taken into account, a viable and feasible business model design should have been reached. Together these steps form the STOF method of business model evaluation, see [5] for an elaboration.

In the following paragraphs we summarize the discussion of Bouwman et al. [3] on the four domains of the STOF framework. Value is the main component of the service domain in the STOF model. Value is further divided into intended and delivered value for the provider, and expected and perceived value for a customer. Service domain has additional components. Context encompasses both concrete situations and larger socio-cultural aspects of the usage environment of the service, and co-determines perceived value of the service. Tariff is the price paid for the service, and effort is made by a customer to use the service, both affecting perceived value of the service. Bundling of services generally increases perceived value of the service. Security can be an essential co-determinant in the value of the service.

The Technology domain focuses on technical architecture, which is used to deliver technical functionality. The technical architecture consists of applications, devices, service platforms, access networks, and backbone infrastructure. All these generate costs and affect the delivered value of the service. Intended value in turn puts requirements on the technical architecture and the value network behind the service. Security is generally a cost item.

The main component in the Organization domain is the value network, which consists of several actors and their interactions. Actors have strategies, goals, resources, and capabilities. They perform value activities which together with organizational arrangements are combined into roles. The organizational arrangements affect both interactions and financial arrangements of the actors. Value activities sett requirements on the technical architecture, and generate investment sources, costs, and delivered value. Security aspects can be included in the strategy of an actor.

The Finance domain determines pricing of the service. It consists of four sources which generate capital, costs, revenues, and risk.

Each domain has Critical Design Issues; see Bouwman et al. [4] for a detailed discussion on them. The Service domain has the following CDIs: targeting, creating value elements, branding and customer retention. The Technology domain CDIs include security, quality of service, system integration, accessibility for customers, and management of user profiles. The Organization domain consists of partner selection, network openness, network governance and network complexity. The Finance domain incorporates pricing, division of investments, division of costs and revenues, and valuation of contributions and benefits. Security aspects affect most CDIs.

Critical Success Factors exist for both customer and network value creation; see Bouwman et al. [4] for an elaborated discussion on them. The CSFs for creating customer value consist of clearly defined target group, compelling value proposition, unobtrusive customer retention, and an acceptable quality of service. The CSFs for creating network value include acceptable profitability, acceptable risks, sustainable network strategy and an acceptable division of roles. Reaching high scores on CSFs in both categories is expected to result in a service capable of generating both customer and network value, i.e., a service capable of meeting user expectations and motivating actor participation. Again, security aspects affect most CSFs.

B. Quantitative Techno-Economic Modeling

Our techno-economic model is depicted in Fig. 3. We calculate net present value (NPV) with revenues, operational expenditure (OPEX) and capital expenditure (CAPEX), tax percentage, and discount rate as inputs. *NPV* is defined as

$$NPV = \sum_{t=1}^{n} \frac{C_t}{(1+r)^t}\,, \qquad (1)$$

where t is the time of the cash flow, r is the discount rate and C_t is the net cash flow at time t. The cash flow is calculated by subtracting OPEX, CAPEX, and tax from revenues:

$$C_t = \sum_i R_i - C_{EX} - r_{tax}\left(\sum_i R_i - C_{EX}\right), \quad (2)$$

where C_{EX} is the sum of OPEX and CAPEX, and r_{tax} is the tax percentage.

The revenues R_i for product i are calculated by multiplying the following factors: number of product i purchased by one user n_i, the price p_i per product i, content legalization percentage, i.e., the percentage of content available to legalization, r_{CL}, and the number of users n_U. Value-added tax (VAT) and reimbursements after VAT to copyright holders and to users based on respective multipliers r_{VAT}, r_{RCH} and r_{RU} are deducted from revenues:

$$R_i = n_i p_i r_{CL} n_U \cdot \frac{1 - r_{RCH} - r_{RU}}{1 + r_{VAT}}. \quad (3)$$

The number of users n_U is calculated by multiplying the total population n by illegal downloading r_{ID}, broadband connectivity r_{BC}, market interest r_{MI}, and market penetration r_{MP} rates:

$$n_U = n r_{ID} r_{BC} r_{MI} r_{MP}. \quad (4)$$

The OPEX C_{OPEX} consists of content delivery network (CDN), user support, and marketing C_M costs:

$$C_{OPEX} = \frac{r_{PH} T \cdot 8}{3600} \mathrm{C}_{\mathrm{Mbps}} + C_F + C_{SR} n_{SR} + C_M. \quad (5)$$

The CDN cost has two elements: cost of a megabit per second (Mbps) C_{Mbps} and fixed cost C_F. The Mbps requirement is a multiplication of the total traffic T in MB and a peak hour load percentage r_{PH}. An even distribution of traffic during the peak hour is assumed. The user support is calculated by multiplying the cost per support request (SR) C_{SR} by the number of SRs n_{SR}.

The CAPEX C_{CAPEX} is a sum of person months (PMs) for both contract negotiation (CN) and software (SW) development & maintenance:

$$C_{CAPEX} = n_{CN} C_{CN} + n_{SW} C_{SW}. \quad (6)$$

Security aspects influence both OPEX and CAPEX, but due to the scope of our techno-economic model, they are only implicitly included.

We also carry out risk analysis by running Monte Carlo simulations. In a Monte Carlo simulation, uncertain variables are assigned random values according to predefined distributions. The simulation is repeated for thousands of trials. The impact on the results of the calculations is recorded for each trial. Based on the records, several statistical variables can be calculated. The statistical variables can then be used to assess the risk related to each scenario.

We perform the risk analysis with 100,000 trials for each scenario. The selected uncertain variables are assigned triangular distributions with expected, minimum and maximum values corresponding to mode, lower limit, and upper limit of the distribution, respectively.

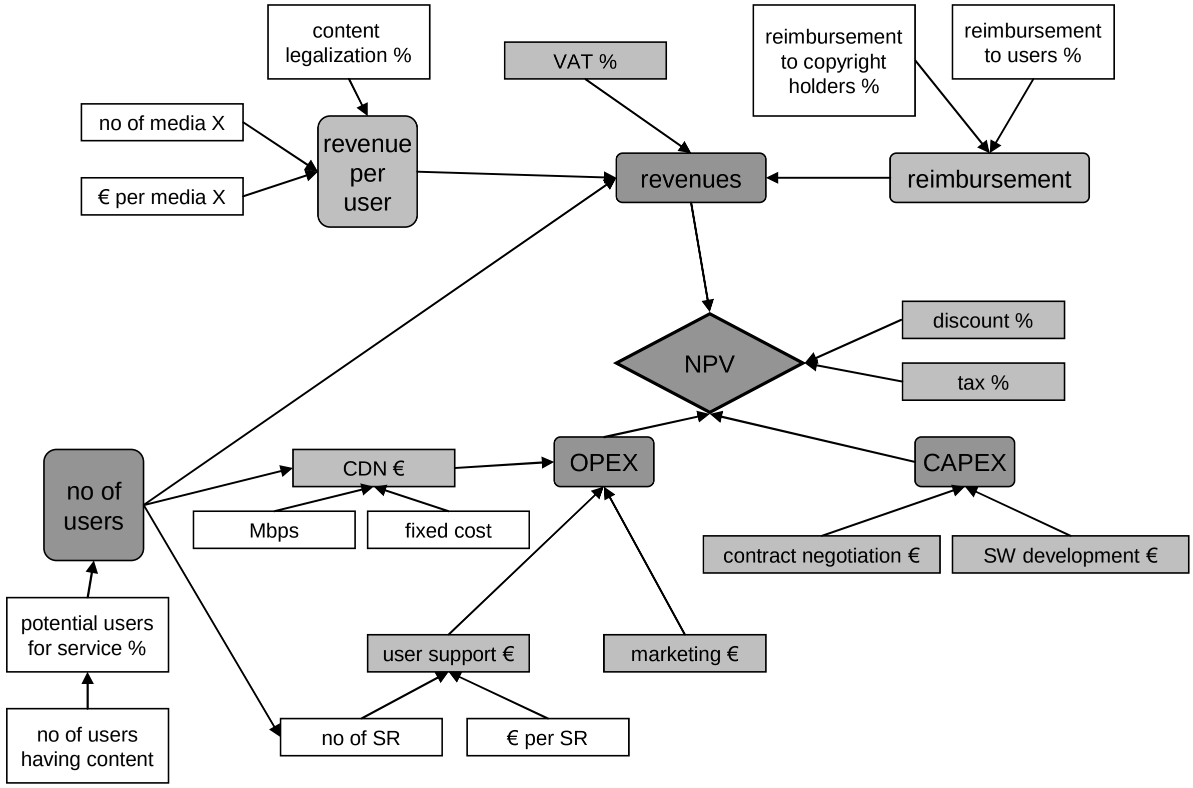


Figure 3. Quantitative techno-economic model for scenario comparison

C. *Privacy Threat Analysis*

To assess the privacy risks facing the users of the post-payment copyright system, we perform a threat analysis of identified key elements in the system. Threat analysis is an important part in security engineering and it forms the basis for the security design of the system [39]. For the threat analysis we form an attack tree as introduced by Schneier [40] covering possible attacks against the privacy of user's data. In our threat analysis, we consider following information items to be of special relevance to the post-payment copyright system: user identity, user contact information, information about the illegal content of the user, and the list of content paid with the post-payment copyright system.

The main goal for attacks, which we assume in our analysis, is to obtain potentially incriminating information about the user. The threats are considered to be related to illegal combining of user records in different parts of the post-payment copyright system, or to the threats introduced by direct external eavesdropping and active intrusion into system components. The attack tree used in our analysis is shown in Fig. 4.

IV. Scenarios

We compare three different online media rights purchase systems. They are an online media shop (A), a post-payment copyright system where the user's downloaded files are replaced with new files (B), and a post-payment copyright system where only immaterial rights are purchased by the user but existing files remain untouched (C).

The consumer price per content item is the same in all three cases. The rights for a song in the post-payment system cost exactly as much as the same song in an online music shop. Also, the reimbursement to the content owners is the same in all three cases.

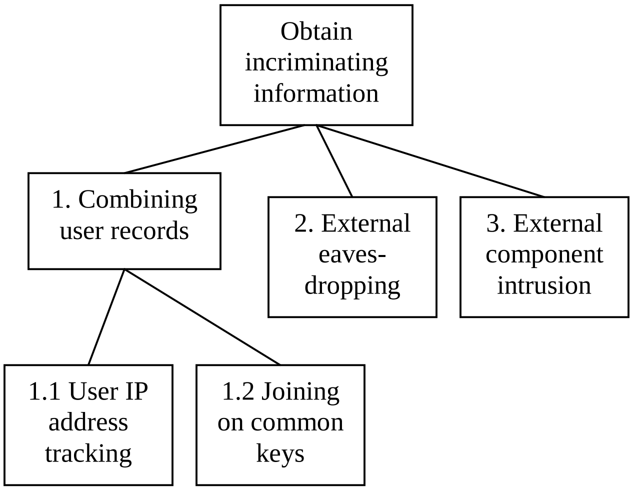

Obtain incriminating information
1. Combining user records
2. External eaves-dropping
3. External component intrusion
1.1 User IP address tracking
1.2 Joining on common keys

Figure 4. Attack tree identifying privacy threats against the post-payment copyright system

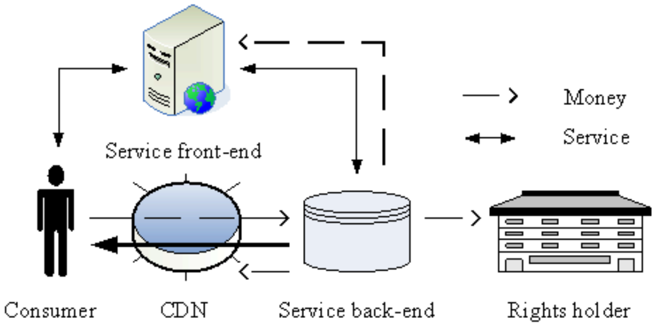


Figure 5. Online shop model

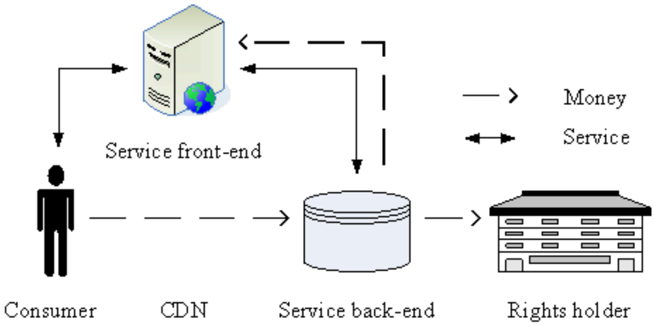


Figure 6. Post-payment copyright model

In the online shop model (see Fig. 5) the media companies and organizations representing artists are called rights holders. The service back-end company makes a distribution contract with the rights holder. The back-end service provider acquires the files, delivers them to the consumer and handles the payment of the user. The service visible for the consumer is provided by the service front-end. The service front-end is a web site marketing the service and building consumer's shopping cart. The service back-end provider uses a CDN to ensure a satisfactory content download service for the consumer.

The post-payment copyright system, where the files are replaced with new files, does not differ from the online music shop model. The post-payment part of the service is just a marketing tool for the content in the service back-end.

In the post-payment copyright system (see Fig. 6) no files are distributed. The service legalizes the unauthorized files on user's hard disks and memory cards. The service front-end handles the rights shopping cart for the user and the service back-end distributes the copyright fees to the rights holders.

The basic differences of the three scenarios compared are based on the market size and delivery cost differences depicted in Table 3. The online shop has all users of the potential market; post-payment variants have only the past users of illegal file sharing systems. The delivery and storage costs are considerable for both online shop and post-payment with file download. In practice the post-payment download services have the market potential of the online shop, but in this simplified comparison we study only the potential of the post-payment download feature.

TABLE III. COMPARISON OF THE SCENARIOS

Scenario	Market size	Delivery cost
Online shop (A)	Full	Full
Post-payment download (B)	P2P users	Full
Post-payment (C)	P2P users	Nominal

CDN adds additional security considerations and cost to the scenarios A and B. Additional cost related to security considerations is also present in contract negotiations, software development and user support for the post-payment scenarios B and C.

We use Finland in 2009-2013 as a case for our study. We use the population *n* of Finland in 2006 (5,276,955) as a basis for our population calculations and assume a 0.4% annual growth [41]. Table 4 summarizes the annual usage input values for our scenarios. In all scenarios, the initial value for the broadband connectivity r_{BC} is 53% in 2006 [42] which is extrapolated using a simple logistic saturation function with a saturation value of 65%.

In the scenarios B and C, the illegal downloading r_{ID} vector is an estimate based on [15]; the market interest r_{MI} is assumed to be a constant 20% based on [15]; the market penetration r_{MP} and its distribution into different user groups are hypothetical; and the penetration of a fourth user group "long tail small" is calculated by subtracting the other user group penetrations from 100%.

In the scenario A, r_{ID} is 100% every year because it is not relevant to the calculation of usage; an estimate of 15% in 2007 [43] with 16% annual growth [44] is used as a basis for r_{MI}; r_{MP} is hypothetical; and no user group distribution is used: only "long tail small" user group is in use.

We use only one product category: song with a price of €0.99 including VAT. A "parent," a "heavy user," a "long tail large" customer and a "long tail small" customer purchase annually 250, 500, 400 and 50 songs, respectively. In scenario A, a user buys annually 50 songs. Content legalization r_{CL} vectors are depicted in Table 5. In the scenario A, r_{CL} is 100%. Users are reimbursed (r_{RU}) 5% of their total purchases in the scenarios B and C, 0% in the scenario A. VAT r_{VAT} is 22% and reimbursement to copyright holders r_{RCH} is 80% in all the scenarios.

We use the following values for the calculation of OPEX. For the CDN cost, a base transfer of 10 MB per user is assumed in the scenarios B and C, 5 MB in the scenario A. In the scenarios A and B, a song transfer generates 5 MB of traffic. In the scenario C, each item generates only 1 kB of traffic (i.e., a transfer of a checksum). The peak hour load r_{PH} is 0.05% of total traffic assuming a 95th-to-mean ratio of 4:1 [45] in all the scenarios. In the scenarios A and B, the data transfer capacity cost C_{Mbps} is €5±0.5 per Mbps and the fixed cost C_F is €24,000±2,400 annually. In the scenario C, the prices are €1,000±100 and €8,000±800, respectively. The data transfer costs decrease annually by 5±0.5% in all the scenarios. For calculating the user support costs, we assume a fixed cost per SR C_{SR}: €5±0.5 in all the scenarios with an annual growth of 5±0.5%. In the scenario B and C, a parent generates 0.6 SRs annually, a heavy user and a long tail large user 0.4 SRs annually, and a long tail small user 0.2 SRs annually. In the scenario A, a user generates 0.2 SRs annually. The post-payment system is more complex, thus generating more SRs. Fixed marketing costs C_M are €50,000±5,000 in all the scenarios in 2009. In the scenario A they decrease by €3,000±300 annually; in the scenarios B and C by €10,000±1,000 annually. The difference in the marketing costs is based on the assumption that the online shop is marketed to a broad audience, whereas the post-payment systems are marketed to a limited audience. We estimate the catalogue matching costs in the scenarios B and C to be marginal; therefore they are not included in our calculations.

PMs for calculating CAPEX are depicted in Table 6. Regarding the cost of a PM, C_{SW} is €10,400±1,040 and C_{CN} is €20,800±2,080. Both have an annual growth rate of 5±0.5%. We assume the online shop is less complex to develop and deploy due to several existing solutions. The tax rate r_{tax} is 26% and the discount rate *r* is 10% in all the scenarios.

V. RESULTS

A. Qualitative Business Modeling

The dynamic business model framework in STOF has three phases, and each of them includes a STOF analysis. The phases are Technology R&D, Roll-out and Market. In this paper we analyze the Technology R&D phase empirically based on the literature presented in previous Sections of this paper. The business model is discussed from the post-payment copyright system operator point of view.

TABLE IV. ANNUAL USAGE INPUT VALUES (%)

	2009	2010	2011	2012	2013
r_{ID}	40±10	45±10	47.5±10	45±10	42.5±10
r_{BC}	55.4	59.9	62.1	63.3	64.0
r_{MI}[a]	20	23	27	32	37
r_{MP}	0	4.5±2.5	7±2.5	10±2.5	12±2.5
r_{MP}[a]	0	5±2.5	7.5±2.5	12.5±2.5	15±2.5
parent[b]	80±5	70±5	60±5	40±5	20±5
heavy user[b]	1±0.5	2±0.5	3±0.5	2±0.5	1±0.5
long tail large[b]	10±2.5	15±2.5	13±2.5	11±2.5	8±2.5

a. in the online shop scenario (A)

b. not in use in the online shop scenario (A)

TABLE V. CONTENT LEGALIZATION R_{CL} VECTORS (%)

	2009	2010	2011	2012	2013
parent	70± 10	74± 10	78± 10	82± 10	86± 10
heavy user	60± 10	62± 10	64± 10	66± 10	68± 10
long tail[a]	40±5	41±5	42±5	43±5	44±5

a. applies to both "long tail large" and "long tail small" user groups

TABLE VI. PERSON MONTH VECTORS

	2009	2010	2011	2012	2013
n_{SW}	24±3	3	1	1	1
n_{SW}[a]	8±1	3	1	1	1
n_{CN}	6±1	2	1	0.5	0.5
n_{CN}[a]	3±0.5	2	1	0.5	0.5

a. in the online shop scenario (A)

1) Service, Technology, Organization and Finance

The Service is the post-payment copyright service, which is discussed in detail in Section 2 of this paper. In the service the user can legalize the earlier illegally copied music files. In the early phase of deployment, the target customer group is Finnish peer-to-peer file sharing users. Most of them are males in the age group of 25-35 [15]. The value for the customer is that they can legalize the illegal music files, to which they listen a lot. The availability of Digital Rights Management free (DRM-free) online music is still very limited in Finland, and the service may attract consumers who prefer DRM-free music files.

The most of the Technology in the system is available in the existing online music shops. The technical architecture is described in Section 2 of this paper. The new technology in the system includes an algorithm to analyze illegal vs. legal status of the user files, catalogue matching server and user client. The user client integrates different system and service components together to form a smooth and compelling user experience. The illegal vs. legal music file classification may have serious legal consequences. Considering the early stage of the technology, the feature is introduced to consumers rather as a tool to help them to select the files they may want to legalize than to prove if the file is from an illegal or a legal source. The catalogue matching is also a critical component. The user payments are made according to the catalogue matches, not according to the real user files. The catalogue matching results are shown to the user and the user accepts to pay for the matching results, not for the music files. It is a considerable challenge to get correct matches from the multitude of music file descriptions to the limited coverage of the music catalogue.

The current state of the music industry impacts on the Organizational setup strongly. The digitalization of media, multimedia computers, and broadband connectivity has lead to a strong trend of illegal downloading and copying. The music industry continuously searches for new models for revenue and profit. At the same time they make effort to get publicity for the copyright issues in order to decrease the impact of illegal copying. From this perspective the rights holders, i.e., record companies and copyright organizations, have research interests in post-payment copyright system in addition to direct revenue and profit. For the existing online shops the incremental effort to take the post-payment copyright system into use is relatively small, and it opens the potential market of legalizing illegally copied media.

However, organizing the different actors into a viable value network is challenging. Users do not have strong incentives to legalize their content unless there is strong legislative pressure for it. Even though copyright legislation has developed partially according to the lobbying of the music industry, consumer rights advocates and the proponents of freedom in the digital domain are resisting with increasing force the sanctions of copyright infringements done by consumers. On the other hand, new actors to the digital media industry are establishing new business models where the ultimate goal is to possess the leading platform for digital media distribution. The new actors include both device manufacturers with their own distribution platforms (e.g., Apple iTunes and Nokia Comes with Music) and independent distribution platform providers (e.g., Amazon and Spotify). The role of devices and distribution platforms as part of digital music experience is under constant change and subject to re-definition.

The Finance of the business model is very similar to the online music shop business model. Artists, composers and technicians have contracts with record labels and copyright organizations. The back-end system provider has contracts with record labels and copyright organizations to pay them a certain portion of the consumer price. The back-end provider has a contract with financial institutions to enable payments in exchange for a fee.

The post-payment system operator has a similar role as the web front-end in the online music shop white label business model. It concentrates on the marketing of the service and operates as a reseller, whereas the back-end system provider acts like a wholesales organization. The consumer is the customer paying for the service. More indirect revenue models also exist: capitalizing on the copyright campaign nature of the service, reselling the information about consumer music preferences, and advertising in various parts of the system, but they are not evaluated in this paper.

2) Critical Design Issues

The first Critical Design Issue is targeting by making the choice between the business to consumer (B2C) and business to business (B2B) models. In B2C the post-payment service provider markets the service to consumers and receives payments from them. In the B2B model the post-payment service provider offers the service to rights holders or to an existing online music shop. In first service trials the model must be B2C, but when the service establishes itself on the market, dedicated B2B post-payment service providers are to be expected.

In creating value elements for the basic post payment copyright, the system can be enhanced by providing access to high quality music files, album art, and song lyrics. An important issue in the value context is Digital Rights Management (DRM). In some cases a file without DRM is of higher value to a consumer, even when its source is illegal, compared to a commercial legal file with DRM.

The service branding is closely related to the choice between the B2B and B2C options. The possible branding alternatives include own branding, use of an existing online music shop brand, or a connection to a rights holder brand, e.g., a brand of a record label or a copyright agency.

The customer retention of a post-payment copyright system to an online music shop can be arranged so that the user account or the payment information of the post-payment copyright system is bundled with the online music shop. The customer retention to the post-payment copyright system itself can be obtained through personalization: the system can store the user classification of files to be legalized, being legal, being illegal, among other possibilities.

The security of the post-payment copyright system has two distinctive parts. The information related to payment has very high security requirements which are met by established solutions. The unique security challenge in the system is privacy. The personal information in the system basically describes evidence for copyright law violation. The collection, communication and storage of the personal information must be well understood both by the operator and the consumer. Generally, it is an advantage if the communication is transparent and it is possible for an advanced user to check if the communication contains only the promised information. On the other hand, unencrypted transfers are prone to eavesdropping, so the data sent should be protected by encryption.

The quality of service consists of the time needed for downloading the application, scanning the hard disk, analyzing the legal status of files, and the catalogue matching. From the privacy point of view it would be beneficial to carry out the catalogue matching in the user application, but it would increase the download time dramatically. In this respect the quality of service consideration exceeds the privacy concern. In order to improve the time consumed at each phase of the process it is possible to develop algorithms which optimize the time used for processing while maintaining the accuracy of the results.

In the early phase when the market is building up, system integration to several other systems may not seem relevant. But especially in the B2C model the possibility to integrate the system to various online backend systems is crucial. A generic modular structure and well defined simple interfaces using common technology components decrease the time needed for individual system integration projects.

The user application providing the best accessibility for customers is web browser. On the other hand, the web browser is not allowed to get full access to the computer where it is running. Plug-ins like Java script engine and Adobe Flash player are a compromise between the full access rights on the local computer and the need to download and run an application. The latest plug-in technology versions normally have the most advanced feature sets and include many functions, which can speed up the development phase, but the support for them may not be available on all common computing platforms.

Management of user profiles gives the possibility for service personalization, but they also create a difficult situation for the service provider. The service provider has knowledge of the copyright violation of the user, and access to the user account which can link to the real personality of the user. In many jurisdictions police may force the service provider to reveal such information to copyright authorities. In order to avoid such procedures it may be preferable for the service provider not to have user accounts even if the absence of them decreases the possibility for service personalization and customer retention.

When considering the partner selection, the added value of the service provider is related to the contribution it makes for the value network. In one extreme the rights holder could run the post-payment copyright service bundled with its own online music shop. In the early phase it might be useful to have a number of partners, which are specialized in certain parts of the value chain, at least in order to learn how those parts of the value chain typically operate. Having partners enables the service provider to concentrate on the novel parts of the system where it most likely can add significant value.

Concerning network openness, network governance and network complexity, a balance between the network growth and control of the network has to be maintained. The post-payment copyright value network has generally better possibilities to grow uncontrolled, but for the post-payment system operator and for the early players, the uncontrolled growth may lead to lost opportunities and lost market share. A realistic and attractive business case could be to specialize in one part of the system when the value network has potential to grow. The specialization could be providing the post-payment system as a back-end service for existing online music shops, delivering catalogue matching system or user application. Also licensing, consulting and system integration services for the entrants can be considered.

Pricing is probably the most important factor for any product. In the post-payment copyright system, the reference price point is the price of a piece of music in an online music shop. As the system does not need to distribute copies of a file, a lower price point could be justified. On the other hand, the system provides a substitute product for the online music shops and it would not be logical to allow very different price points for substituting product formats. Special pricing according to the quantities should at least be considered so that the system would encourage the users to legalize as many files as possible.

The division of investments for the system includes the development of the system and marketing efforts. Our assumption is that these costs are shared by the rights holder, the online music shop providing the contracts and the back-end system, and the post-payment copyright system provider. The development costs are most naturally carried by the post-payment copyright system provider, because the service is its own initiative. The marketing costs should be shared more evenly.

The post-payment copyright business contracts define the division of costs and earnings. The post-payment system operator may not be able to include all its development costs into new contracts, but at least the amount it would cost for a new player to develop the same system can be taken into account. Basically the same applies for the contribution of the other partners as well.

In the contract negotiations between the partners probably the most important issue is how the valuation of contributions and benefits is carried out. A joint venture can be created where the partners are investors, or more typically each link in the value chain forms a customer – service provider relationship. Each link can be negotiated as a fixed fee, transaction based revenue sharing, or a combination of them.

3) *Critical Success Factors*

Several Critical Success Factors are not positive by default in the service. The compelling value proposition to the customer depends on the viewpoint. On the other hand, no other way to legalize illegal downloads exists in most markets. But as the consumer already has the music file and does not get any concrete value by paying for it after obtaining it, the value proposition is not strong.

The clearly defined target customer group consists of users who have downloaded and copied music illegally in the past, and of the parents of children who have illegal copies of music files. The target groups are rather well defined, although convincing them to use the service is very challenging.

While a compelling value proposition is not strongly present, the value analysis of the service design provides additional insight to the case. The intended value for the end-user customer is the possibility to legalize illegal copies of music files. This gives a covenant not to sue protection to the customer. The delivered value depends on the contracts between the service provider and the rights holders. We can expect that the contracts are professionally made and the jurisdiction has the concept of covenant not to sue, or the freedom of the contract prevails, so that the delivered value matches well to the intended value. The expected value is a more challenging aspect. In most services, the user gets something when he pays money for the service. In the case of a post-payment service, the user gets a receipt. The receipt may not be tangible enough to drive the user to make a purchase decision. Additional material like CD covers, track lyrics, or other material about the artist could improve the perceived value for the customer. On the other hand, the music file to be legalized is known by the consumer. He knows the content, how much he listens to the file, what is the coding quality is, and in which devices the file can be played. Hence, the perceived value of the content equals to the delivered value in the post-payment copyright system.

Unobtrusive customer retention is a challenging aspect. The core offering is to give a unique opportunity to legalize illegal downloads. Furthermore, the user is encouraged to use legal music shops instead of illegal means. From this perspective, customer retention is realized only when the user legalizes just a part of the illegal files in possession when the service is used for the first time. Another possibility for customer retention is that the service is bundled with an online music shop, and customers accessing the music shop are considered as a part of the overall customer retention.

The acceptable quality of service is guaranteed in the web front-end and the payment services, as they follow industry standards. The specific areas to develop the quality of service are illegal vs. legal file recognition and music catalogue matching. Illegal file classification is a recent area of forensics, and major improvements can still be expected. The implementation of the catalogue matching is a trade-off between accuracy, number of methods in use, and resources. The accuracy can be improved by adding more methods for matching. For example, fingerprint recognition can be used in addition to metadata analysis. The accuracy of a method can be increased by adding execution cycles, i.e., increasing the number of times the method is applied. Additional methods require increased development resources. Running methods in parallel and increasing execution cycles require processing resources.

The acceptable profitability of the post-payment copyright service is quantitatively analyzed in Section 5.2.

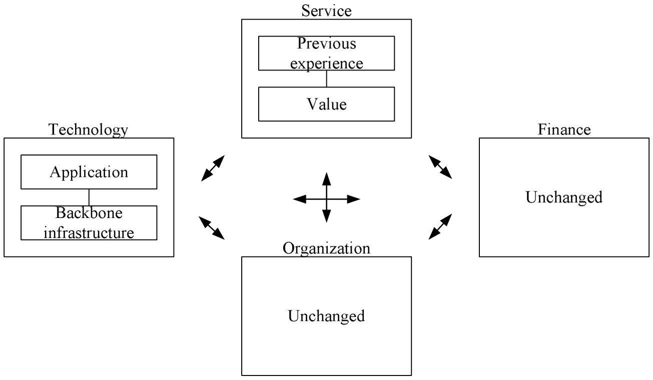


Figure 7. Differences in post-payment system and online music shop STOF models

The acceptable risks are gained by re-using the existing technology assets of the partners. The main investments in the beginning are related to client development, catalogue matching server development, and marketing. In this analysis, we expect that the development risk is taken by the post-payment system operator, and the marketing risk is shared between the participating organizations.

The sustainable network strategy has a good basis, because for all others except for the post-payment system operator, the service is an extension in their current business operations. For most of the participating organizations in the value chain adding the new service does not require new technical development and could be described as “business as usual”.

The acceptable division of roles is achieved by having the same roles as in the online music shop value chain. The roles follow industry standards, and the same applies for the profitability and risk. Post-payment system providers have potential to negotiate better contract terms than online music shop providers due to the advantages of the post-payment copyright system, including the opportunity to increase consumers’ moral regarding copyright.

4) Differences in post-payment and online shop STOF models

As part of the STOF analysis, we compare the Service, Technology, Organization, and Finance domain descriptive models of the post-payment copyright system and the online music shop. For both of them, the Organization and Finance domains are similar according to our analysis. The differences are visible in the Service and Technology domains, see Fig 7.

In the Service domain, the customer most likely has previous experience of online music. For the post-payment system only the payment experience exists. The post-payment copyright experience is new for the customer. The customer does not know how long scanning, or catalogue matching takes. Also, the accuracy of matching and illegal vs. legal classification is without earlier references. The intended value of the online music shop is that the user is able to listen to the purchased music file. The perceived value of the DRM protected content can be lower, if the customer would like to play the files in a device without the DRM client of the online music shop. In the post-payment system, the content itself is well know by the customer, as it is already in her possession, but due to that the delivered value of the system is abstract rather than practical.

In the Technology domain the payment and customer data platforms are similar in both online shop and post-payment systems. The main differences are in the client and in the backbone infrastructure. The online music shop client has a download feature, and often it also has its own music player with the DRM client installed. The post-payment system client includes the scanning and illegal vs. legal classification but no download or player features. The main difference in the network components is the catalogue matching service existing only in the post-payment copyright system.

B. *Quantitative Techno-Economic Modeling*

With the base case parameters, the scenario A produces total revenue of €1.49 million and the scenarios B and C €1.43 million. Therefore, the scenarios are at a comparable revenue level. The online shop scenario (A) has a linear revenue curve, whereas the post-payment scenarios (B and C) have a peak curve.

Table 7 depicts the results of break-even analysis in the base cases. The break-even point is reached when NPV is zero. The break-even market penetration rate $r^{MP}_{B\text{-}E}$ in the table is defined as the number of users in the break-even situation. It is calculated as the number of users n_U multiplied by a break-even multiplier $r_{B\text{-}E}$, which is set so that NPV is zero divided by the number of potential users, i.e., the number of users n_U divided by market penetration rate r_{MP}:

$$r^{MP}_{B-E} = \frac{r_{B-E} n_U}{n_U / r_{MP}} = r_{B-E} r_{MP} \,. \tag{7}$$

The post-payment scenario (C) reaches the lowest break-even market penetration rates, followed by the post-payment download scenario (B). Thus, post-payment scenarios require less market penetration among potential users than the online shop scenario (A) to reach a break-even situation in the base case.

The results of the NPV analysis are depicted in Fig. 8. The first bars illustrate the mean values, whereas their error bars display the minimum and maximum values. The second bars are base values. The mean, minimum, and maximum values are calculated based on the risk analysis, whereas base values represent the results without risk analysis. The results clearly indicate that the post-payment scenario (C) is the most profitable and has mid-level risk, whereas the post-payment download scenario (B) is the least profitable and has the most risk.

According to our sensitivity analysis, the usage parameters have the largest effect on the outcome in all the scenarios. Because we did not perform extensive sensitivity analysis eliminating the effect of potential cross-correlations of variables, we do not present the results in detail.

Fig. 9 depicts the results of the cost analysis. Mean values and error bars are displayed. Reimbursement to copyright holders is the most significant cost item in all the scenarios. The other items are notably less substantial. The differences in scenario definitions are clearly visible in the cost structure of each scenario.

TABLE VII. BREAK-EVEN MARKET PENETRATION RATES $R^{MP}_{B\text{-}E}$ (%) IN THE BASE CASE SCENARIOS

Scenario	2009	2010	2011	2012	2013
A	0.0	4.4	6.6	11.0	13.3
B	0.0	4.2	6.6	9.4	11.3
C	0.0	3.9	6.1	8.7	10.4

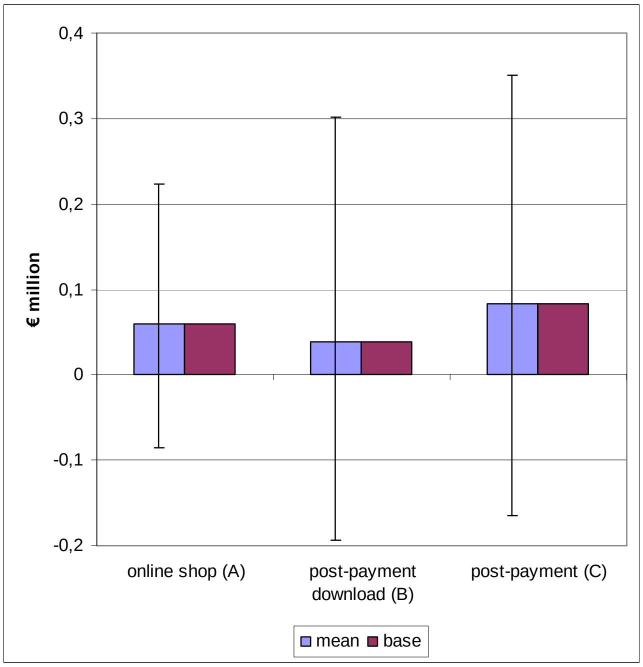


Figure 8. Results of NPV analysis

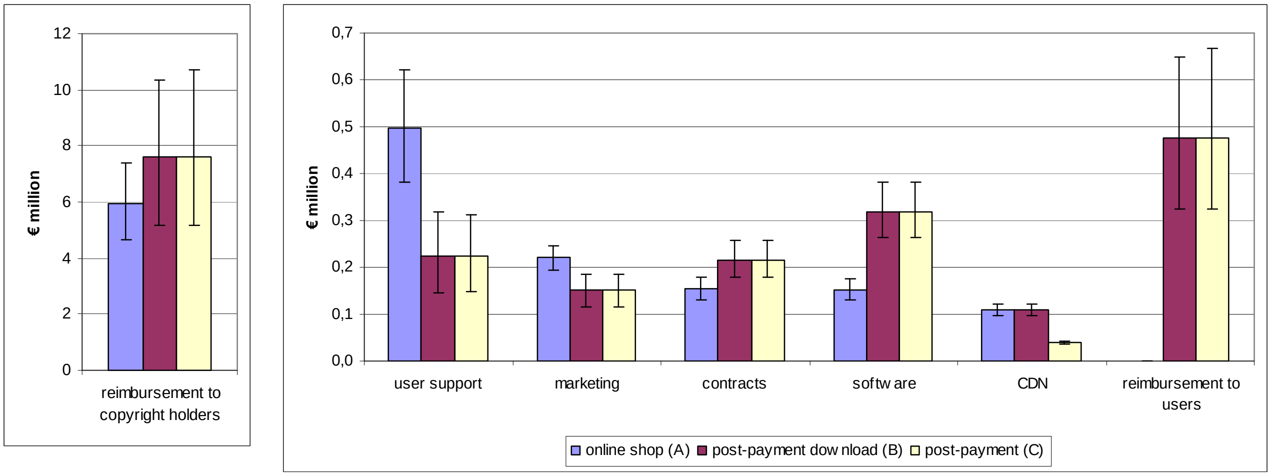


Figure 9. Results of cost analysis

C. *Privacy Threat Analysis*

In this section, we present the results of the attack tree method -based privacy threat analysis, see Fig. 4. We describe how the system provider can try to assure the user that private information is not compromised. For the success of the post-payment copyright system, it is particularly important that the users of the system can be confident that the user identity in combination with the list of illegal content in the user's possession does not leak out of the system. The goal of the privacy protection measures is to limit the privacy exposure of the users.

In our analysis, the threat class 1 *combining user records*, relates to the improper re-identification of the user and linking of data records between system components. The most critical re-identification risk is related to the requests to the catalogue matching server. If the requester is identified and the requests tracked, one can obtain an extensive list of potentially illegal content on the requester's device. That can obviously be incriminating evidence against the user. While the request protocol as such is based on anonymous operation, the source IP address of the requests still remains traceable. In some cases, the IP address may be linkable to a specific device. That can potentially reveal the requester's identity, described as *user IP address tracking*, threat class 1.1.

Even if the IP address cannot be traced back to a specific device or person, the IP address provides a potential key, based on which information between the system components could be linked. If the front-end and the catalogue matching server colluded by comparing the IP addresses of requests they serve, it would be possible to link the shopping cart information with the actual list of illegal content on the user's device. For a comparison of privacy risks arising from collusion of individual system components, see Table 8.

TABLE VIII. COLLUSION RISKS AGAINST USER PRIVACY

Collusion risks	**Worst-case privacy exposure**	**Colluding components**		
		Catalogue matching server	***Front-end and back-end***	***Payment system***
Linking of catalogue matching requests to purchase sessions based on IP address	User contact information + list of content to purchase licence for + list of illegal content in user's possession	X	X	
Linking of shopping cart information to payment information leading to full user identification	User identity + User contact information + list of content to purchase licence for		X	X
Linking of catalogue matching request to payment transactions based on IP addresses	User identity + list of illegal content in user's possession	X		X
Linking of catalogue matching requests to shopping carts and payment information leading to full privacy exposure	User identity + User contact information + list of content to purchase licence for + list of illegal content in user's possession	X	X	X

Even when there is no malicious intent against the users in the system itself, there are external privacy threats. They are related to the user information that external attackers can get by eavesdropping. The external communication threats are threat class 2 *external eavesdropping* of unprotected user communications, and threat class 3 *external component intrusion*. Note that also law enforcement officials using a search warrant can be regarded as such external attackers against the users' privacy. Once an intrusion happens, the intruder can of course learn user information, which is hosted on the system component. Increase in privacy exposure by combining this information to other data is much more challenging, as can be seen in Table 9.

From the privacy point of view, threats in class 2 related to eavesdropping are as problematic as the collusion threats in class 1 *combining user records*. Unlimited eavesdropping on the system communications would enable an external attacker to know the same information as the system components. However, as all communications between the user’s device and the system are protected by using encryption, we estimate that the probability of successful eavesdropping attacks against the user is very low. The eavesdropping will not impact users’ privacy perceptions so much that it would have a negative impact on the adoption of the post-payment copyright system.

After analyzing the threats, we present solutions for the recognized privacy threats of the post-payment copyright system. We also go through the tradeoffs of the selected solutions, see Table 10.

Distributing the user information in the system decreases the impact and risk of the threat class 1 *combining user records*. Storing payment information and receipt information is regulated by the legislation. The information is distributed in the system: backend provides clearing the payments for the rights holders; payment system operator handles the payment; and the service front-end operator works with the shopping basket. The distribution may decrease the consumer privacy concerns, as no one in the system has information about the user files, real identity, and credit card information. All these transactions can be collected and linked together for example in a copyright infringement claim investigation. If the user loses the receipt and likes to get another copy of the receipt, the distribution of the data makes the task more challenging.

Academic research is a valid reason to collect private and sensitive information in copyright legislation. At this state of the work, the post-payment copyright system is a research project. The research part is clearly separated from the commercial service, it has its own privacy statements, and the difference is clearly communicated to the user.

TABLE IX. PRIVACY EXPOSURE IN EXTERNAL INTRUSION THREATS

Linking scenarios	Privacy exposure
Linking of catalogue matching requests to purchase sessions	Linking cannot be done, since catalogue matching requests are anonymous and IP addresses are not tracked
Linking of shopping cart information to payment information leading to full user identification	Back-end cannot identify user, since payment system provides only information about the success of the transaction not about payment information
Linking of catalogue matching request to payment transactions	Linking is not possible since there is no linking key. IP addresses are not tracked.

TABLE X. PRIVACY SOLUTIONS, TRADEOFFS AND BENEFITS

Solution	Threat	Tradeoff	Benefit
Distributed architecture	1	More privacy statements, service agreements, and other legal documents. More complicated management of the system.	All information related to the user is not available at one point
Research data separated	1	More complicated data storage system. More complicated data structures in the research analysis.	Less private information stored long periods
No user accounts	1	Less possibilty for personalization of the service	Less information can be linked to a user
Anonymizing proxy	1.1	Service cannot use any information gained during the previous session. Research of data gets more difficult.	Extremely difficult to link any service use to the user
Encrypted communication	2	Difficult for user to verify the data sent to the network	No eavesdropping at intermediaries
Illegal vs. legal analysis at user device	1	The accuracy of analysis is decreased	The most sensitive information is not transmitted in the Internet or stored on the servers
Catalogue matching only for items which user considers paying	1	The only a coarse price of the music available in the client for all user files	Minimized personal data transfers from the device to the network
User data not stored	1.1	Tedious tracking of all purchase data afterwards at the service provider. No possibility for service personalization.	As little as possible personal data is stored in the system
Rights holders clearly visible	Adds trust	May raise worries about hidden agenda of the rights holders	Increases trust for the validity of the service

Storing information about the users' past illegal activities is naturally a very sensitive matter. It should be handled with similar care as the patient registers in health care. Collection of such information would be an attractive target for internal and external attacks, and it could potentially be interesting for authorities. From a user point of view, this would represent threats that could deter users from using the service in the first place. Therefore, we decided not to store any of the information that links a specific user to the illegal past actions. The result of with this decision is that we don’t have any user accounts on the web frontend. The user accounts could be used to help the user during the consecutive usage sessions and to provide a possibility for service personalization.

A commonly used method to protect against the threat class 1.1 *user IP tracking*, is that the service is used through an anonymizing proxy. If all system components maliciously collude against the user of the system, user privacy is compromised leading in the worst case to full exposure of the user's private information. This is clearly unacceptable for the users of the system. From the user point of view, the best improvement here would be that the users would apply an anonymizing proxy or an anonymizing network in all interactions with the system.

As a solution for the threat class 2 *external eavesdropping*, we apply encryption in all communication in the service. We use Secure Socket Layer (SSL) to decrease the concerns about potential eavesdropping or network monitoring by authorities. Also here, the selection of the encrypted communication is a privacy concern tradeoff. The positive impact is that the users need not to be so worried about their sensitive information being transmitted in clear text through the Internet. The negative side is that for people like civil liberties and web activists, who would really like to check what is communicated between the application and the servers, it is challenging to verify that the service provider promises about the communicated content match with reality.

The exposure to the threat class 1 *combining user records* is decreased by carrying out the illegal vs. legal analysis in the user device only. As a tradeoff of the client based analysis we lose the centralized analysis help in researching the accuracy of the analysis. The centralized analysis would also allow faster deployment of the improvements to the users. But as that information is the most sensitive in the system, the user privacy was selected as the dominating factor in the implementation.

An opposite tradeoff was accepted with the catalogue matching functionality. It is implemented on the network server. In this case, we therefore trade the privacy concern for a small application download size and for a more reliable catalogue matching in the system. The impact of the threat 1 combining user records with the selected architecture is decreased by carrying out the catalogue matching only for the items, which user considers paying, and by clearing all private data like IP addresses in the logs of the catalogue matching server. The threat class 2 *external eavesdropping* is made very difficult by encrypting all communication with the catalogue matching server. It is however to be noted that encryption alone cannot hide the fact that a specific source IP address has been in interaction with the catalogue matching server. Even this information might be considered incriminating. The alternative of the adopted solution would have been to download the entire catalog of available content to the user device and to perform the catalogue matching locally, thus minimizing the exposure of the user.

In addition to protecting the user against the negative privacy threats, we try to use a positive approach to improve the user trust in the post-payment system. The privacy concerns of the users are greatly diminished if they can consider the post-payment copyright service as a trusted third party that does not forward sensitive data to the rights holders. The challenge of the post-payment copyright service is to act as a reliable trusted third party between the users and the rights holders. The trusted third party should be credible in the users' perception and make the users willing to use the service.

Communication about the system plays a very important role. We build the system architecture and operating process to minimize privacy concerns. The consumers become aware of these solutions through communication. With successful communication about the selected solutions we can build trust in the consumers and lower their privacy concerns. Peer-to-peer is a very sensitive issue from the communication point of view. In the communication, it might be wise not to mention peer-to-peer as a source of illegal copies in order to avoid the wave of emotional bursts in the peer-to-peer user groups, leading to a decreased trust within other post-payment users.

Showing clearly which rights holders are behind the post-payment system builds trust in the consumer. It convinces the users that the service really delivers what it promises. At the same time, quite a few consumers may wonder if the rights holders have a hidden agenda in the service. The users might get increased privacy concerns about what data is really collected and to whom it is given.

VI. Discussion

Our quantitative study shows that a post-payment copyright system is potentially a more profitable business than an online music shop. However, the study is limited by the definition of the inputs to the model and the simplifications used in the model. In reality, the outcome may differ significantly from our results.

In our study, the models are separated, whereas in commercial systems the post-payment system will be linked to other models. We have assumed that the users of a post-payment system receive electronic vouchers to online shops as reimbursements of their transactions. The vouchers encourage post-payment users to buy their digital music in online shops.

The largest source of inaccuracy is the number of people using the service. The popularity of the service is very difficult to estimate due to the following factors: the development of P2P networks, broadband connections, digital rights management, upcoming legal implications on P2P networks, popularity of digital media, the perceived usability of the service, and the benefit from the service.

Comparing the scenarios, CDN cost for file storage and delivery is very deterministic, because the post-payment without download scenario benefits significantly from the absence of media retention and transfer. There are ways to the cost: limiting the market area geographically, replacing CDN with own servers, locating servers close to an Internet exchange point, having point of presence in the Internet exchange point, and leasing own fibers.

Our research is a continuation to the trend of studies suggesting P2P networks as a part of a viable business model for media distribution [15]-[18]. We also demonstrate the usefulness of techno-economic modeling and associated risk analysis when making decisions regarding the development and deployment of a new online service. Our model could be generalized for the analysis of different types of media, for instance the online distribution of movies, and potentially extended with real options analysis [46].

The qualitative analysis with STOF helps to form a working business model and to identify the relevant partners. The main benefits of using STOF instead of other potential analysis frameworks are its holistic and systematic approach and its fit to novel online services. With the help of STOF analysis, we were able to create a more solid business model for post-payment copyright. The post-payment system architecture (Fig. 1) and the post-payment process (Fig. 2) support directly the new business model.

The Critical Success Factor evaluation reveals that the service design and enrollment of the service may have a great impact on the revenue generated by the service. The main challenges in the post-payment copyright system are low value for customers and low customer retention rate. Service bundling with an online music shop offering and careful consideration of marketing message are suggested as solutions for value perception and customer retention.

Our analysis does not concentrate on the Roll-out and Market phases of the STOF framework. In the actual deployment of the service, a current and detailed analysis of the market situation and the relevant competing services should be made. De Reuver et al. [47] demonstrated that addressing Critical Design Issues in sufficient detail leads to better scores in Critical Success Factors, i.e., a viable service design improves the chances for actual success. Furthermore, according to de Reuver et al. [48], having a balanced business model internally is not sufficient for success. The business model also has to be continually balanced to changing market, regulation, and technology conditions in different phases of the service lifetime.

Based on our threat analysis, it seems that trustworthiness of the system is a key factor for users to be willing to use the post-payment copyright system. The users need to be able to trust the system. It must not maliciously act against the user in any way, and it has to protect the user against external threats. If we presume that the post-payment copyright system implements the suggested privacy solutions, we can conclude that it provides quite sufficient protection for the user's privacy even against external threats.

Our study does not confirm that a post-payment copyright system will be the winning model, but the study shows that in favorable conditions post-payment copyright is a very competent model compared to the online shop model. The privacy challenges play in an important role in the user adoption of the service, and solutions for the most important challenges are available. Our recommendation is to do further evaluation among industry experts and end-users and finally to test the validity of our results with a live post-payment service.

Acknowledgment

An earlier version of this work was presented at the ICDS 09 conference, see [1]. We would like to thank our colleagues, especially Mark de Reuver, Asta Kärkkäinen, Sakari Luukkainen, Ben Mellin and Jukka K. Nurminen, the reviewers of this journal, and the reviewers and participants of the ICDS 09 conference for their feedback on previous versions of this work.

The research was partially conducted in the DECICOM project funded by the Finnish Funding Agency for Technology and Innovation (Tekes), Nokia, Ericsson, and NetHawk and partially within the Econ@Tel (COST605) network.

References

[1] M. V. J. Heikkinen and H. Kokkinen, “Techno-Economic Modeling of Post-Payment Copyrights,” Proc. Intl. Conf. on Digital Society (ICDS 09), IEEE Press, Feb. 2009, pp. 217-222, doi:10.1109/ICDS.2009.12.

[2] H. Bouwman, H. De Vos, and T. Haaker, Eds., Mobile Service Innovation and Business Models, Springer, May 2008, doi:10.1007/978-3-540-79238-3.

[3] H. Bouwman, E. Faber, T. Haaker, B. Kijl, and M. De Reuver, “Conceptualizing the STOF Model,” in Mobile Service Innovation and Business Models, H. Bouwman, H. De Vos, and T. Haaker, Eds. Springer, May 2008, pp. 31-70, doi:10.1007/978-3-540-79238-3_2.

[4] H. Bouwman, E. Faber, E. Fielt, T. Haaker, and M. De Reuver, “STOF Model: Critical Design Issues and Critical Success Factors,” in Mobile Service Innovation and Business Models, H. Bouwman, H. De Vos, and T. Haaker, Eds. Springer, May 2008, pp. 71-88, doi:10.1007/978-3-540-79238-3_3.

[5] H. De Vos and T. Haaker, “The STOF Method,” in Mobile Service Innovation and Business Models, H. Bouwman, H. De Vos, and T. Haaker, Eds. Springer, May 2008, pp. 115-136, doi:10.1007/978-3-540-79238-3_5.

[6] P. Timmers, “Business Models for Electronic Markets,” Electronic Markets, vol. 8, Apr. 1998, pp. 3-8, doi:10.1080/10196789800000016

[7] R. Amit and C. Zott, “Value Creation in E-Business,” Strategic Management J., vol. 22, June 2001, pp. 493-520, doi: 10.1002/smj.187.

[8] H. Chesbrough and R. S. Rosenbloom, “The Role of the Business Model in Capturing Value from Innovation: Evidence from Xerox Corporation’s Technology Spin-Off Companies,” Industrial and Corporate Change, vol. 11, pp. 529-555, 2002.

[9] R. Alt and H.-D. Zimmerman, “Introduction to Special Section: Business Models,” Electronic Markets, vol. 11, Jan. 2001, pp. 3-9, doi:10.1080/713765630.

[10] A. Osterwalder, Y. Pigneur, and C. L. Tucci, “Clarifying Business Models: Origins, Present, and Future of the Concept,” Communications of the Association for Information Systems, vol. 16, pp. 1-25, 2005.

[11] S. M. Shafer, H. J. Smith, and J. C. Linder, “The Power of Business Models,” Business Horizons, vol. 48, 2005, pp. 199-207, doi:10.1016/j.bushor.2004.10.014.

[12] H. Kokkinen, J. E. Ekberg, and J. Noyranen, “Post-Payment System for Peer-to-Peer Filesharing,” Proc. Consumer Communications and Networking Conf. (CCNC 08), IEEE Press, Jan. 2008, pp. 134-135, doi:10.1109/ccnc08.2007.37.

[13] H. Hietanen, A. Huttunen, and H. Kokkinen, “Laila: File Sharing Indulgence Service,” NIR Nordic Intellectual Property Law Review, vol. 78, pp. 175-180, 2009.

[14] H. Kokkinen and J. Nöyränen, “Forensics for Detecting P2P Network Originated MP3 Files on the User Device,” in Forensics in Telecommunications, Information and Multimedia, LNICST 8, M. Sorell, Ed. Springer, May 2009, pp. 10-18, doi: 10.1007/978-3-642-02312-5_2.

[15] H. Hietanen, A. Huttunen, and H. Kokkinen, “Criminal Friends of Entertainment: Analysing Results from Recent Peer-to-Peer Surveys,” SCRIPTed, vol. 5, 2008, pp. 31-49, doi:10.2966/scrip.050108.31.

[16] D. Y. Cohn and V. L. Vaccaro, “A Study of Neutralization Theory’s Application to Global Consumer Ethics: P2P File-Trading of Musical Intellectual Property on the Internet,” Intl. J. Internet Marketing and Advertising, vol. 3, pp. 68-88, 2006.

[17] M. Peitz and P. Waelbroeck, “An Economist’s Guide to Digital Music,” CESifo Economic Studies, vol. 51, pp. 359-428, 2005.

[18] S. Bhattacharjee, R. D. Gopal, K. Lertwachara, and J. R. Marsden, “Economic of Online Music,” Proc. Intl. Conf. on Electronic Commerce, ACM Intl. Conf. Proc. Series, vol. 50, 2003, pp. 300-309, doi:10.1145/948005.948045.

[19] L. A. Ims, Ed., Broadband Access Networks: Introduction Strategies and Techno-Economic Evaluation, London: Chapman & Hall, 1998.

[20] N. K. Elnegaard and K. Stordahl, “Analysing the Impact of Forecast Uncertainties in Broadband Access Rollouts by the Use of Risk Analysis,” Teletronikk, vol. 4, pp. 157-167, 2004.

[21] T. Monath, N. Kristian, P. Cadro, D. Katsianis, and D. Varoutas, “Economics of Fixed Broadband Access Network Strategies,” IEEE Communications Magazine, vol. 41, Sept. 2003, pp. 132-139, doi:10.1109/MCOM.2003.1232248.

[22] B. Jerman-Blažič, “Techno-Economic Analysis and Empirical Study of Network Broadband Investment: The Case of Backbone Upgrading,” Information Systems Frontiers, vol. 10, 2008, pp. 103-110, doi:10.1007/s10796-007-9059-y.

[23] K. R. R. Kumar and V. Y. H. Kueh, “Techno-Economic Analysis of International Mobile Roaming,” IEEE Wireless Communications, vol. 15, June 2008, pp. 73-80, doi:10.1109/MWC.2008.4547526.

[24] T. Smura, A. Kiiski, and H. Hämmäinen, “Virtual Operators in the Mobile Industry: A Techno-Economic Analysis,” Netnomics, vol. 8, Oct. 2007, pp. 25-48, doi:10.1007/s11066-008-9012-3.

[25] E. Kivisaari, T. Autio, T. Smura, and H. Hämmäinen, “Operator Roles in Mobile Broadcast,” Nordic and Baltic J. of Information and Communication Technologies, vol. 2, pp. 48-60, 2008.

[26] T. Rokkas, D. Varoutas, D. Katsianis, T. Smura, R. Kumar, M. Heikkinen, J. Harno, M. Kind, D. Von Hugo, and T. Monath, “On the Economics of Fixed-Mobile Convergence,” Info, vol. 11, 2009, pp. 75-86, doi:10.1108/14636690910954999.

[27] M. V. J. Heikkinen and S. Luukkainen, “Value Analysis of Technology Evolution: Case Mobile Peer-to-Peer Communications,” Proc. Wireless Telecommunications Symposium (WTS 2009), IEEE Press, in press.

[28] J. Schäfer, K. Malinka, and P. Hanáček, “Peer-to-Peer Networks: Security Analysis,” Intl. J. on Advances in Security, vol. 2, pp. 53-61, 2009.

[29] J. Suomalainen, A. Pehrsson, and J. K. Nurminen, “A Secure P2P Incentive Mechanism for Mobile Devices,” Intl. J. on Advances in Security, vol. 2, pp. 42-52, 2009.

[30] P. Merz, F. Kolter, and M. Priebe, “A Distributed Reputation System for Super-Peer Desktop Grids,” Intl. J. on Advances in Security, vol. 2, pp. 30-41, 2009.

[31] H. Koshutanski, M. Ion, and L. Telesca, “Towards User-Centric Identity Interoperability for Digital Ecosystems,” Intl. J. on Advances in Security, vol. 1, pp. 26-38, 2008.

[32] T. Buchanan, C. Paine, A. N. Joinson, and U-D. Reips, ”Development of Measures of Online Privacy Concern and Protection for Use on the Internet,” J. of the American Society for Information Science and Technology, vol. 58, Nov. 2006, pp. 157-165, doi: 10.1002/asi.20459.

[33] H. Wang, M. K. O. Lee, and C. Wang, “Consumer Privacy Concerns about Internet Marketing,” Commun. ACM, vol. 41, Mar. 1998, pp. 63-70, doi:10.1145/272287.272299.

[34] A. Kobsa, “Privacy-Enhanced Personalization,” Commun. ACM, vol. 50, Aug. 2007, pp. 24-33, doi:10.1145/1278201.1278202.

[35] Y. Lu, W. Wang, B. Bhargava, and D. Xu, “Trust-Based Privacy Preservation for Peer-to-Peer Data Sharing,” IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, vol. 36, May 2006, pp. 498-502, doi:10.1109/TSMCA.2006.871795.

[36] D. L. Hoffman, T. P. Novak, and M. Peralta, “Building Consumer Trust Online,” Commun. ACM, vol. 42, Apr. 1999, pp. 80-85, doi:10.1145/299157.299175.

[37] J. W. Palmer, J. P. Bailey, and S. Faraj, “The Role of Intermediaries in the Development of Trust on the WWW: The Use and Prominence of Trusted Third Parties and Privacy Statements,” J. of Computer-Mediated Communication, vol. 5, Jun. 2006, doi:10.1111/j.1083-6101.2000.tb00342.x.

[38] I. Clarke, S. G. Miller, T. W. Hong, O. Sandberg, and B. Wiley, “Protecting Free Expression Online with Freenet,” IEEE Internet Computing, vol. 6, Jan 2002, pp. 40-49, doi:10.1109/4236.978368.

[39] S. Lipner, “The Trustworthy Computing Security Development Lifecycle,” Proc. 20th Annual Computer Security Applications Conf. (ACSAC 04), IEEE Press, Dec. 2004, pp. 2-13, doi:10.1109/CSAC.2004.41.

[40] B. Schneier, “Attack trees,” Dr. Dobb’s J., Dec. 1999.

[41] Statistics Finland, “Population Projection,” May 2007.

[42] Ministry of Transport and Communications Finland, “National Broadband Strategy: Final Report,” Jan. 2007.

[43] IFPI, “Digital Music Report,” Jan. 2008.

[44] Jupiter Research, “US Digital Music Forecast,” Jan. 2007.

[45] W. B. Norton, “Video Internet: The Next Wave of Massive Disruption to the U.S. Peering Ecosystem (v1.5),” Apr. 2007.

[46] J. Alleman, G. Madden, and H. Kim, “Real Options Methodology Applied to the ICT Sector: A Survey,” Communications & Strategies, no. 70, pp. 27-44, 2008.

[47] M. de Reuver, H. Bouwman, and T. Haaker, “Mobile Business Models: Organizational and Financial Design Issues that Matter,” Electronic Markets, vol. 19, Mar. 2009, pp. 3-13, doi:10.1007/s12525-009-0004-4.

[48] M. de Reuver, H. Bouwman, and I. MacInnes, “Business Models Dynamics for Start-Ups and Innovating E-Businesses,” Intl. J. Electronic Business, vol. 7, pp. 269-286, 2009.

A Holistic Approach to Open Source VoIP Security: Results from the EUX2010SEC Project

Lothar Fritsch, Arne-Kristian Groven, Lars Strand, Wolfgang Leister, and Anders Moen Hagalisletto Norsk Regnesentral
Oslo, Norway
email: {lothar.fritsch, groven, lars.strand, wolfgang.leister, anders.moen}@nr.no

Abstract— **The present paper describes the approach and preliminary results from the research project EUX2010SEC. The project works closely with Voice over IP (VoIP) companies and users. The project aims at providing better security of open source VoIP installations. The work towards this goal is organized by gathering researchers and practitioners around scientific activities that range from security modeling and verification up to testbed testing. The expected outcomes of the project are a solid scientific and practical understanding of the security options for setting up VoIP infrastructures, particular guidance on secure, typical setups of such infrastructure. The project's special focus is on producing results relevant to the practitioners in the project, aiming at the stimulation of innovation and the provision of highest quality in open source based VoIP products and services. The article describes the research-based innovation approach used.**

Index Terms— **VoIP, SIP, security model, security requirements, testbed testing, formal protocol analysis.**

I. Introduction

This article provides overview of the VoIP security research project EUX2010SEC[1] which has its roots in the Nordic resource network *Enterprise Unified Exchange (EUX2010).* The project is partly funded by the Norwegian Research Council, and runs from 2007 until 2011. The project provides a forum for researchers[2], user representatives from Norwegian public administration[3], and small and medium sized companies representing both the VoIP and open source software industry in Norway[4]. The current work is based on a conference article on the International Conference on Networking in 2009 [1].

A. Research-based innovation in Norway

The EUX2010SEC project is placed in Norwegian Research Council's technological programme "Kjernekompetanse og verdiskapning i IKT" (VERDIKT), a public funding scheme for user-driven, research-based innovation which targets Norwegian industry and research institutions. The principal tool in the VERDIKT programme is the user-driven project.

The EUX2010SEC project aims at the analysis and development of open source technologies used in VoIP infrastructures. As means towards the goal we implemented a testbed laboratory for the industrial users, and applied user-need based research and problem-solving activities for the VoIP stakeholders in the project. The outcomes shall widen understanding of VoIP, promote secure infrastructures, and strengthen the competitiveness of the Norwegian industry partners in the project. It uses the *Empathic Design* [2] approach and rapid prototyping strategies among other innovation strategies. In addition to industry research work and publication, the project educates a PhD student in the field. Thus EUX2010SEC uses the three most successful industry-oriented innovation strategies considered by MIT researchers [3].

B. Project goals

The overall research goal of the project is to improve the level of security and awareness when developing, installing, and using open source VoIP solutions, such as the open source Asterisk PBX[5]. The main objectives of VoIP-oriented security are to preserve the availability of VoIP services, to protect VoIP transmissions and stored information from disclosure and theft, to prevent fraudulent usage of voice communication, so called toll fraud with financial losses, and to preserve the integrity of the VoIP system, e.g., that the system logs to be stored by the providers on behalf of the authorities are correct.[6]

As one of the fastest growing Internet technologies today, Voice over IP (VoIP) can provide a number of additional services compared to traditional telephony. These services include conferencing, events notification, presence, instant messaging, video telephony and other multimedia transmissions, and location independence (location mobility). Such wide flexibility imposes challenges on how security is handled [4], [5].

Our experience from work with the industry partners is that in many cases the security model applied to VoIP networks is a model of isolation, physically separating voice and data or using virtual LANs or VPNs to separate VoIP traffic from any other IP traffic. This separation sacrifices many of the benefits of VoIP and makes the integration of communication

This paper is based on the conference article "A holistic approach to Open Source VoIP security: Results from the EUX2010SEC project", presented at the ICN 2009 conference.

[1] Project homepage: http://eux2010sec.nr.no

[2] Norwegian Computing Center, UNU-MERIT

[3] Buskerud County Municipality (Buskerud fylkeskommune).

[4] Redpill Linpro AS, Freecode AS, Nimra Norge AS, Ibidium Norden AS

[5] Asterisk is a central component in the VoIP networks we are interested in. Asterisk homepage: http://www.asterisk.org

[6] In many countries the telephony providers must store the connection logs of for a specified time, typically several months.

applications hard or even impossible. Hence, the potential of VoIP systems is often not utilized. One goal of the project is to look into other possible VoIP network topologies and approaches to security. This would enable the adoption of innovative functions, such as mobile software phones on laptops and PDAs being used on open Public IP networks, much easier.

When analyzing VoIP security and vulnerability different perspectives are used in the project:

- Analysis at device level, focusing on a particular device, e.g., a PBX (Private Branch Exchange);
- Analysis at system level, focusing on the VoIP infrastructure components and VoIP topologies, or;
- Analysis focusing on the flow of data and signals in VoIP systems.

Vulnerabilities in VoIP have many causes [6] which may be related to weaknesses in the applied protocols, the software, or the configurations of the various VoIP applications and equipment in use. EUX2010SEC provides analysis, testing and guidance of many possible options to the suppliers and users of VoIP services, and in addition researches the security consequences.

The EUX2010SEC project aims at transferring innovation to the market by supporting the practitioners with scientific security knowledge. This knowledge is provided by analysis of topologies and usage patterns of VoIP systems; analysis of the systems using both formal methods and testbed testing; the collection of realistic security requirements from practitioners and users; and the development and testing of secure configurations, which will be recommended as base configurations for various basic VoIP setups.

C. State of knowledge

This section provides an overview of general VoIP security literature. The following sections on verification, testing and security modeling might introduce more specialized background references where needed.

Security of VoIP systems has received much attention in national security bodies and in academia. Analysis focused on technical security issues, and availability considerations of VoIP-based critical communications services. The VOIPSA taxonomy is our starting point for a systematic exploration of known VoIP security threats [6]. The VOIPSA taxonomy is less detailed in the description of problems and fixes, but it is superior in its taxonomic description over many of the hands-on guidebooks such as or [4]. Various governments information security institutions or standards institutes have issued warnings or guidance, for example the U.S-National Institute of Standards and Technology [7] and others [8].

Some scientific publications overlook the topic, but mainly discovered classic attack patterns such as man-in-the-middle attacks, the exploitation of misconfigurations, and reachability control issues [9], [10]. Some work has been done to analyze security vulnerabilities in VoIP implemented technologies [11]. Among others, the SIP protocol [12], [13] has the important role of connection establishment and management. SIP is vulnerable to authentication and hijacking problems [14], and others [15], [16].

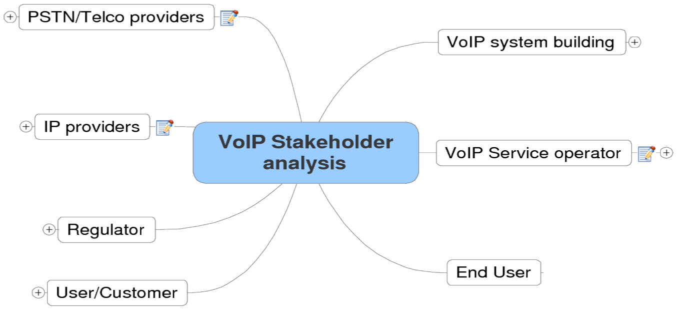


Fig. 2. VoIP Stakeholder analysis

II. Methodology and approach

The research activities in EUX2010SEC focus on three areas of activity, as shown in Fig. 1

The ***security model*** activity analyzes stakeholders' requirements towards security and stability of VoIP systems. Its goal is to derive typical requirements' profiles, and to provide security models and default configurations for them. This is shown in the right part of Fig. 1.

Testbed systems with the partners' technology, and real user requirements: These testbed systems will have VoIP traffic routed through them for testing the system properties and the consequences of configuration options. They will additionally be used for the deployment of a set of attacks and attack tools. The testbed activity is depicted in the middle part of Fig. 1.

Formal protocol analysis*:* The function, usage and real configuration and implementation of security-relevant protocols used in the Asterisk family of VoIP systems is formalized and then tested with a protocol verification tool that attacks the protocol model. This approach can reveal unknown protocol failures, and wrongful implementation of protocols. The formal analysis approach is shown in the left part of Fig. 1.

A. Requirements & security model

The stakeholder and requirements gathering approach is inspired by the privacy design process outlined in [17], and was used in [18]. It is modified in EUX2010SEC to find and elaborate VoIP security requirements for the identified basic scenarios of VoIP usage.

The security model activity is carried out in consecutive steps. A basic stakeholder model and initial scenario profiles is derived from the state of the art literature. Various VoIP project partners and possibly their customers are contacted for empiric research. Steps to be carried out are as follows:

Stakeholder Analysis: The stakeholders are identified and contacted, and their main interests in the VoIP market be captured by means of a stakeholder analysis [19].

Requirements Elicitation: The stakeholders are interviewed concerning their usage scenarios and requirements concerning VoIP security.

- The interviews collect anecdotic accounts of problems and requirements.
- The interviewees are presented with scenarios and use cases to single out their typical scenarios.

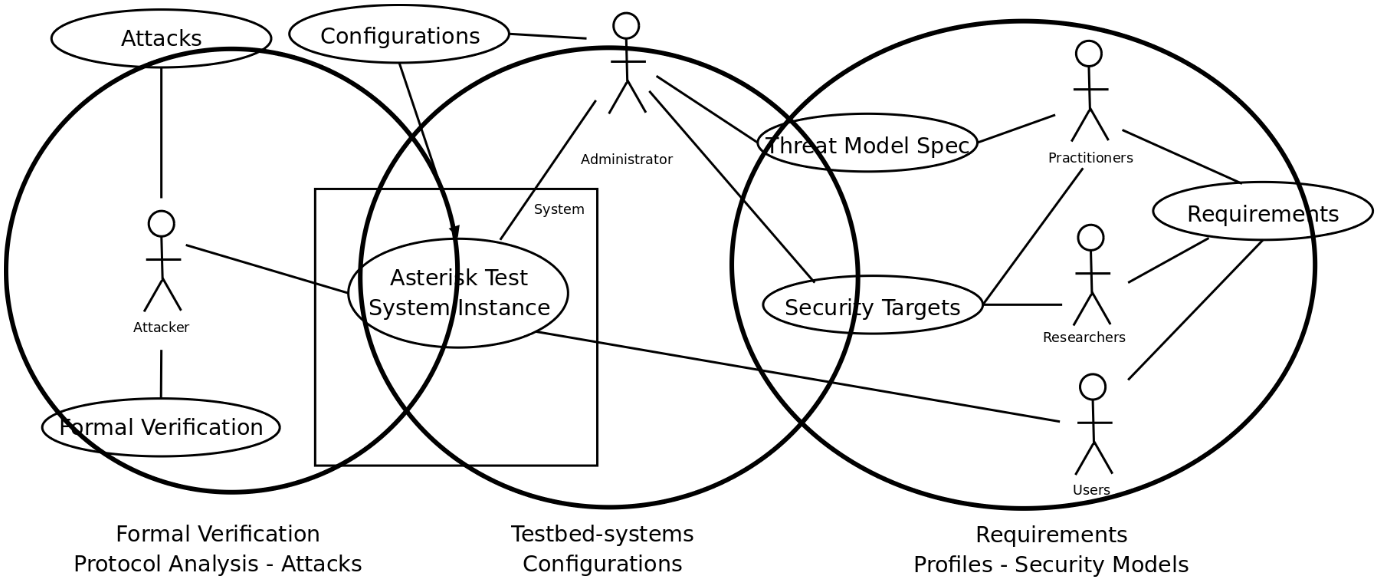


Fig. 1. EUX2010SEC research approach

Scenario Profiles: From the steps above, one or more profiles for typical VoIP usage scenarios will be generated. The profiles should create the basis for further analysis, testbed creation, and verification activities.

- A profile is based on a use case description.
- A profile contains a description of security, reliability, quality-of-service and scalability needs.

Multilateral Security Analysis: For each of the profiles, a multilateral security analysis is performed [20] to ensure that all stakeholders' views and needs are contained. Its goal is to gather security and privacy requirements for the infrastructure in question, and to make suggestions for improvement of the requirements specification. Multilateral security analysis takes into account all stakeholders' requirements relevant to security and privacy issues.

Security Models: Finally, security models are developed for the VoIP profiles. A security model is based on security goals, and a trust model. It contains a description of:

- Subjects
- Objects
- Rules and policies
- Security functions

It is hard to retrieve stable, unified requirements from interviews with stakeholders. Therefore, it is necessary to have several cycles of interaction with the stakeholders to verify the requirements, profiles and models. Our approach to this problem is similar to rapid prototyping in software development: a fast, parallel development of requirements, to be presented and discussed with the stakeholders in several loops of interaction, such as *Maieutik* [21] and *Empathic Design* [2].

B. Configurations testbed and attacking

For testing VoIP configurations and security profiles from our project partners we have developed a dedicated VoIP testbed. Testbeds as a research approach enable us to do prospective analysis of VoIP technology and to effectively gain knowledge about VoIP capabilities, limitations and benefits in different conditions [22]. This provides us with an advantage over a theoretical approach alone, since VoIP is tested in different contexts. The testbed is used as a controlled environment using strict configuration management to ensure scientific measurements. Specifically, we test various VoIP installations, where we launch predefined, reproducible attacks to uncover security vulnerabilities.

Real life VoIP has many deciding factors that have an impact on performance and security, such as the network topology, network congestion, and the protocols used. A theoretical approach alone cannot be employed to consider all these factors because of their complex relationships. Simulation is often used to study computer networks, since it offers a convenient combination of flexibility and controllability. The disadvantage of using simulations is that results may not be applicable to the reality, since often an inappropriate level of abstraction has been applied. The testbed creates an environment where the project researchers can experiment with different VoIP configurations in a low-risk environment, prior to real-world testing and deployment.

We pursue the following goals with the VoIP testbed:

(1) Given VoIP configurations are validated in the testbed against security requirements resulting from the previous analysis steps outlined above in Section II-A. Specifically, the experiments in the testbed shall show conformance between a given VoIP installation, configuration or architecture, and specified security requirements defined by the stakeholders.
While the testbed can be used in various ways, our work hypothesis is as follows: VoIP-specific security mechanism are deployed and tested to see if they are in accordance with the stakeholder's security policy. In this environment, the deployment of attacks will be launched to uncover

potential vulnerabilities. Data gathered from these tests will be used as input to formal modeling and verification, as outlined below in Section II-C.

(2) We use an automated VoIP testbed attack tool to scan a given VoIP installation for known vulnerabilities according to the threat model, and to launch VoIP related attacks.

(3) To be able to re-use a given testbed configuration as a reference configuration management is an important aspect of testbed testing. Especially the handling of a wide range of configuration files is considered as a challenge.

(4) Using the results from the tests we create VoIP configurations that are arguable more secure, based on our findings in the preceding three goals. These configurations, along with recommended best practices, are then presented to the stakeholders for discussion and further refinement.

Various VoIP configurations containing Asterisk PBXs as one of the components are used as target test systems in this testbed. These configurations are copies of real systems deployed in different organizations. When performing tests tests real traffic data are provided by mirroring data traffic into the testbed.

C. Formal analysis of protocols

Formal protocol analysis is an important part, in addition to extensive security testing of real-world VoIP systems and traffic in the project's experimental testbed. We perform formal protocol analysis in combination with experiments in the testbed using the following methodological approach:

- Real-world production systems are installed and configured in the testbed.
- Network traffic from the testbed is recorded/logged (at a certain level of detail).
- Based on the logged network traffic and additional information, like RFCs, formal specifications are constructed.
- These specifications are further analyzed in a formal analysis tool, capable of identifying potential attacks and vulnerabilities affecting system security.
- In order to validate the results from the formal protocol analysis, attempts are made to reconstruct in the testbed on real-world systems the error conditions found in the formal analysis.

In Fig. 3 the work approach and data flow of our formal protocol analysis is illustrated in more detail. So far, the formal protocol analysis has been looking into the properties of SIP [12], [13]. SIP is used for signaling and is working together with other protocols that take care of the media stream, using, e.g., RTP (Real-time transfer protocol) [23]. SIP is a text-based protocol that needs to be strengthened to enhance security. We have been looking into SIP with digest authentication when analyzing SIP-based traffic [14].

In order to gain initial knowledge of the behavior of the SIP implementation of Asterisk, traffic is recorded from real phone sessions going through an Asterisk server. This is done by using VoIP-targeted IP network monitoring and interception tools such as *Wireshark*[7]. The traces of sessions produced by

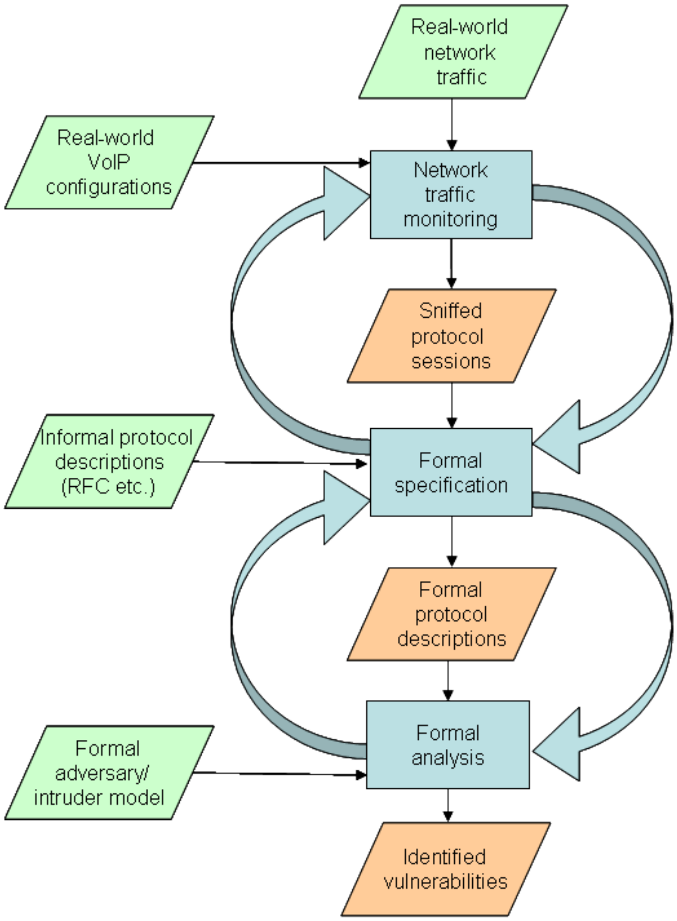

Real-world network traffic
Real-world VoIP configurations
Network traffic monitoring
Sniffed protocol sessions
Informal protocol descriptions (RFC etc.)
Formal specification
Formal protocol descriptions
Formal adversary/ intruder model
Formal analysis
Identified vulnerabilities

Fig. 3. Formal analysis of VoIP systems

Wireshark can be presented both textually and as interaction diagrams, at various levels of detail.

Based on the output from *Wireshark*, formal models/formal specifications are then produced. In this process, the SIP RFC specifications are used as additional guidelines and references. This transformation from the traces of SIP-sessions to formal specifications of the same sessions requires manual intervention, and several rounds of quality assurance.

Having produced the formal models from the SIP traces these are further analyzed in a formal protocol analyzer. We used a locally developed experimental tool for formal protocol analysis, PROSA [24], in the analysis so far. PROSA is based on temporal epistemic logic, and includes a module for automated refinement and validation of protocols.

PROSA is not the sole alternative, and different types of formal protocol analysis tools and methods are today available, of which some are listed below: Process calculus with probability and complexity [25], symbolic execution models/multiset rewriting [26], Protocol logics like BAN logic [27], model checking, either symbolic analysis like strand spaces or (exhaustive) finite state analysis like Murphi [28] or CASPER/FDR [29], and finally search using symbolic representation of states, e.g., the NRL analyzer [29]. Tools similar to PROSA include OFMC [30] and Scyther [31].

[7] Wireshark web page: *www.wireshark.com*

The purpose of formal protocol analysis is to look for non-intuitive attacks, omissions in specifications, or errors in different products implementation of protocols. During the analysis of VoIP protocols, networks are assumed to be hostile, in that they may contain intruders that can read, modify, or delete traffic, and that may have control of one or more network principals. Many of these attacks do not depend upon flaws or weaknesses in the underlying cryptographic algorithm, but can be exploited by an attacker. The results of the formal protocol analysis are validated in the testbed.

PROSA is a tool developed for the specification, static analysis and simulation of security protocols. PROSA consists of three main modules: (a) a specification language based on temporal epistemic logic; (b) a static analysis module; and (c) a simulator for executing intended protocols and attacks on protocols.

The language in the PROSA tool contains constructs for specification and reasoning about message transmission, cryptographic operations, and agent beliefs. Below is listed an excerpts of the PROSA language to be used later in this article Here $\mathscr{L}_P$ is the smallest language such that:

(i) Each of the following atomic formulas are in $\mathscr{L}_P$

ε	the empty sentence
$a = b$	equality
$\mathsf{Agent}(a)$	a is an agent
$\mathsf{isKey}(k)$	k is a key
$\mathsf{isNonce}(\mathsf{n}(N, a))$	$\mathsf{n}(N, a)$ is a nonce
$\mathsf{playRole}(a, x, \mu)$	a plays the x-role in protocol μ
$\mathsf{role}(a)$	a is a role in a protocol

(ii) If φ, ψ, $\xi^{\mathcal{T}}$, $\xi^{\mathcal{A}}$, $\xi^{\mathcal{S}}$ $\in \mathscr{L}_P$, then so are;

$\neg\varphi$, $\varphi \rightarrow \psi$	propositional logic
$a \longrightarrow b : \varphi$	a sends the message φ to b
$\mathsf{Bel}_a(\varphi)$	a believes φ
$\mathsf{Hash}[\varphi]$	hash φ
$\mathsf{Enforce}_{t^A}(\varphi)$	enforce agent t^A to do φ
$\mathsf{protocol}[\mu, N, \xi^{\mathcal{T}}, \xi^{\mathcal{A}}, \xi^{\mathcal{S}}, \Phi]$	protocol operator

In addition there are constructs for, e.g., time stamps, quantifiers, encrypt, decrypt, and constructs that explain succession.

The static analysis module consists of algorithms for *automated refinement* of both protocol specifications and attack descriptions. The automated refinement results in an explicit specification that contains assumptions local to each agent participating, i.e. pre- and postconditions, for each transmission clause. Refined specifications can then be *validated.*

The validation process of a trace specification is performed in two steps in PROSA: First, a tool-supported refinement of the specification is generated. This will give a specification that contains information about the agents beliefs and construction of credentials, like the generation of nonces, timestamps, assumptions about keys, and cryptographic operation like encryption, decryption and hashing. Secondly, the refined specification is validated to check whether a participant in the protocol setting possesses any beliefs that have not been legally obtained through communication or cryptography.

The PROSA language is defined to be close to practical protocol specification and design, understandable for both software developers as well as system architects. The same language is also the metalanguage for reasoning about the protocol specification. In this way it differs from state-the-art tools like OFMC [30] and Scyther [31]. Here a specification is written in one language that is later preprocessed to an intermediate language serving as input to the reasoning tools.

The PROSA language has similarities with, e.g., BAN logic, yet there are some significant differences. The meaning of the belief operator is defined by the detailed definition of the protocol machine, which is a central part of the operational semantics of PROSA. Hence the belief operator is interpreted as part of the execution of protocols. Contrary to a purely logical explanation of abstract security properties and mechanisms, the belief construct is given a concrete operational meaning. Belief means possession, there is no other operator for reasoning about beliefs and data-content. Other logics, e.g., BAN logic, have several operators.

Although beliefs in PROSA are rather complex, in the way they are explained by many rules in the operational semantics, it is still possible to have a rather standard logical understanding of beliefs.

III. Results and progress

In this section, we summarize the results and the progress so far with an emphasis on the areas of *formal protocol analysis*, *security modeling*, and *laboratory security testing*. Since the project will continue to work into 2011 we expect more results during its course.

A. Formal protocol analysis

The SIP protocol specification, as described in RFC 3261 [13], is implemented differently in the various VoIP systems. We explored how Asterisk implements the SIP protocol by using formal protocol analysis. Real-world Asterisk configurations originating from an industrial partner were used as basis for our analysis.

Traffic was then monitored and recorded as a basis for the formal analysis, hence capturing the specifics of how SIP is implemented in Asterisk. The fact that Asterisk is implementing a B2BUA (a back to back user agent) instead of a SIP proxy became clear to us during the analysis.

Transforming a representation of a session from network traffic monitoring tool trace to a formal model in standard notation requires manual intervention. We identified the need of a tool that is able to export the traces representing real data traffic from *tcpdump* or *Wireshark* into a formal specification readable for the protocol analysis tool. Until such a tool is developed the transformation must be performed carefully in order to avoid errors in that process.

The PROSA syntax is using a standard Alice-Bob notation, [32] and standard notations for describing security protocols [33]. Hence the PROSA formulas presented in this section should be readable to those familiar with the above mentioned notations. We explain a few constructs using Fig. 4 as an example: The header of a protocol specification consists of a protocol name, then a session number – since there might be several instances – followed by specification of all roles, the

$$\begin{aligned}
&\mathsf{protocol}\mathrm{SIP}, 0,\\
&\quad \mathsf{role}(A) \wedge \mathsf{role}(S) \wedge \mathsf{role}(B),\\
&\quad \mathsf{role}(A) \wedge \mathsf{role}(S) \wedge \mathsf{role}(B),\\
&\quad \mathsf{role}(A),
\end{aligned}$$

$$\begin{aligned}
\mathsf{Enforce}_A(\mathsf{Bel}_A(&\mathrm{startProtocol}(\mathrm{SIP-CANCEL},\\
&\mathsf{playRole}(B, C, \mathrm{SIP-CANCEL}) \wedge\\
&\mathsf{playRole}(S, T, \mathrm{SIP-CANCEL}) \wedge\\
&\mathsf{playRole}(A, D, \mathrm{SIP-CANCEL}),\\
&\mathsf{Text}(\text{Refer to session }))))
\end{aligned}$$

$$\begin{aligned}
A \longrightarrow S :\ & \mathsf{Text}(\text{INVITE}) \wedge \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge\\
& \mathsf{Text}(\text{Contact}, A) \wedge \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
S \longrightarrow A :\ & \mathsf{Text}(\text{Proxy Authentication Required}) \wedge \mathsf{Text}(\text{Username}, A) \wedge\\
& \mathsf{Text}(\text{Realm}) \wedge \mathsf{isNonce}(\mathsf{n}(DIGESTCHALLENGE, S)) \wedge\\
& \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
A \longrightarrow S :\ & \mathsf{Text}(\text{ACK}) \wedge \mathsf{Agent}(B) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
A \longrightarrow S :\ & \mathsf{Text}(\text{INVITE}) \wedge \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge\\
& \mathsf{Text}(\text{Contact}, A) \wedge \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A)) \wedge\\
& \mathsf{Hash}[\mathsf{Hash}[\mathsf{Text}(\text{Username}, A) \wedge \mathsf{Text}(\text{Realm}) \wedge \mathsf{isKey}(\mathsf{key}(\mathsf{s}, A, S))] \wedge\\
& \qquad \mathsf{isNonce}(\mathsf{n}(DIGESTRESPONSE, A)) \wedge\\
& \qquad \mathsf{isNonce}(\mathsf{n}(DIGESTCHALLENGE, S)) \wedge\\
& \qquad \mathsf{Hash}[\mathsf{Text}(\text{INVITE}) \wedge \mathsf{Text}(\text{URI}, B)]]
\end{aligned}$$

$$\begin{aligned}
S \longrightarrow A :\ & \mathsf{Text}(\text{100 TRYING}) \wedge \mathsf{Text}(\text{Contact}, B) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
S \longrightarrow B :\ & \mathsf{Text}(\text{INVITE}) \wedge \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge\\
& \mathsf{Text}(\text{Contact}, A) \wedge \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
B \longrightarrow S :\ & \mathsf{Text}(\text{100 TRYING}) \wedge \mathsf{Text}(\text{Contact}, B) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
B \longrightarrow S :\ & \mathsf{Text}(\text{180 RINGING}) \wedge \mathsf{Text}(\text{Contact}, B) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
S \longrightarrow A :\ & \mathsf{Text}(\text{180 RINGING}) \wedge \mathsf{Text}(\text{Contact}, B) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$B \longrightarrow S : \mathsf{Text}(\text{200 OK})$$

$$S \longrightarrow A : \mathsf{Text}(\text{200 OK})$$

$$\begin{aligned}
A \longrightarrow S :\ & \mathsf{Text}(\text{ACK}) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
S \longrightarrow B :\ & \mathsf{Text}(\text{ACK}) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
A \longrightarrow B :\ & \mathrm{start}(\mathrm{MediaTrans},\\
& \mathsf{playRole}(\mathrm{Alice}, A, \mathrm{MediaTrans}) \wedge\\
& \mathsf{playRole}(\mathrm{Bob}, B, \mathrm{MediaTrans}))
\end{aligned}$$

Fig. 4. Specification of the SIP call setup sub-protocol

$$\begin{aligned}
&\mathsf{protocol}[\mathrm{SIPAttack}, 0,\\
&\quad \mathsf{role}(A) \wedge \mathsf{role}(S) \wedge \mathsf{role}(B) \wedge \mathsf{role}(I) \wedge \mathsf{role}(F),\\
&\quad \mathsf{role}(A) \wedge \mathsf{role}(S) \wedge \mathsf{role}(B) \wedge \mathsf{role}(I) \wedge \mathsf{role}(F),\\
&\quad \mathsf{role}(A),
\end{aligned}$$

$$\begin{aligned}
A \longrightarrow I(S) :\ & \mathsf{Text}(\text{INVITE}) \wedge \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
I(A) \longrightarrow S :\ & \mathsf{Text}(\text{INVITE}) \wedge \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge\\
& \mathsf{Text}(\text{Contact}, A) \wedge \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
S \longrightarrow I(A) :\ & \mathsf{Text}(\text{Proxy Authentication Required}) \wedge \mathsf{Text}(\text{Username}, A) \wedge\\
& \mathsf{Text}(\text{Realm}) \wedge \mathsf{isNonce}(\mathsf{n}(DIGESTCHALLENGE, S)) \wedge\\
& \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
I(S) \longrightarrow A :\ & \mathsf{Text}(\text{Proxy Authentication Required}) \wedge \mathsf{Text}(\text{Username}, A) \wedge\\
& \mathsf{Text}(\text{Realm}) \wedge \mathsf{isNonce}(\mathsf{n}(DIGESTCHALLENGE, S)) \wedge\\
& \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
A \longrightarrow I(S) :\ & \mathsf{Text}(\text{ACK}) \wedge \mathsf{Agent}(B) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
I(A) \longrightarrow S :\ & \mathsf{Text}(\text{ACK}) \wedge \mathsf{Agent}(B) \wedge \mathsf{Text}(\text{Contact}, A)\\
& \wedge\ \mathsf{Text}\ (\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
A \longrightarrow I(S) :\ & \mathsf{Text}(\text{INVITE}) \wedge \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A)) \wedge\\
& \mathsf{Hash}[\mathsf{Hash}[\mathsf{Text}(\text{Username}, A) \wedge \mathsf{Text}(\text{Realm}) \wedge \mathsf{isKey}(\mathsf{key}(\mathsf{s}, A, S))] \wedge\\
& \mathsf{isNonce}(\mathsf{n}(DIGESTRESPONSE, A)) \wedge\\
& \mathsf{isNonce}(\mathsf{n}(DIGESTCHALLENGE, S)) \wedge\\
& \mathsf{Hash}[\mathsf{Text}(\text{INVITE}) \wedge \mathsf{Text}(\text{URI}, B)]]
\end{aligned}$$

$$\begin{aligned}
I(A) \longrightarrow S :\ & \mathsf{Text}(\text{INVITE}) \wedge \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A)) \wedge\\
& \mathsf{Hash}[\mathsf{Hash}[\mathsf{Text}(\text{Username}, A) \wedge \mathsf{Text}(\text{Realm}) \wedge \mathsf{isKey}(\mathsf{key}(\mathsf{s}, A, S))] \wedge\\
& \mathsf{isNonce}(\mathsf{n}(DIGESTRESPONSE, A)) \wedge\\
& \mathsf{isNonce}(\mathsf{n}(DIGESTCHALLENGE, S)) \wedge\\
& \mathsf{Hash}[\mathsf{Text}(\text{INVITE}) \wedge \mathsf{Text}(\text{URI}, B)]]
\end{aligned}$$

$$I(S) \longrightarrow A : \mathsf{Text}(\text{CANCEL}) \wedge \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))$$

$$\begin{aligned}
A \longrightarrow I(S) :\ & \mathsf{Text}(\text{487 Request Terminated}) \wedge \mathsf{Text}(\text{URI}, A) \wedge\\
& \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$I(S) \longrightarrow A : \mathsf{Text}(\text{ACK}) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))$$

$$\begin{aligned}
S \longrightarrow I(A) :\ & \mathsf{Text}(\text{100 TRYING}) \wedge \mathsf{Text}(\text{Contact}, B) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
S \longrightarrow I(B) :\ & \mathsf{Text}(\text{INVITE}) \wedge \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
I(S) \longrightarrow B :\ & \mathsf{Text}(\text{INVITE}) \wedge \mathsf{Agent}(A) \wedge \mathsf{Agent}(B) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
B \longrightarrow I(S) :\ & \mathsf{Text}(\text{100 TRYING}) \wedge \mathsf{Text}(\text{Contact}, B) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
I(B) \longrightarrow S :\ & \mathsf{Text}(\text{100 TRYING}) \wedge \mathsf{Text}(\text{Contact}, B) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
B \longrightarrow I(S) :\ & \mathsf{Text}(\text{180 RINGING}) \wedge \mathsf{Text}(\text{Contact}, B) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
I(B) \longrightarrow S :\ & \mathsf{Text}(\text{180 RINGING}) \wedge \mathsf{Text}(\text{Contact}, B) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
I(S) \longrightarrow B :\ & \mathsf{Text}(\text{CANCEL}) \wedge \mathsf{Agent}(A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
B \longrightarrow I(S) :\ & \mathsf{Text}(\text{487 Request Terminated}) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$I(S) \longrightarrow B : \mathsf{Text}(\text{ACK}) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))$$

$$\begin{aligned}
S \longrightarrow I(A) :\ & \mathsf{Text}(\text{180 RINGING}) \wedge \mathsf{Text}(\text{Contact}, B) \wedge\\
& \mathsf{Text}(\text{URI}, B) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$I(B) \longrightarrow S : \mathsf{Text}(\text{200 OK})$$

$$S \longrightarrow I(A) : \mathsf{Text}(\text{200 OK})$$

$$\begin{aligned}
I(A) \longrightarrow S :\ & \mathsf{Text}(\text{ACK}) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLID, A))
\end{aligned}$$

$$\begin{aligned}
S \longrightarrow I(B) :\ & \mathsf{Text}(\text{ACK}) \wedge \mathsf{Text}(\text{Contact}, A) \wedge\\
& \mathsf{Text}(\text{URI}, A) \wedge \mathsf{isNonce}(\mathsf{n}(CALLIDB, S))
\end{aligned}$$

$$\begin{aligned}
\mathsf{Enforce}_I(\mathsf{Bel}_I(&\mathrm{startProtocol}(\mathrm{MediaTransAttack},\\
&\mathsf{playRole}(\mathrm{Malice}, I, \mathrm{MediaTransAttack}) \wedge\\
&\mathsf{playRole}(\mathrm{Frank}, F, \mathrm{MediaTransAttack}), \mathsf{Text}(\text{Reference to callid's}))))
\end{aligned}$$

$$\begin{aligned}
\mathsf{Enforce}_F(\mathsf{Bel}_F(&\mathrm{startProtocol}(\mathrm{MediaTransAttack},\\
&\mathsf{playRole}(\mathrm{Malice}, I, \mathrm{MediaTransAttack}) \wedge\\
&\mathsf{playRole}(\mathrm{Frank}, F, \mathrm{MediaTransAttack}), \mathsf{Text}(\text{Reference to callid's}))))
\end{aligned}$$

Fig. 5. Call hijacking attack on the SIP call setup sub-protocol

agent specific roles, and the start role. The Enforce construct builds instances of tear down subprocesses within each agent making them able to listen for CANCEL messages. In SIP a CANCEL message can appear whenever an agent hangs up the phone, from any state in a call setup process. Following the Enforce statement, in the specification in Fig. 4, 14 transmissions in sequential order are representing the SIP call setup signaling sequence.

In the PROSA tool a static analysis can be performed as follows. The initial protocol specification is automatically, by the tool, augmented with pre- and postconditions expressing beliefs and trust at each stage in the specification. Some statistics taken from the static analysis of the SIP call setup protocol specification is presented in Table I. The length of the protocol indicates the number of statements in the original specification. In our case the Enforce statement is followed

TABLE I
STATISTICS ON THE SUB-PROTOCOLS.

protocol	Length	Refined	Crypto	Validation
...	...	...	...	...
Call Setup	15	88	3	27812
...	...	...	...	...

TABLE II
STATISTICS ON THE SIMULATIONS.

Simulation scenario	PROSA rewrites	Time (milli seconds)
SIP without Digest	18 239	41
SIP Digest simulation	82 987	164
SIP with eavesdropper	83 365	188
Active call-hijacking attack	364 969	472

by 14 transmissions, totalling 15 statements. The length of the automatically refined protocol quantifies the number of statements plus the additional pre-and post conditions. The number of cryptographic operations involved are 3 instances of a hash functions while the last column is a count of the number of rewrites in PROSA tool performed to validate the specification.

After finishing static analysis and validation, the next step is simulation. Our simulation scenario included three components, two calling parties *Alice*, *Bob*, and a proxy server. Each agent runs an instance of the SIP sub-protocols described above. Here, we assume that Alice initiates a phone call with Bob in three variations:

(a) without Digest Access Authentication;
(b) using Digest Access Authentication; and
(c) using Digest Access Authentication, but with an attacker eavesdropping the messages.

A standard digest simulation without an attacker, (b), is augmented with an attacker on the line just forwarding the messages, (c). The number of computation steps required to perform an eavesdropping differs insignificantly from the "good" simulation. This augmenting is automatically done. The results of the PROSA simulations are reported in Table II. The first simulation is without Digest Access Authentication, while the latter three include Digest Access Authentication. The last one is manually derived from (c). This due to the need for the intruder- and adversary model for PROSA to be extended. What takes place in the latter simulation scenario is the following: An attack where an intruder Ivory (denoted I) hijacks a call-setup session and establishes a phone call with another agent Frank (denoted F), as described in Fig. 5.

Initial results of our work indicates potential vulnerabilities in SIP authentication [14] and call-setup [34] that can lead to attacks, based on analysis under the Dolev-Yao attacker/intruder model [35].

B. Security modeling

In the following we show the characterization of VoIP scenarios. Six different scenario patterns were visualized graphically in a metaphor as islands. These depict different VoIP basic setups as shown in Fig. 3: *Island, Archipelagos, Nomadic Islanders, Nomadic Libertarians, Fortress, Maginot Line*. These have been verified in a pre-study with selected stakeholders in the project. These profiles are used as a basis for classification of VoIP setups, and will be the basis for the development of security models. The first round of stakeholder interviews was performed in 2008 and early 2009. Through our industry connections, we got access to one VoIP system vendor, and five VoIP system operators which include universities, public administrations, and service providers.

TABLE III
INTERVIEWEES AND THEIR ROLES

Stakeholder	Role
1	VoIP service provider / system vendor
2	municipality
3	university
4	municipality
5	county administration
6	VoIP service provider / energy provider

We observed that most of the stakeholders were acting in more than one role. The vendor offered both system-building and service operation. The service operators originated either from public administration, such as municipalities and counties, or power companies. Both forms of operators own rights to operate telecommunication cable.

The interviews focused on the business model, the customer and user profiles, and security needs and incidents. The interviews were performed as conversations with moderated discussion, where the topics were raised, discussed along the contributions of the interviewees, and terminated with a list of questions from the interviewers. The interviews aimed at classifying the interviewees into the island metaphors, at learning the security requirements and conceptions and the realities. The island metaphors were introduced early to enable an abstraction away from particular details of the telecommunications infrastructure or security technology, as the interviewees mostly had a background in telecommunication technology or network administration. The interviews were following an outline made for each stakeholder category. An example for the outline is shown in Fig. 6.

Concerning their business models, all interviewees shown in Table III provide VoIP-based telephony to their customers. While Stakeholder 1 operates on the open telecommunications market, Stakeholder 6 targets consumers along the power network they operate. Stakeholder 3 is a large university, where VoIP is currently built up to replace PSTN in the offices and laboratories. Generally, the municipal or county organizations seek to replace their own phone infrastructure with an Internet-based infrastructure motivated by cost of ownership. As a side effect, many organizations begin to include users outside the public administration offices, such as schools or medical service centers that are under their governance.

The major reason for choosing VoIP – and in particular Asterisk-based solutions – was the favorable costs of Asterisk-based telecommunications infrastructures. Many of the interviewees were operating old telephony switches, and were facing high maintenance cost and expensive offers for

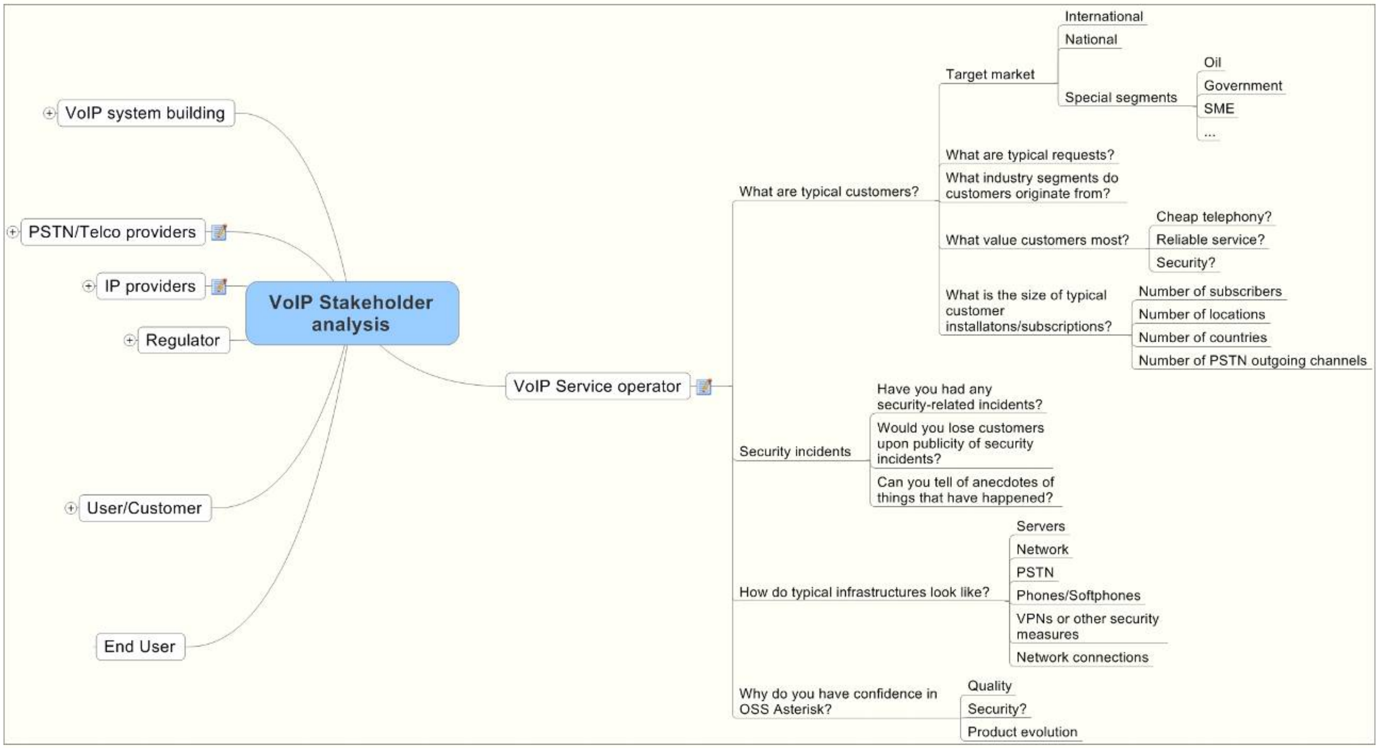

VoIP system building
PSTN/Telco providers
IP providers
Regulator
VoIP Stakeholder analysis
User/Customer
End User
VoIP Service operator
What are typical customers?
Target market
International
National
Special segments
Oil
Government
SME
...
What are typical requests?
What industry segments do customers originate from?
What value customers most?
Cheap telephony?
Reliable service?
Security?
What is the size of typical customer installatons/subscriptions?
Number of subscribers
Number of locations
Number of countries
Number of PSTN outgoing channels
Security incidents
Have you had any security-related incidents?
Would you lose customers upon publicity of security incidents?
Can you tell of anecdotes of things that have happened?
How do typical infrastructures look like?
Servers
Network
PSTN
Phones/Softphones
VPNs or other security measures
Network connections
Why do you have confidence in OSS Asterisk?
Quality
Security?
Product evolution

Fig. 6. Stakeholder interview outline for "VoIP system operator"

replacement of their PSTN switches. At the same time, they had already built up their own IP infrastructure. For most of the customers the seamless replacement of the ordinary desk or cordless phones with the same functionality was in focus. Only one of the stakeholders is actually deploying softphones on laptops for a particular user segment – school teachers who share offices that do not have personally assigned phones. In summary, most of the stakeholders' activities were targeted at migrating the switch-based phone functionality to VoIP.

The typical infrastructure is composed of one or more Asterisk servers, one or more PSTN trunks, and many pre-configured desktop VoIP phones for the end users.

Security concepts go along the lines of dedicated data connections, special routing or VPN tunneling. Probed for security measures and threat scenarios, the interviewees mainly responded that they were shielding their cable, or using dedicated IP addressing, MAC verification and on occasion VPN routers to "keep the VoIP traffic in its own network". This, in addition to the user-side need for the "old" telephony network, reinforces the insight that VoIP is built and used as if it was the PSTN. Asked for security incidents, the stakeholders reported a few billing fraud incidents, mainly based on successful ID theft based subscriber sign-ups. Some mentioned cost induced with 0900 service usage by their legitimate telephony users. The largest worries concerning security have been stated around the topic of identity fraud, fraudulent service usage, and losses due to fraudulent outgoing calls into a billed long-distance network – problems that pre-existed the times of VoIP. For some stakeholders availability of service, in particular of emergency calling, was an issue. None of the stakeholders mentioned IP-based attacks, session hijacking, break-ins into voice mail systems, SPIT calling or eavesdropping problems. There was a considerably low enthusiasm to discuss regulatory issues such as police wiretapping, data retention and crime investigation issues.

Some stakeholders, in particular the system builder, agreed that the complexity of configuration options in Asterisk and the related protocols and the options in the infrastructure is too high. Configuration errors are believed to provide greatly to potentials for unavailability of service or security problems.

Further interviewing and infrastructure inspection in EUX2010SEC will reveal whether some of the existing security threats on the Internet are known to the stakeholders, and help in the development of security concepts for VoIP infrastructures.

C. Laboratory security testing

We work in close interaction with the industry partners participating in the project on how to set up, use, and test different VoIP configurations in the testbed. For this we install and configure different scenarios. For complex scenarios to be rolled out in real life, the industry partners install and configure the scenario in the testbed in order to get an implementation as close to reality as possible.

The routines for the VoIP testbed are as follows: After having installed and configured the lab to a given scenario, the setup is documented and the relevant configuration files are included into the configuration management. The testbed provides our partners with a VoIP infrastructure for experimentation, analysis, testing and prototyping of SIP/VoIP components in a controlled environment before deployment.

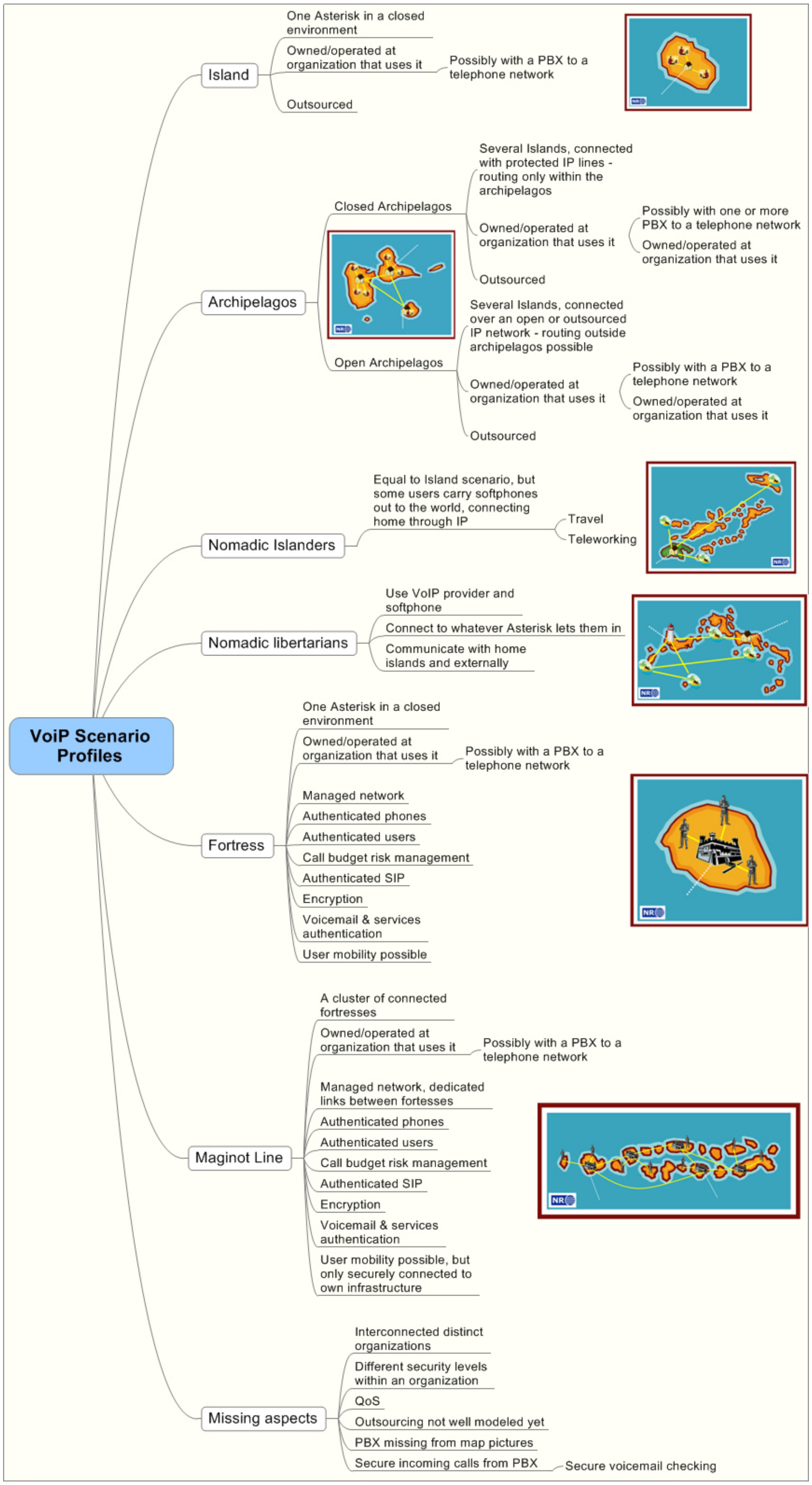

VoIP Scenario Profiles
Island
One Asterisk in a closed environment
Owned/operated at organization that uses it
Possibly with a PBX to a telephone network
Outsourced
Archipelagos
Closed Archipelagos
Several Islands, connected with protected IP lines - routing only within the archipelagos
Owned/operated at organization that uses it
Possibly with one or more PBX to a telephone network
Owned/operated at organization that uses it
Outsourced
Open Archipelagos
Several Islands, connected over an open or outsourced IP network - routing outside archipelagos possible
Owned/operated at organization that uses it
Possibly with a PBX to a telephone network
Owned/operated at organization that uses it
Outsourced
Nomadic Islanders
Equal to Island scenario, but some users carry softphones out to the world, connecting home through IP
Travel
Teleworking
Nomadic libertarians
Use VoIP provider and softphone
Connect to whatever Asterisk lets them in
Communicate with home islands and externally
Fortress
One Asterisk in a closed environment
Owned/operated at organization that uses it
Possibly with a PBX to a telephone network
Managed network
Authenticated phones
Authenticated users
Call budget risk management
Authenticated SIP
Encryption
Voicemail & services authentication
User mobility possible
Maginot Line
A cluster of connected fortresses
Owned/operated at organization that uses it
Possibly with a PBX to a telephone network
Managed network, dedicated links between fortesses
Authenticated phones
Authenticated users
Call budget risk management
Authenticated SIP
Encryption
Voicemail & services authentication
User mobility possible, but only securely connected to own infrastructure
Missing aspects
Interconnected distinct organizations
Different security levels within an organization
QoS
Outsourcing not well modeled yet
PBX missing from map pictures
Secure incoming calls from PBX
Secure voicemail checking
NR

Fig. 7. Island metaphors for VoIP scenario profiles

The project partners responded entirely positive to having a testbed provided by the project. We observed a high interest to use the lab, and we conclude that there is a need for telecom testbeds available for research and experimentation.

A typical test-run proceeds as follows: The industry partner's request can range from a specific configuration they would like to test to a more broad "we want a more secure authentication". We then identify research questions that can be applied to this test. Examples for such research questions might be how to evaluate the performance difference between two authentication mechanisms, or to evaluate their vulnerability to remote attacks. After having configured the testbed, we execute a test, alongside which we have a range of different methods to measure on the testbed. For network performance tests, we use tools like *tcpdump, MRTG* and *Munin*, while for VoIP specific tests, we can use tools like *SIPP, SIPvicious, SIPSak, sip-kill, Scapy* or similar [36].

To implement the call-hijack attack [34] shown in Section III-A, we used three different tools: (1) the VoIP attack tool "sip-kill" to send the `SIP CANCEL` messages which block out Alice and Bob from the phone call, (2) the generic attack tool "Scapy" to send the remaining SIP messages between Frank and Ivory, and (3) the multimedia stream-server "VLC" to set up a RTP media stream between the attackers.

Technical setup: Our lab today consists of a wide variety of components with different hardware and software. The main platform for VoIP servers is Asterisk on Linux. See Table IV for a list of the equipment currently used in the lab.

We carefully document all different setups in an internal wiki, and keep all relevant configurations files under revision control. Using configuration management enables us to deploy repeatable, accurate test frameworks, to repeat a particular test under the same conditions for reproducibility, or to test a particular scenario with added functionality. In our lab we have set up and installed other standard services, such as internal DNS, email, LDAP, DHCP, and monitoring tools. These services are part of VoIP infrastructures, and therefore must be included in the testbed.

To capture raw network traffic from our testbed, we can use *tcpdump* on the participating hosts. However *tcpdump* can inflict a severe performance penalty at high network throughput, and thus (potentially) affects the measurement itself. To avoid this, we have enabled "port spanning", also called "port mirroring", on the network switch. This functionality duplicates network traffic from one network port to another. On the mirrored network port, we have a high end server running *tcpdump* that captures all network traffic.

We are aware that realistic VoIP experiments require a distributed testbed running over the Internet. Therefore, we have a permanent SIP trunk over the Internet to a public telephony provider in Norway. This enables us to make real-world phone calls. We have also performed VoIP tests to other project partners over the Internet using VoIP servers installed and configured at their locations.

Our current lab scenario setup is depicted in Fig. 9. The system layout is a replica of a large scale VoIP installation from one of our project partners. This configuration involves three SIP servers, 16 SIP phones as well as ordinary infrastructure services like DNS, email and so forth. In this scenario, all the phones have real-world phone numbers (reachable from the outside). The two different network segments, labeled as "Company A" and "Company B" can also represent two different departments inside a larger company.

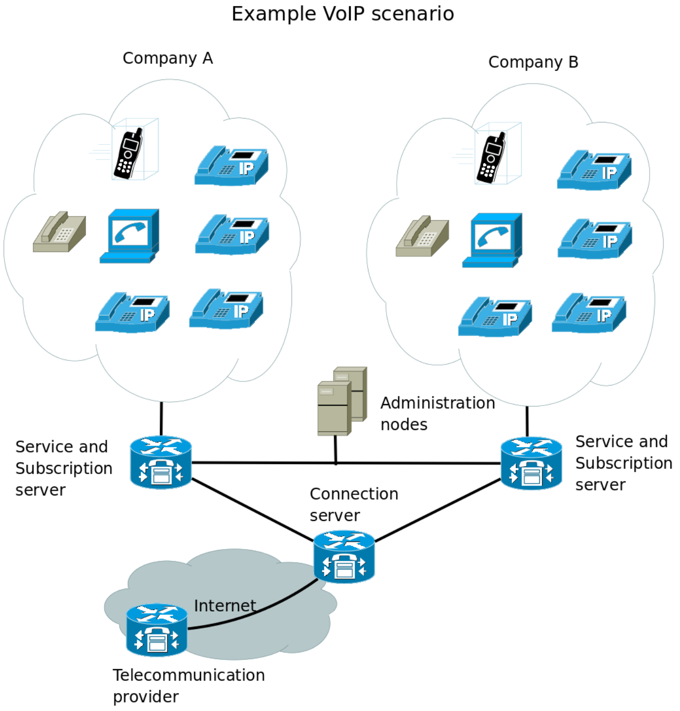

Example VoIP scenario
Company A
Company B
Administration nodes
Service and Subscription server
Service and Subscription server
Connection server
Internet
Telecommunication provider

Fig. 9. A real-life VoIP scenario replicated in testbed

Penetration testing: An ongoing penetration test with external and internal attacks uses several security consultants as hired "evil hackers" trying to attack and compromise the installation. For this test we have set up an automatic phone conversation with a pre-recorded message setting up a new conversation every fifth minute, in which both participants play a pre-recorded message and then hang up. The conversation is between our testlab and a smaller lab located at one of our industry partners.

Each attacker gets an allocated time-slot (usually a day) where he can perform his attacks. The attackers are free to do whatever attack they can think of, but we instruct them to log every command (and output) with a timestamp, and we require that they write down a report of their method and findings. We will also debrief them after each attack attempt. At our side we carefully monitor the system for any changes, and we do a full network sniffing of all raw network traffic.

We plan several iterations using this scenario: We envisage first an external attack, and second an attack from the inside, impersonating a disgruntled employee of an organization. When the attackers perform an external attack, they are given two phone numbers and one external IP address of a VoIP server. Attackers on the inside also can log in and access the network infrastructure. As a usual action-pattern the attackers first gather information ("footprinting") about the victim, in terms of network infrastructure, VoIP platform, version num-

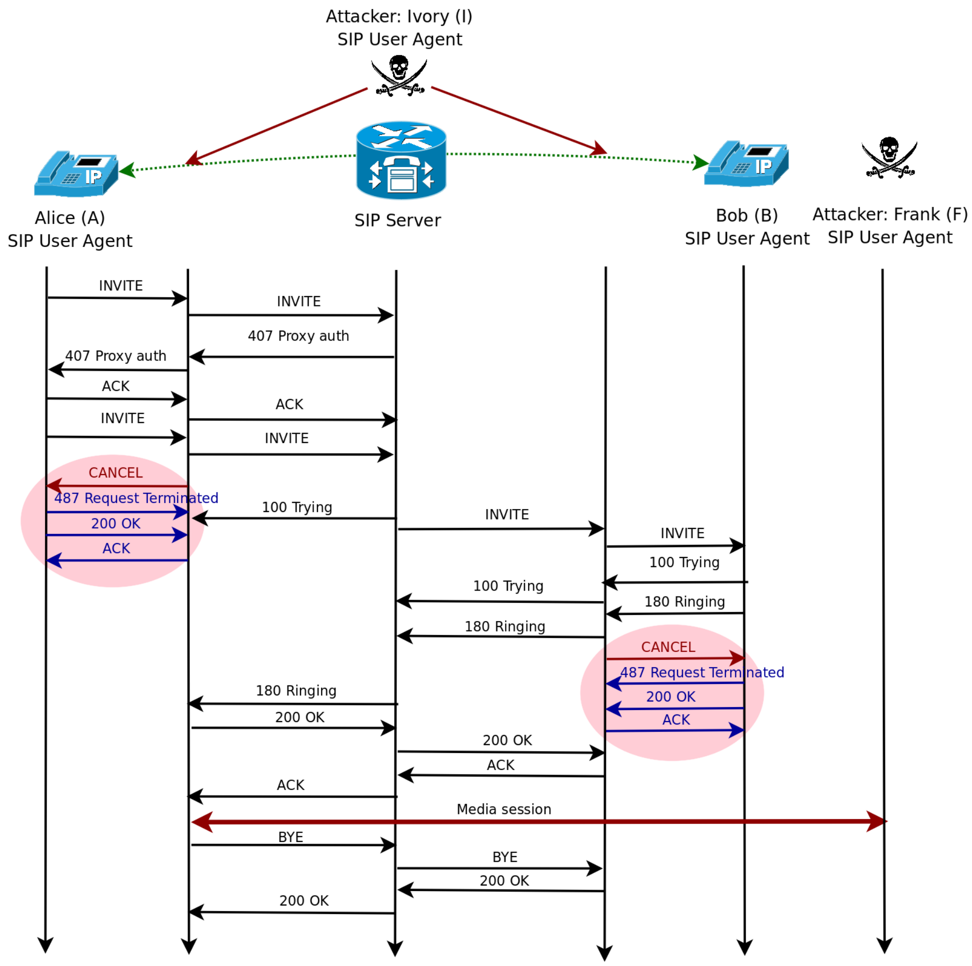


Fig. 8. Hijacking the initiator and the responder.

TABLE IV
LIST OF TESTBED EQUIPMENT AND FUNCTIONS

Function	Equipment	Software	Comment
VoIP servers (UAS)	3 high-end servers	Asterisk or OpenSIPS on Linux	Hardware typically used by several of our project partners
VoIP clients (UAC)	16 SIP hardphones (8 different models), 2 different SIP softphones, 2 soft switchboards on laptop computers	Proprietary; softphones are free software.	Phone models typically used by our project partners.
Administrative functions	1 high-end server, 1 desktop machine	DNS, LDAP, email, Subversion, Munin, Nagios, MRTG, Wiki	Relevant IT infrastructure services and monitoring
Network sniffing	1 high-end server	tcpdump	Network sniffing to disk.
Attack nodes	2 desktop machines	various	Various VoIP and network attack tools.
Connectivity	Internet, VPN		Mobile users normally use VPN. Test of UAC over VPN.

bers, and so on, before they perform any active attack.

To rank the attacks, we have set up a score board that is handed out to the attackers, with a prioritized list of security goals. The highest goal is modification of voice messages, i.e., to change one participants media stream (voice) in real-time undetected. We do not have any expectation that the attackers will be able to achieve this, but other more trivial attacks could be plausible, such as attacks on availability (DoS attack) or various SIP methods (registration, call-setup etc).

An external attack iteration is currently ongoing. During our experiment, one attacker was able to uncover a misconfigured service on the Asterisk VoIP server and log in. He did not manage to exploit this configuration error, but others might. Unless the attackers are able to compromise the VoIP server, we expect limited results from this iteration. Since the attackers do not have control over any relevant network infrastructure, it is hard or even impossible to intercept and modify the VoIP traffic.

IV. CONCLUSION AND FUTURE WORK

As an outlook into the crystal ball for 2011, we see that EUX2010SEC will have developed security guidelines, best practices and configurations for several VoIP scenarios that reflect business or user needs, and innovative options of VoIP technology. The configurations have been tested in the testbed, and aspects of them have been formally modeled and checked. The methodology of formal-methods based protocol analysis and implementation verification has been applied, improved and advanced. Thus we enable the practitioners to roll out better products and innovative services with high security levels.

From the interviews with stakeholders, we have had easy access to scenarios leading to only few of the profiles metaphors we have come up with. Is this due to our inability of covering all the different predefined profiles, or is this also reflecting the status, maturity, or majority of the (Norwegian) market? After having frequent contact with the VoIP market in Norway the last couple of years, it seems that replication of old telephony concepts onto VoIP infrastructure is where most organizations are today. The desire for enhanced functionality will sooner or later be pushing the limits in many organizations. The requirements elicitiation process therefore has to take into consideration both the requirements elicited from the interviews, but also near-future trends regarding the functionality. This makes it easier to help and guide organizations that are going to move from a conservative profile to a more challenging one.

The models described in this paper are based on security goals. Some of these goals might deduce sub-goals that are related to the selection of protocols and associated security mechanisms. Having the ability to use (deduced) security goals from the security models when performing formal protocol analysis, represents an added value when it comes to validation of systems against security models. Likewise, having a library of verfied protocols will also be valuable. Having a formal analysis of a protocol, the results are further taken into the testbed for validation. This to see if potential vulnerabilities identified in the formal analysis can be constructed at the system level, and under which conditions. Since the Asterisk systems are fully flexible the various configurations have to be validated against the security goals of the security models.

V. Acknowledgements

This research is funded by the EUX2010SEC project in the VERDIKT framework of the Research Council of Norway (Norges forskingsråd, project 180054). The authors would like to thank the anonymous reviewers for comments on earlier drafts of this paper.

References

[1] Lothar Fritsch, Arne-Kristian Groven, and Lars Strand. A holistic approach to open-source VoIP security: Preliminary results from the EUX2010sec project. In *Proceedings of the Eight International Conference on Networking (ICN2009)*, pages 275–280. IEEE Computer Society, March 2009.

[2] Dorothy Leonard and Jeffrey Rayport. Spark innovation through emphatic design. *Harvard Business Review*, 75(6):102, 1997.

[3] Richard Lester. Universities, Innovation,and the competitiveness of local economies - MIT-IPC-05-010. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA, December 2005.

[4] Thomas Porter. *Practical VoIP Security*. Syngress, March 2006.

[5] David Endler and Mark Collier. *Hacking Exposed VoIP: Voice over IP Security Secrets and Solutions*. McGraw-Hill Osborne Media, 2006.

[6] Jonathan Zar. VOIPSA VoIP Security and Privacy Threat Taxonomy - Public release 0.1. Technical report, October 2005.

[7] Richard Kuhn, Thomas Walsh, and Steffen Fries. Security Considerations for Voice over IP Systems - Recommendations of the National Institute of Standards and Technology. Technical report, US Nat'l Inst. Standards and Technology, Gaithersburg, MD, USA, 2005.

[8] M. Manulis, A. Adelsbach, A. Alkassar, K-H. Garbe, M. Luzaic, E. Scherer, J. Schwenk, and E. Siemens. VoIPSEC – Studie zur Sicherheit von Voice over Internet Protocol. Technical report, Godesberger Allee 185-189, 53175 BONN, 2005.

[9] Patrick Hung and Miguel Martin. Through the looking glass: Security issues in VoIP applications. In *IADIS International Conference on Applied Computing*, San Sebastian, Spain, 2006.

[10] Prateek Gupta and Vitaly Shmatikov. Security Analysis of Voice-over-IP Protocols. In *Proceedings of the 20th IEEE Computer Security Foundations Symposium, 2007. CSF '07*, pages 49–63. IEEE, 2007.

[11] Angelos D. Keromytis. Voice over ip: Risks, threats and vulnerabilities. In *Proceedings of the Cyber Infrastructure Protection (CIP) Conference*, New York, June 2009.

[12] Henry Sinnreich and Alan B. Johnston. *Internet communications using SIP: Delivering VoIP and multimedia services with Session Initiation Protocol*. John Wiley Sons, Inc., New York, NY, USA, second edition, August 2006.

[13] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. RFC 3261 - SIP: Session Initiation Protocol. Technical Report 3261, Internet Engineering Task Force, June 2002.

[14] Anders Moen Hagalisletto and Lars Strand. Formal modeling of authentication in SIP registration. In *Second International Conference on Emerging Security Information, Systems and Technologies SECURWARE '08*, pages 16–21. IEEE Computer Society, August 2008.

[15] S. El Sawda and P. Urien. SIP Security Attacks and Solutions: A state-of-the-art review. In *Proc. 2nd conference on Information and Communication Technologies, ICTTA '06*, volume 2, pages 3187–3191. IEEE, 2006.

[16] Geneiatakis D., Kambourakis G., Dagiuklas T., Lambrinoudakis C., and Gritzalis S. SIP Security Mechanisms: A state-of-the-art review. In *Fifth International Network Conference (INC 2005)*, pages 147–155. July 2005.

[17] Lothar Fritsch. Privacy-Respecting Location-Based Service Infrastructures: A Socio-Technical Approach to Requirements Engineering. *Journal of Theoretical and Applied E-Commerce research*, 2(3):1–17, December 2007.

[18] Lothar Fritsch and Tobias Scherner. A Multilaterally Secure, Privacy-Friendly Location-based Service for Disaster Management and Civil Protection. In Pascal Lorenz and Petre Dini, editors, *Networking - ICN 2005 - Proceedings of the 4th International Conference on Networking, Reunion Island (LNCS 3421), France, April 17-21, 2005*, volume 3421 of *Lecture Notes on Computer Science*, pages 1130–1137. Springer, Berlin, Heidelberg, New York, 2005.

[19] Benjamin L. Crosby. Stakeholder Analysis: A vital tool for strategic managers. *USAID IPC Technical Notes*, 2, March 1991.

[20] Günter Müller and Kai Rannenberg. *Multilateral Security in Communications - Technology, Infrastructure, Economy*. Addison-Wesley-Longman, München, 1999.

[21] Tom Sommerlatte. *Angewandte Systemforschung: ein interdisziplinärer Ansatz*. Gabler, Wiesbaden, first edition, 2002.

[22] J. Ramon Gil-Garcia, Theresa A. Pardo, and Andrea Baker. Understanding Context through a Comprehensive Prototyping Experience: A Testbed Research Strategy for Emerging Technologies. In *40th Hawaii International Conference on System Sciences (HICSS)*, page 104, Hawaii, 2007.

[23] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A Transport Protocol for Real-Time Applications. RFC 3550 (Standard), July 2003.

[24] Anders-Moen Hagalisletto. *Automated support for the Design and Analysis of Security Protocols*. PhD thesis, University of Oslo, Oslo, June 2007.

[25] John C. Mitchell, Ajith Ramanathan, Andre Scedrov, and Vanessa Teague. A probabilistic polynomial-time process calculus for the analysis of cryptographic protocol. *Theoretical Computer Science*, 353(1-3):118–164, March 2006.

[26] Nancy A. Durgin, Patrick Lincoln, and John C. Mitchell. Multiset rewriting and the complexity of bounded security protocols. *Journal of Computer Security*, 12(2):247–311, 2004.

[27] Michael Burrows, Martin Abadi, and Roger Needham. A logic of authentication. *ACM Trans. Comput. Syst.*, 8(1):18–36, 1990.

[28] J. C. Mitchell, M. Mitchell, and U. Stern. Automated analysis of cryptographic protocols using Murphi. In *IEEE Symposium on Security and Privacy 1997*, pages 141–151. IEEE Compuer Society, 1997.

[29] Gavin Lowe. Casper: a compiler for the analysis of security protocols. *Journal of Computer Security*, 6(1-2):53–84, 1998.

[30] David Basin, Sebastian Mödersheim, and Luca Viganò. OFMC: A symbolic model checker for security protocols. *International Journal of Information Security*, 4(3):181–208, June 2005.

[31] C.J.F. Cremers. The Scyther Tool: Verification, falsification, and analysis of security protocols. In *Computer Aided Verification, 20th International Conference, CAV 2008, Princeton, USA, Proc.*, volume 5123/2008 of *Lecture Notes in Computer Science*, pages 414–418. Springer, 2008.

[32] Sebastian Mödersheim. Algebraic properties in alice and bob notation. *Availability, Reliability and Security, International Conference on*, 0:433–440, 2009.

[33] Jennifer G. Steiner, B. Clifford Neuman, and Jeffrey I. Schiller. Kerberos: An authentication service for open network systems. In *USENIX Winter*, pages 191–202, 1988.

[34] Anders Moen Hagalisletto, Lars Strand, Wolfgang Leister, and Arne-Kristian Groven. Analysing protocol implementations. In Feng Bao, Hui Li, and Guilin Wang, editors, *The 5th Information Security Practice and Experience Conference (ISPEC 2009)*, volume LNCS 5451, pages 171–182. Springer Berlin / Heidelberg, April 2009.

[35] D. Dolev and A. Yao. On the security of public key protocols. *IEEE Transactions on Information Theory*, 29(2):198–208, Marc-1983 1983.

[36] Dorgham Sisalem, John Floroiu, Jiri Kuthan, Ulrich Abend, and Henning Schulzrinne. *SIP Security*. WileyBlackwell, March 2009.

Encoding Subsystem Codes

Pradeep Sarvepalli
Department of Physics and Astronomy
University of British Columbia
Vancouver, BC V6T 1Z1, Canada
Email: pradeep@phas.ubc.ca

Andreas Klappenecker
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843, USA
Email: klappi@cse.tamu.edu

***Abstract*— Encoding quantum codes is an important component of quantum error correction and of relevance in some cryptographic protocols such as entanglement distillation and quantum secret sharing. In this paper, we investigate the encoding of subsystem codes and propose efficient encoding methods for them. We show that encoding of subsystem codes can be reduced to encoding of a related stabilizer code making it possible to use the known results on encoding of stabilizer codes. Along the way we also show how Clifford codes can be encoded. We present two systematic methods to encode subsystem codes and suggest optimizations for lower complexity. These encoding schemes can tolerate initialization errors on the so-called gauge qubits. This tolerance can be traded for reduced encoding complexity.**

***Index Terms*—Clifford codes; encoding; quantum codes; stabilizer codes; subsystem codes;**

I. Introduction

Though originally conceived to protect quantum information in the context of quantum computation, quantum codes have found important applications in other areas such as quantum cryptography and have strong connections with many cryptographic protocols such as entanglement distillation [4], key distribution [35] and secret sharing [9], [14]. A secret sharing scheme, for instance, could be viewed as a quantum error correcting code and the method of generating the shares and the subsequent reconstruction procedures are directly related to the encoding and decoding of the associated quantum code. Compared to decoding, the problem of efficient encoding has received less attention from the research community. Our interest in this paper is the encoding of subsystem codes, for two reasons. Firstly, because of their importance in quantum error correction and because this topic has not received a comprehensive treatment in the literature before. Secondly, subsystem codes generalize some of the cryptographic protocols based on standard error correcting codes [26] and the results of this paper will be potentially useful in such protocols.

In most error correction paradigms such as the stabilizer codes [5]–[7], [11]–[13], [25], one protects the information by encoding into a subspace Q of $\mathcal{H}$, the system Hilbert space of n qubits or more generally qudits, *i.e.*, q-level systems, thus $\mathcal{H} \cong \mathbb{C}^{q^n}$. We refer to Q as code subspace or codespace; Q induces a decomposition of Hilbert space as $\mathcal{H} = Q \oplus Q^\perp$, where $Q^\perp$ is the complement of Q. If $\dim Q = q^k$ and the code can detect errors on $d-1$ qudits or fewer, then we denote this code as an $[[n,k,d]]_q$ code.

Subsystem codes, or operator quantum error correcting codes as they are also called, generalize the standard notion of protecting by encoding into subspaces of the system Hilbert space, see [3], [24], [28], [29]. In this case the subspace Q can be factored into a tensor product of two subspaces, thus $\mathcal{H} = (A \otimes B) \oplus Q^\perp$. Only the subsystem A carries the information to be protected, while the auxiliary subsystem B does not and it is referred to as the gauge subsystem. The gauge subsystem could, in some cases, lead to a simplification of the error recovery schemes. If $\dim A = q^k$, $\dim B = q^r$ and the code can detect all errors on $d-1$ qudits or less, then we say it is an $[[n,k,r,d]]_q$ subsystem code. Such a code is said to encode k qudits with r gauge qudits.

We pause to clarify the terminology of gauge qudits. These are not as the name might suggest a new type of qudits. Let us elaborate a little further. In the standard error correction model, every input state is encoded to a unique state. On the other hand, in the subsystem model we do not have such a unique map between input states and the corresponding encoded states. Instead every input state is identified with a unique subspace, but the state itself can be mapped to any state within the subspace. These degrees of freedom are what we equate with the gauge qudits. If the subspace is q^r-dimensional, then we say that the subsystem code has r gauge qudits, although one might prefer to use the term gauge degrees of freedom. The terminology is not entirely unjustified though, as we can exchange these degrees of freedom for encoding more qudits, see for instance [2].

The development of subsystem codes was motivated to a great extent by the search for efficient error recovery schemes. Subsequent research has demonstrated that indeed in some cases these codes afford simpler recovery schemes that led to an improvement in the threshold for fault tolerant quantum computation [1].

Our first result is that encoding of a subsystem code can be reduced to the encoding of a related stabilizer code, thereby making use of the previous theory on encoding stabilizer codes [8], [13], [16]. We shall prove this in two steps. We begin by showing that Clifford codes [20] can be encoded using the same methods used for stabilizer codes. Subsequently, we shall show how these methods can be adapted to encode Clifford subsystem codes. Since subsystem codes subsume stabilizer codes, noiseless subsystems and decoherence free subspaces, these results imply that we can essentially use the same methods to encode all these codes. In fact, while the exact details were not provided, Poulin suggested in [33] that encoding of subsystem codes can be achieved by Clifford unitaries. Our treatment is comprehensive and gives proofs for all the claims.

Our second result is more pragmatic in nature and is concerned with actual circuits for encoding. It is partly motivated by the idea that just as subsystem codes can potentially lead to simpler error recovery schemes, they can also simplify the encoding process, though perhaps not as dramatically. These simplifications have not been investigated thoroughly, neither have the gains in encoding been fully characterized. Essentially, these gains are in two forms. In the encoded state there need not exist a one to one correspondence between the gauge qubits and the physical qubits. However, prior to encoding such a correspondence exists. We can exploit this identification between the gauge degrees of freedom and the physical qubits before encoding to tolerate errors on the gauge qubits, a fact which was recognized in [33]. Alternatively, we can optimize the encoding circuits by eliminating certain encoding operations. The encoding operations that are saved correspond to the encoded operators on the gauge qubits. This is a slightly subtle point and will become clear later. We argue that optimizing the encoding circuit for the latter is much more beneficial than simply allowing for random initialization of gauge qubits.

It must be noted that our encoding schemes are quite general in that we do not tailor our encoding schemes to any specific noise process. We assume that the code construction has already taken the noise process into account while designing the code. At this point, we are not concerned with the code design but with the the implementation of the code or more precisely, implementation of a specific task with respect to code namely, encoding. We do assume that the noise process is local, *i.e.*, independent on each qudit. Although, error correction is not primarily the objective of encoding per se, we do show that in case of subsystem codes, the encoding schemes can offer a significant benefit viz., they can tolerate initialization errors on some qubits, namely the gauge qubits. These qubits can be completely corrupted. The results presented here could perhaps be optimized for a specific noise process to get additional benefits but we do not investigate these possibilities.

This paper is structured as follows. In Section II we give a brief sketch of the representation theoretic framework of quantum codes, for the benefit of readers who are not familiar with this approach. We then deal with the problem of encoding in its most general setting in Section III, where we make no reference to the alphabet of the codes. Subsequently, in Section IV we address in detail the differences that arise in encoding into a subspace versus a subsystem. In Sections V and VI, we give two different methods to encode the subsystem codes.

Notation. We shall denote a finite field with q elements by $\mathbb{F}_q$. Following standard convention we use $[[n,k,d]]_q$ for stabilizer codes and $[[n,k,r,d]]_q$ for subsystem codes. The inner product of two characters of a group N, say χ and θ, is defined as $(\chi,\theta)_N = 1/|N|\sum_{n\in N}\chi(n)\theta(n^{-1})$. We shall denote the center of a group N by $Z(N)$. Given a subgroup $N \leq E$, we shall denote the centralizer of N in E by $C_E(N)$. Given a matrix A, we consider another matrix B obtained from A by column permutation π as being equivalent to A and denote this by $B =_\pi A$. In other words, B can be obtained from A after applying a permutation π. Often we shall represent the basis of a group by the rows of a matrix. In this case we regard another basis obtained by any row operations or permutations as being equivalent and by a slight abuse of notation continue to denote $B =_\pi A$. The commutator of two operators A, B is defined as $[A,B] = AB - BA$.

We note that this paper is an expanded version of [34]. It addresses in more detail the differences between encoding into a subspace and subsystem, additionally it includes alternative methods to encode subsystem codes and discusses further variations.

II. Quantum Codes from a Representation Theoretic Perspective

In part of the paper we lean heavily on a representation theoretic framework of quantum codes. So we sketch the relevant ideas of this framework for those readers unfamiliar with this approach, postponing some of the mathematical details to the appropriate juncture in the paper; interested readers can find further details in [22], [23] and [20]. For an introduction to quantum error correction in general we refer the readers to [10], [15], [18], [30], [31]. An introduction to subsystem codes can be found in [27], [33], while interested readers are referred to [21], [24], [28], [29], [32].

Recall that quantum states are unit vectors in the Hilbert space $\mathcal{H}$, which is a q^n-dimensional complex vector space. Protecting a set of quantum states implies that we are required to protect the subspace spanned by them because, typically, quantum algorithms manipulate not just a set of logical states but also their complex superpositions. So quite naturally the computational space is a vector space. We designate this space to be protected as the codespace, Q. While there are important differences, just as in classical error correction, redundancy is a key ingredient of quantum error correction. Consequently, the codespace cannot be the entire Hilbert space, if it were, then every state would be a valid state and we would not know if that state had been corrupted by noise or not. Put differently, this means that the codespace Q is a proper subspace of $\mathcal{H}$. If the dimension of Q is q^k, we denote this as an $[[n,k]]_q$ quantum code.

Errors are simply operators on $\mathcal{H}$, in other words they are elements of the matrix algebra of $q^n \times q^n$ matrices. It suffices to consider only a basis for this matrix algebra; this basis is called the error basis and denoted as $\mathcal{E}$. It is convenient to work with an orthonormal basis, assuming a suitable definition of the inner product between the matrix elements. The group of operators generated by the elements of $\mathcal{E}$ is called the error group and denoted as $\overline{\mathcal{E}}$. Knill [23] introduced the concept of nice error basis which is an orthonormal error basis with these additional restrictions: (a) it contains contains the identity, (b) its elements are unitary operators, and (c) the product of any two basis elements is another basis element up to a scalar multiple. The motivation for these particular conditions can be found in [23].

A nice error basis induces a group called the abstract error group, which we denote by E. The abstract error group is isomorphic to the error group $\overline{\mathcal{E}}$. Additionally, the elements of $\mathcal{E}$ can be indexed by elements of $E/Z(E)$, therefore $E/Z(E)$ is called the index group. Furthermore, $\overline{\mathcal{E}}$ is a faithful, irreducible unitary representation of E.

The error basis formulation, although abstract, is quite useful in that we can construct and study codes using the machinery of group theory and representations. The connection with codes is as follows. A quantum code is a "suitably defined eigenspace" of the operators of a subgroup of $\overline{\mathcal{E}}$. Let G be a subgroup of $\overline{\mathcal{E}}$. If G is an abelian group that does not contain $Z(E)$, then the codespace is defined as the "joint eigenspace" of all the operators in G. These codes are precisely the stabilizer codes in the sense of Gottesman [12]. et al [5]. Knill [22] considered quantum codes derived from normal subgroups of $\overline{\mathcal{E}}$ which are not necessarily abelian; these codes, called Clifford codes, are the objects underlying our study in this paper. In this case, the codespace is defined in a slightly more complex manner involving the characters of the representation of E. The codespace is an eigenspace of each operator in G, but the eigenvalues may now vary from operator to operator.

As we mentioned earlier subsystem codes are quantum codes which afford a tensor product decomposition of the codespace. Clifford codes turn out to have a tensor product decomposition which makes them a natural candidate for constructing the subsystem codes, (thereby prompting our study of Clifford codes in this paper). This decomposition is related to the representation and the characters of the abstract error group. Precise construction of subsystem codes from Clifford codes can be found in [21].

It is not possible to correct all errors that occur on the code space. We usually attempt to correct those errors that are most likely to occur. Assume that the code corrects a set of errors in $\mathcal{A}$. Then it also corrects the linear span of those errors. Therefore, it suffices to correct only a basis of errors. An error basis of the state space of n qudits is a tensor product of the error basis on a single quantum system. One can therefore meaningfully speak of a local error model where we assume that the errors on each qudit are independent. We can quantify the error correcting capabilities of the code in terms of the errors the code can correct or equivalently in terms of the errors it can detect. If the code can detect errors on any $d-1$ or fewer quantum systems we say that the code has a distance d.

From the point of view of code construction, for optimal performance, one should take into account the source and characteristics of noise. In the absence of exact knowledge about the noise characteristics it is common to assume a pessimistic noise model. The noise model affects the choice of the code and in this paper we assume that the code has been constructed factoring the noise model.

III. Encoding Clifford Codes

In this section, we show that a Clifford code can be encoded using its stabilizer and therefore the methods used for encoding stabilizer codes are applicable. We briefly recapitulate some facts about Clifford subsystem codes, see [21] for more details. Let E be an *abstract error group*, *i.e.*, it is a finite group with a faithful irreducible unitary representation ρ of degree $|E : Z(E)|^{1/2}$. Denote by ϕ, the irreducible character afforded by ρ. Let N be a normal subgroup of E. Further, let χ be an irreducible character χ of N such that $(\phi_N, \chi)_N > 0$, where ϕ_N is the restriction of ϕ to N.

Definition 1 (Clifford code)**.** *The Clifford code defined by (E, ρ, N, χ) is the image of the orthogonal projector*

$$P = \frac{\chi(1)}{|N|} \sum_{n \in N} \chi(n^{-1})\rho(n). \tag{1}$$

Under certain conditions we can construct a subsystem code from the Clifford code. In particular when the *index group*, *i.e.*, $E/Z(E)$ is abelian and $C_E(Z(N)) = LN$, the Clifford code C has a tensor product decomposition[1] as $Q = A \otimes B$, where B is an irreducible $\mathbb{C}N$-module and A is an irreducible $\mathbb{C}L$-module. In this case we can encode information only into the subsystem A, while the co-subsystem B provides additional protection. When encoded this way we say Q is a *Clifford subsystem code*. The normal subgroup N consists of all errors in E that act trivially on A. It is also called the *gauge group* of the subsystem code. Our main goal will be to show how to encode into the subsystem A. The dimensions of A and B can be computed using [21, Theorems 2,4] but, since we are interested in encoding we focus on the projectors for the Clifford code and the subsystem code and not so much on the parameters of the codes themselves.

An alternate projector for a Clifford code with data (E, ρ, N, χ) can be defined in terms of $Z(N)$, the center of N. This projector is given as, see [20, Theorem 6] for proof,

$$P' = \frac{1}{|Z(N)|} \sum_{n \in Z(N)} \varphi(n^{-1})\rho(n), \tag{2}$$

where φ is an irreducible (linear) character of $Z(N)$, that satisfies $(\chi \downarrow Z(N))(x) = \chi(1)\varphi(x)$, where $(\chi \downarrow Z(N))(x)$ is the restriction of χ to $Z(N)$. In this case Q can be thought of as a stabilizer code in the sense of [5], *i.e.*,

$$\rho(m)\,|\psi\rangle = \varphi(m)\,|\psi\rangle \text{ for any } m \text{ in } Z(N). \tag{3}$$

We pause to mention that stabilizer codes can be viewed in two equivalent ways. We could view them as the joint +1-eigenspaces of an abelian subgroup of the error group, this is the sense in which stabilizer codes are defined by Gottesman [12]. Alternatively, we could augment this subgroup by the center of the error group to define the code, as in [5]. In

[1] Strictly speaking the equality should be replaced by an isomorphism.

the latter case the codespace is not anymore the joint +1-eigenspace of the operators of the subgroup. We account for the varying eigenvalues by a character of the subgroup.

Our goal is to use the stabilizer of Q for encoding and as a first step we will show that it can be computed from $Z(N)$. The usefulness of such a projector is that it obviates the need to know the character φ.

Lemma 2. *Let (E,ρ,N,χ) be the data of a Clifford code and φ an irreducible character of $Z(N)$, the center of N, satisfying $(\chi\downarrow Z(N))(x)=\chi(1)\varphi(x)$. Let e be the exponent of E and let e divide $|Z(E)|$. Then for all n in $Z(N)$, $\varphi(n)\in\{\zeta^k \mid \zeta=e^{j2\pi k/e}, 0\le k<e\}$. Further, if $Z(E)\le N$, then for any $n\in Z(N)$, we have $\varphi(n^{-1})\rho(n)\in\rho(Z(N))$.*

Proof: First we note that the irreducibilty of ρ implies that for any z in $Z(E)$ we have $\rho(z)=\omega I$ for some $\omega\in\mathbb{C}$ by Schur's lemma, (or see [17, Prop. 9.14, pg. 84]). The assumption that ρ is also faithful implies that $Z(E)$ is cyclic, [17, Prop. 9.16, pg. 85] and e divides $|Z(E)|$ forces $|Z(E)|=e$; consequently, $\omega\in\{\zeta^k \mid 0\le k<e\}$ where $\zeta=e^{j2\pi/e}$. Since ρ is faithful $\rho(Z(E))=\{\zeta^l I \mid 0\le l<e\}$. Secondly, we observe that φ is an irreducible additive character of $Z(N)$ (an abelian group with exponent at most e) which implies that we must have $\varphi(n)=\zeta^l$ for some $0\le l<e$. From these observations with the fact ρ is faithful, we infer that $\varphi(n^{-1})I=\zeta^l I=\rho(z)$ for some $0\le l<e$ and $z\in Z(E)$. Since $Z(E)\le N$, it follows that $Z(E)\le Z(N)$ and $\varphi(n^{-1})\rho(n)=\rho(zn)$ is in $\rho(Z(N))$. ■

Theorem 3. *Let Q be a Clifford code with the data (E,ρ,N,χ) and φ an irreducible character of $Z(N)$ satisfying $(\chi\downarrow Z(N))(x)=\chi(1)\varphi(x)$. Let E and N be as in Lemma 2 and*

$$S=\left\{\varphi(n^{-1})\rho(n)\mid n\in Z(N)\right\};\quad P=\frac{1}{|S|}\sum_{s\in S}s. \tag{4}$$

Then S is the stabilizer of Q and Im $P=Q$.

Proof: We will show this in a series of steps.

1) First we will show that $S\le\rho(Z)$. By Lemma 2 we know that $\varphi(n^{-1})\rho(n)$ is in $\rho(Z)$, therefore $S\subseteq\rho(Z)$. Let $Z=Z(N)$, for short. For any two elements $n_1,n_2\in Z$, we have $s_1=\varphi(n_1^{-1})\rho(n_1), s_2=\varphi(n_2^{-1})\rho(n_2)\in S$ and we can verify that $s_1^{-1}s_2=\varphi(n_1)\rho(n_1^{-1})\varphi(n_2^{-1})\rho(n_2)=\varphi(n_2^{-1}n_1)\rho(n_1^{-1}n_2)\in S$, as $\rho(n_1^{-1}n_2)$ is in $\rho(Z)$. Hence $S\le\rho(Z)$.
2) Now we show that S fixes Q. Let $s\in S$ and $|\psi\rangle\in Q$. Then $s=\varphi(n^{-1})\rho(n)$ for some $n\in Z$. The action of s on $|\psi\rangle$ is given as $s\,|\psi\rangle=\varphi(n^{-1})\rho(n)\,|\psi\rangle=\varphi(n^{-1})\varphi(n)\,|\psi\rangle=|\psi\rangle$, in other words S fixes Q.
3) Next, we show that $|S|=|Z|/|Z(E)|$. If two elements n_1 and n_2 in Z map to the same element in S, then $\varphi(n_1^{-1})\rho(n_1)=\varphi(n_2^{-1})\rho(n_2)$, that is $\rho(n_2)=\varphi(n_1^{-1}n_2)\rho(n_1)$. By Lemma 2, it follows that $\rho(n_2)=\zeta^l\rho(n_1)$ for some $0\le l<e$. Since $\rho(Z(E))=\{\zeta^l I\mid 0\le k<e\}$ and ρ is faithful, we must have $n_2=zn_1$ for some $z\in Z(E)$. Thus, $|S|=|Z|/|Z(E)|$.
4) Let T be a transversal of $Z(E)$ in Z, then every element in Z can be written as zt for some $z\in Z(E)$ and $t\in T$. From step 3) we can see that all elements in a coset of $Z(E)$ in Z map to the same element in S, therefore,
$$S=\{\varphi(t^{-1})\rho(t)\mid t\in T\}.$$
Recall that a projector for Q is also given by
$$\begin{aligned}P' &= \frac{1}{|Z|}\sum_{n\in Z}\varphi(n^{-1})\rho(n),\\ &= \frac{1}{|Z|}\sum_{t\in T}\sum_{z\in Z(E)}\varphi((zt)^{-1})\rho(zt).\end{aligned}$$
But we know from step 3) that if $z\in Z(E)$, then $\varphi(n^{-1})\rho(n)=\varphi((zn)^{-1})\rho(zn)$. So we can simplify P' as
$$\begin{aligned}P' &= \frac{1}{|Z|}\sum_{t\in T}\sum_{z\in Z(E)}\varphi(t^{-1})\rho(t),\\ &= \frac{|Z(E)|}{|Z|}\sum_{t\in T}\varphi(t^{-1})\rho(t)=\frac{1}{|S|}\sum_{s\in S}s=P.\end{aligned}$$
Thus the projector defined by S is precisely the same as P' and P is also a projector for Q.

From step 3) it is clear that $S\cap Z(E)=\{\mathbf{1}\}$ and by [19, Lemma 10], S is a closed subgroup of E. By [19, Lemma 9], Im $P=Q$ is a stabilizer code. Hence S is the stabilizer of Q. ■

The essence of Theorem 3 is that if one were to ignore the underlying structure of the subspace that is associated to a Clifford code, then it can be also identified with a stabilizer code.

Corollary 4. *Let Q be an $[[n,k,r,d]]_q$ Clifford subsystem code and S its stabilizer. Let*

$$P=\frac{1}{|S|}\sum_{s\in S}s. \tag{5}$$

Then P is a projector for the subsystem code, i.e., $Q=$ Im P.

Proof: By [21, Theorem 4][2], we know that an $[[n,k,r,d]]_q$ Clifford subsystem code is derived from a Clifford code with data (E,ρ,N,χ). Since as subspaces the Clifford code and subsystem code are identical, by Theorem 3 we conclude that the projector defined from the stabilizer of the subspace is also a projector for the subsystem code. ■

Theorem 3 shows that any Clifford code can be encoded using its stabilizer. As to encoding a subsystem code, while Corollary 4 shows that there exists a projector that can be defined from its stabilizer, it is not clear how to use it so that one respects the subsystem structure during encoding. More precisely, how do we use the projector defined in Corollary 4 to encode into the information carrying subsystem A and not the gauge subsystem. This will be the focus of the next section.

[2] Though [21, Theorem 4] assumes that E is an extraspecial p-group it also holds with the error groups with the conditions we have in Lemma 2 and Theorem 3.

IV. Encoding into Subspaces versus Encoding into Subsystems

For ease of presentation and clarity henceforth we will focus on binary codes, though the results can be extended to nonbinary alphabet using methods similar to stabilizer codes, see [16]. We briefly review some of the relevant background and point out the differences in encoding into subspaces and subsystems. Before we get into the details we reiterate that the result obtained in the previous section with respect to encoding for arbitrary alphabet does not yet give us explicit circuits for subsystem codes, such that they respect the subsystem structure of the code. This section prepares the way to giving those circuits for binary subsystem codes, by highlighting the differences that must be accounted for when one is encoding a stabilizer code as against a subsystem code. The results on binary subsystem codes in Sections VI and V give a concrete expression to the abstract result obtained in Section III. Additionally, they also discuss how subsystem encoding can exploit the freedom provided by the gauge qubits to either tolerate initialization errors or reduce complexity.

A. Encoding Stabilizer Codes

We shall now briefly, review the standard form encoding of stabilizer codes, due to Cleve and Gottesman, see [8], [13]. Recall the Pauli matrix operators

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} = iXZ.$$

Let $\mathcal{P}_n$ be the Pauli group on n qubits. An element element $e = i^c X^{a_1} Z^{b_1} \otimes \cdots \otimes X^{a_n} Z^{b_n}$ in $\mathcal{P}_n$, can be mapped to $\mathbb{F}_2^{2n}$ by $\tau : \mathcal{P}_n \to \mathbb{F}_2^{2n}$ as

$$\tau(e) = (a_1, \ldots, a_n | b_1, \ldots, b_n). \tag{6}$$

Given an $[[n,k,d]]_2$ code with stabilizer S, we can associate to S (and therefore to the code), a matrix in $\mathbb{F}_2^{(n-k)\times 2n}$ obtained by taking the image of any set of its generators under the mapping τ. We shall refer to this matrix as the *stabilizer matrix*. We shall refer to the stabilizer as well as any set of generators as the stabilizer. Additionally, because of the mapping τ, we shall refer to the stabilizer matrix or any matrix obtained from it by row reduction or column permutations also as the stabilizer. The stabilizer matrix can be put in the so-called "standard form", [8], [13], see also Lemma 6. This form also allows us to compute the encoded operators for the stabilizer code. Recall that the encoded operators allow us to perform computations on the encoded data without having to decode the data and then compute.

Definition 5 (Encoded operators)**.** *Given a $[[n,k,d]]_2$ stabilizer code with stabilizer S, let $\overline{X}_i$, $\overline{Z}_i$ for $1 \leq i \leq k$ be a set of $2k$ linearly independent operators in $C_{\mathcal{P}_n}(S) \setminus SZ(\mathcal{P}_n)$. The set of operators $\{\overline{X}_i, \overline{Z}_i \mid 1 \leq i \leq k\}$ are said to be encoded operators for the code if they satisfy the following requirements.*

i) $[\overline{X}_i, \overline{X}_j] = 0$

ii) $[\overline{Z}_i, \overline{Z}_j] = 0$

iii) $[\overline{X}_i, \overline{Z}_j] = 2\delta_{ij}\overline{X}_i\overline{Z}_i$

The operators $\overline{X}_i$ and $\overline{Z}_j$ are referred to as encoded or logical X and Z operators on the ith and jth logical qubits, respectively. The choice of which of the $2k$ linearly independent elements of $C_{\mathcal{P}_n}(S) \setminus SZ(\mathcal{P}_n)$ we choose to call encoded X operators and Z operators is arbitrary; as long as the generators satisfy the conditions above, any choice is valid. Different choices lead to different sets of encoded logical states; alternatively, a different orthonormal basis for the codespace. Often, as in Lemma 6 below, we refer to the binary representations of the encoded operators also as the encoded operators.

Lemma 6 (Standard form of stabilizer matrix [8], [13])**.** *Up to a permutation π, the stabilizer matrix of an $[[n,k,d]]_2$ code can be put in the following form,*

$$S =_\pi \left[\begin{array}{ccc|ccc} I_{s'} & A_1 & A_2 & B & 0 & C \\ 0 & 0 & 0 & D & I_{n-k-s'} & E \end{array}\right], \tag{7}$$

while the associated encoded operators can be derived as

$$\left[\begin{array}{c} \overline{Z} \\ \hline \overline{X} \end{array}\right] =_\pi \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & A_2^t & 0 & I_k \\ \hline 0 & E^t & I_k & C^t & 0 & 0 \end{array}\right]. \tag{8}$$

Remark 7. *Encoding using essentially same ideas is possible even if the identity matrices ($I_{s'}$ in the stabilizer matrix or I_k in the encoded operators) are replaced by upper triangular matrices.*

The standard form of the stabilizer matrix prompts us to distinguish between two types of the generators for the stabilizer as they affect the encoding in different ways (although it can be shown that they are of equivalent complexity).

Definition 8 (Primary generators)**.** *A generator $G_i = (a_1, \ldots, a_n | b_1, \ldots, b_n)$ with at least one nonzero a_i is called a primary generator.*

In other words, primary generators contain at least one X or Y operator on some qubit. As we shall see in Lemma 10, the primary generators determine to a large extent the complexity of the encoding circuit along with the encoded X operators. The operators $\overline{X}$ are also called seed generators and they also figure in the encoding circuit. The encoded Z operators do not.

Definition 9 (Secondary generators)**.** *A generator of the form $(0, \ldots, 0 | b_1, \ldots, b_n)$ is called secondary generator.*

In the standard form encoding, the complexity of the encoded X operators is determined by the secondary generators. Therefore they indirectly contribute[3] to the complexity of encoding.

We mentioned earlier that different choices of the encoded operators amounts to choosing different orthonormal basis for the codespace. However, the choice in Lemma 6 is particularly suitable for encoding. We can represent our input in the form

[3] Indirect because the submatrix E, figures in both the secondary generators, see equation (7), and also the encoded X operators, see equation (8).

$|0\rangle^{\otimes^{n-k}}|\alpha_1 \ldots \alpha_k\rangle$ which allows us to make the identification that $|0\rangle^{\otimes^n}$ is mapped to $|\overline{0}\rangle$, the logical all zero code word. This state is precisely the state stabilized by the stabilizer generators and logical Z operators, (which in Lemma 6 can be seen to be consisting of only Z operators). Given the stabilizer matrix in the standard form and the encoded operators as in Lemma 6, the encoding circuit is given as follows.

Lemma 10 (Standard form encoding of stabilizer codes [8], [13])**.** *Let S be the stabilizer matrix of an $[[n,k,d]]_2$ stabilizer code in the standard form,* i.e.*, as in equation (7). Let G_j denote the j^{th} primary generator of S and $\overline{X_l}$ denote the l^{th} encoded X operator as in equation (8). Then G_j is in the form*[4]

$$(0,\ldots,0,a_j=1,\ldots,a_n|b_1,\ldots,b_{s'},0,\ldots,0,b_{n-k+1},\ldots,b_n),$$

while $\overline{X}_l$ is in the form

$$(0,\ldots,0,c_{s'+1},\ldots,c_{n-k}0,\ldots,0,c_{n-k+l},0,\ldots,0|d_1,\ldots,d_n),$$

where $d_m = 0$ for $m \geq s'+1$. Let $P = diag(1,i)$ and $\sigma(a_l,b_l) = (-i)^{a_l b_l}X^{a_l}Z^{b_l}$. To encode the stabilizer code we implement the circuits corresponding to each of the primary generators and the encoded operators. as shown in Figure 1. The generator G_j is implemented after G_{j+1}. The encoded operators precede the primary generators in their implementation but we can implement $\overline{X}_l$ before or after $\overline{X}_{l+1}$.

To encode a stabilizer code, we first put the stabilizer matrix in the standard form, then implement the seed generators, *i.e.*, the encoded X operators, followed by the primary generators $j=s'$ to $j=1$ as per Lemma 10. The complexity of encoding the j^{th} primary generator is at most $n-j$ two qubit gates and two single qubit gates. The complexity of encoding an encoded operator is at most $n-k-s'$ CNOT gates. This means the complexity of standard form encoding is upper bounded by $O(n(n-k))$ gates.

B. Encoding Subsystem Codes

Theorem 3 shows that in order to encode Clifford codes we can use a projector derived from the underlying stabilizer to project onto the codespace. But in case of Clifford subsystem codes we know that $Q = A \otimes B$ and the information is to be actually encoded in A. Hence, it is not sufficient to merely project onto Q, we must also show that we encode into A when we encode using the projector defined in Corollary 4.

Let us clarify what we mean by encoding the information in A and not in B. Suppose that P maps $|0\rangle$ to $|\psi\rangle_A \otimes |0\rangle_B$ and $|1\rangle$ to $|\psi\rangle_A \otimes |1\rangle_B$. Then the information is actually encoded into B. Since the gauge group acts nontrivially on B, this particular encoding does not protect information. Of course a subsystem code should not encode (only) into B, but we have to show that the projector defined by P_s does not do that.

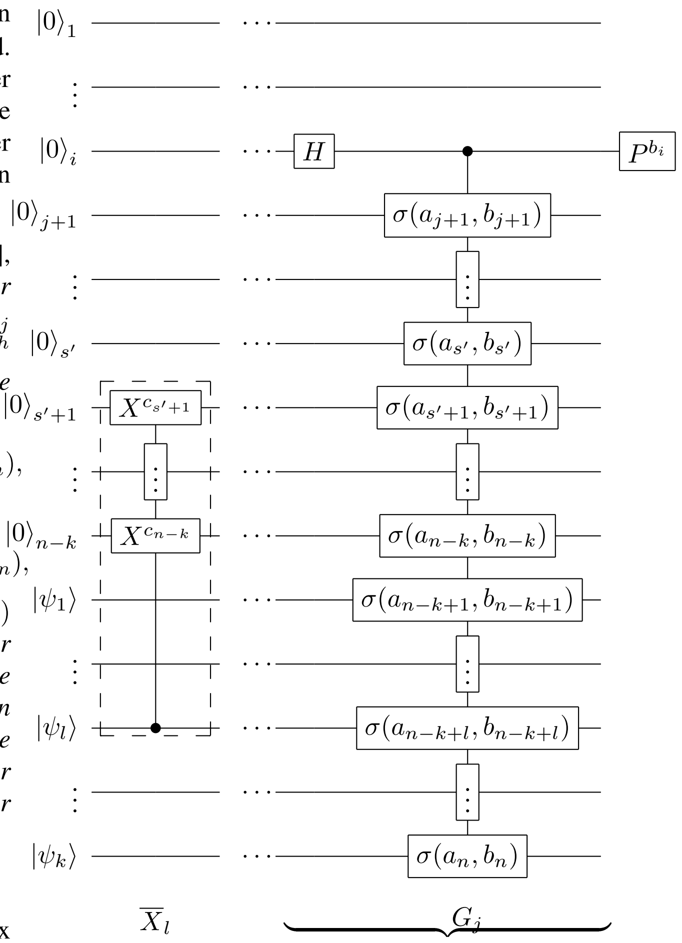


Fig. 1. Building blocks for standard form encoding of stabilizer codes.

We need the following result on the structure of the gauge group and the encoded operators of a subsystem code. Poulin [32] proved a useful result on the structure of the gauge group and the encoded operators of the subsystem code. But first a little notation. A basis for $\mathcal{P}_n$ is X_i, Z_i, $1 \leq i \leq n$, where X_i and Z_i are given as

$$X_i = \bigotimes_{j=1}^{n} X^{\delta_{ij}} \quad \text{and} \quad Z_i = \bigotimes_{j=1}^{n} Z^{\delta_{ij}}.$$

They satisfy the relations $[X_i, X_j] = 0 = [Z_i, Z_j]$; $[X_i, Z_j] = 2\delta_{ij}X_iZ_j$. However, we can choose other generating sets $\{x_i, z_i \mid 1 \leq i \leq n\}$ for $\mathcal{P}_n$ that satisfy similar commutation relations, *i.e.*, $[x_i, x_j] = 0 = [z_i, z_j]$ and $[x_i, z_j] = 2\delta_{ij}x_iz_j$. These operators may act nontrivially on many qubits. We often refer to the pair of operators x_i, z_i that satisfy the commutation relations similar to the Pauli operators as a *hyperbolic pair*. Given an $[[n,k,r,d]]$ code we could view the state space of the physical n qubits as that of n virtual qubits on which these x_i, z_i act as X and Z operators. In particular k of these virtual qubits are the logical qubits and r of them gauge qubits. The usefulness of these operators is that we can specify the structure of the stabilizer, the gauge group

[4] We allow some freedom in the primary generators, in that instead of $I_{s'}$ in equation (7), we allow it be an upper triangular matrix also.

and the encoded operators. The following lemma makes this specification precise.

Lemma 11. *Let Q be an $[[n,k,r,d]]_2$ subsystem code with gauge group, G and stabilizer S. Denote the encoded operators by $\overline{X}_i, \overline{Z}_i$, $1 \leq i \leq k$, where $[\overline{X}_i, \overline{X}_j] = 0 = [\overline{Z}_i, \overline{Z}_j]; [\overline{X}_i, \overline{Z}_j] = 2\delta_{ij}\overline{X}_i\overline{Z}_j$. Then there exist operators $\{x_i, z_i \in \mathcal{P}_n \mid 1 \leq i \leq n\}$ such that*

i) $S = \langle z_1, z_2, \ldots, z_s \rangle$,
ii) $G = \langle S, z_{s+1}, x_{s+1}, \ldots, z_{s+r}, x_{s+r}, Z(\mathcal{P}_n) \rangle$,
iii) $C_{\mathcal{P}_n}(S) = \langle G, \overline{X}_1, \overline{Z}_1, \ldots, \ldots, \overline{X}_k, \overline{Z}_k \rangle$,
iv) $\overline{X}_i = x_{s+r+i}$ and $\overline{Z}_i = z_{s+r+i}$, $1 \leq i \leq k$,

where $[z_i, z_j] = [x_i, x_j] = 0; [x_i, z_i] = 2\delta_{ij}x_i z_i$. Further, S defines an $[[n, k+r]]$ stabilizer code encoding into the same space as the subsystem code and its encoded operators are given by $\{x_{s+1}, z_{s+1}, \ldots, x_{s+r}, z_{s+r}, \overline{X}_1, \overline{Z}_1, \ldots, \overline{X}_k, \overline{Z}_k\}$

Proof: See [32] for proof on the structure of the groups. Let $Q = A \otimes B$, then $\dim A = 2^k$ and $\dim B = 2^r$. From Corollary 4 we know that the projector defined by S also projects onto Q (which is 2^{k+r}-dimensional) and therefore it defines an $[[n, k+r]]$ stabilizer code. From the definition of the operators x_i, z_i and $\overline{X}_i, \overline{Z}_i$ and the fact that $C_{\mathcal{P}_n}(S) = \langle S, x_{s+1}, z_{s+1}, \ldots, x_{s+r}, z_{s+r}\overline{X}_1, \overline{Z}_1, \ldots, \overline{X}_k, \overline{Z}_k, Z(\mathcal{P}_n) \rangle$ we see that x_i, z_i, for $s+1 \leq i \leq r$ act like encoded operators on the gauge qubits, while $\overline{X}_i, \overline{Z}_i$ continue to be the encoded operators on the information qubits. Together they exhaust the set of $2(k+r)$ encoded operators of the $[[n, k+r]]$ stabilizer code. ■

We observe that the logical operators of the subsystem code are also logical operators for the underlying stabilizer code. So if the stabilizer code and the subsystem code have the same logical all zero state, then Lemma 11 suggests that in order to encode the subsystem code, we can treat it as stabilizer code and use the same techniques to encode. If the logical all zero code word was the same for both the codes, then because they have the same logical operators we can encode any given input to the same logical state in both cases. Using linearity we could then encode any arbitrary state. Encoding the all zero state seems to be the key. Now, even in the case of the stabilizer codes, there is no unique all zero logical state. There are many possible choices. Given the encoded operators it is easy to define the logical all zero state as the following definition shows:

Definition 12. *A logical all zero state of an $[[n,k,r,d]]$ subsystem code is any state that is fixed by its stabilizer and k logical Z operators.*

This definition is valid in case of stabilizer codes also. This definition might appear a little circular. After all, we seem to have assumed the definition of the logical Z operators. Actually, this is a legitimate definition because, depending on the choice of our logical operators, we can have many choices of the logical all zero state. In case of the subsystem codes, this definition implies that the logical all zero state is fixed by $n-r$ operators, consequently it can be any state in that 2^r-dimensional subspace. If we consider the $[[n, k+r]]$ stabilizer code that is associated to the subsystem code, then its logical zero is additionally fixed by r more operators. So any logical zero of the stabilizer code is also a logical all zero state of the subsystem code. It follows that if we know how to encode the stabilizer code's logical all zero, we know how to encode the subsystem code. We are interested in more than merely encoding the subsystem code of course. We also want to leverage the gauge qubits to simplify and/or make the encoding process more robust. Perhaps a few examples will clarify the ideas.

C. Illustrative Examples

Consider the following $[[4,1,1,2]]_2$ subsystem code, with the gauge group G, stabilizer S and encoded operators given by L.

$$S = \begin{bmatrix} X & X & X & X \\ Z & Z & Z & Z \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$

$$G = \left[\begin{array}{cccc} X & X & X & X \\ Z & Z & Z & Z \\ \hline I & X & I & X \\ I & I & Z & Z \end{array}\right] = \left[\begin{array}{c} z_1 \\ z_2 \\ \hline x_3 \\ z_3 \end{array}\right].$$

The encoded operators of this code are given by

$$L = \begin{bmatrix} I & I & X & X \\ I & Z & I & Z \end{bmatrix} = \begin{bmatrix} \overline{X}_1 \\ \overline{Z}_1 \end{bmatrix}.$$

The associated $[[4,2]]$ stabilizer code has the following encoded operators.

$$T = \begin{bmatrix} I & X & I & X \\ I & I & X & X \\ I & I & Z & Z \\ I & Z & I & Z \end{bmatrix} = \begin{bmatrix} x_3 \\ \overline{X}_1 \\ z_3 \\ \overline{Z}_1 \end{bmatrix}.$$

It will be observed that the encoded X operators of $[[4,2]]$ are in a form convenient for encoding. We treat the $[[4,1,1,2]]$ code as $[[4,2]]$ code and encode it as in Figure 2. The gauge qubits are permitted to be in any state.

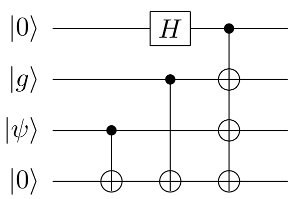


Fig. 2. Encoding the $[[4,1,1,2]]$ code (Gauge qubits can be in any state).

Assuming $g = a|0\rangle + b|1\rangle$, the logical states up to a normalizing constant are

$$\begin{aligned} |\overline{0}\rangle &= a(|0000\rangle + |1111\rangle) + b(|0101\rangle + |1010\rangle), \\ |\overline{1}\rangle &= a(|0011\rangle + |1100\rangle) + b(|0110\rangle + |1001\rangle). \end{aligned}$$

It can be easily verified that S stabilizes the above state and while the gauge group acts in a nontrivial fashion, the resulting states are still orthogonal. In this example we have encoded as if we were encoding the $[[4,2]]$ code. Prior to encoding the gauge qubits can be identified with physical qubits. After the encoding however such a correspondence between the physical qubits and gauge qubits does not necessarily exist in a nontrivial subsystem code. Since the encoded operators of the subsystem code are also encoded operators for the stabilizer code, we are guaranteed that the information is not encoded into the gauge subsystem.

As the state of gauge qubits is of no consequence, we can initialize them to any state. Alternatively, if we initialized them to zero, we can simplify the circuit as shown in Figure 3.

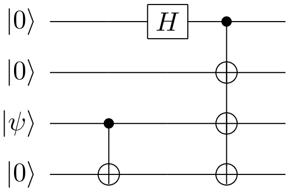


Fig. 3. Encoding the $[[4,1,1,2]]$ code (Gauge qubits initialized to zero).

The encoded states in this case are (again, the normalization factors are ignored)

$$\begin{aligned}|\bar{0}\rangle &= |0000\rangle + |1111\rangle, \\ |\bar{1}\rangle &= |0011\rangle + |1100\rangle.\end{aligned}$$

The benefit with respect to the previous version is that at the cost of initializing the gauge qubits, we have been able to get rid of all the encoded operators associated with them. This seems to be a better option than randomly initializing the gauge qubits. Because it is certainly easier to prepare them in a known state like $|0\rangle$, rather than implement a series of controlled gates depending on the encoded operators associated with those qubits.

At this point we might ask if it is possible to get both the benefits of random initialization of the gauge qubits as well as avoid implementing the encoded operators associated with them. To answer this question let us look a little more closely at the previous two encoding circuits for the subsystem codes. We can see from them that it will not work in general. Let us see why. If we initialize the gauge qubit to $|1\rangle$ instead of $|0\rangle$ in the encoding given in Figure 3, then the encoded state is

$$\begin{aligned}|\bar{0}\rangle &= |0100\rangle + |1011\rangle, \\ |\bar{1}\rangle &= |0111\rangle + |1000\rangle.\end{aligned}$$

Both these states are not stabilized by S, indicating that these states are not in the code space.

In general, an encoding circuit where it is simultaneously possible initialize the gauge qubits to random states and also avoid the encoded operators is likely to be having more complex primary generators. For instance, let us consider the following $[[4,1,1,2]]$ subsystem code:

$$S = \begin{bmatrix} X & Z & Z & X \\ Z & X & X & Z \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$

$$G = \left[\begin{array}{cccc} X & Z & Z & X \\ Z & X & X & Z \\ \hline Z & I & X & I \\ I & Z & Z & I \end{array}\right] = \left[\begin{array}{c} z_1 \\ z_2 \\ \hline x_3 \\ z_3 \end{array}\right].$$

The encoded operators of this code are given by

$$L = \begin{bmatrix} I & Z & I & X \\ Z & I & I & Z \end{bmatrix} = \begin{bmatrix} \overline{X}_1 \\ \overline{Z}_1 \end{bmatrix}.$$

The associated $[[4,2]]$ stabilizer code has the following encoded operators.

$$T = \begin{bmatrix} Z & I & X & I \\ I & Z & I & X \\ I & Z & Z & I \\ Z & I & I & Z \end{bmatrix} = \begin{bmatrix} x_3 \\ \overline{X}_1 \\ z_3 \\ \overline{Z}_1 \end{bmatrix}.$$

The encoding circuit for this code is given by

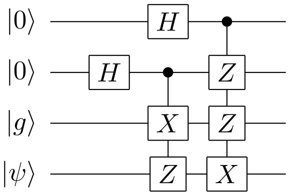


Fig. 4. Encoding $[[4,1,1,2]]$ code (Encoded operators for the gauge qubits are trivial and gauge qubits can be initialized to random states).

In this particular case, the gauge qubits (as well as the information qubits) do not require any additional encoding circuitry. In this case we can initialize the gauge qubits to any state we want. But, the reader would have observed we did not altogether end up with a simpler circuit. The primary generators are two as against one and the complexity of the encoded operators has been shifted to them. So even though we were able to get rid of the encoded operator on the gauge qubit and also get the benefit of initializing it to a random state, this is still more complex compared to either of encoders in Figures 2 and 3. Our contention is that it is better to initialize the gague qubits to zero state and not implement the encoded operators associated to them.

V. Encoding Subsystem Codes by Standard Form Method

The previous two examples might lead us to conclude that we can take the stabilizer of the given subsystem code and form the encoded operators by reducing the stablizer to its standard form and encode as if it were a stabilizer code. However, there are certain subtle points to be kept in mind. When we form the encoded operators we get $k + r$ encoded operators; we cannot from the stabilizer alone conclude which are the encoded operators on the information qubits and which on the gauge qubits. Put differently, these operators belong to the space $C_{\mathcal{P}_n}(S) \setminus S = GC_{\mathcal{P}_n}(G) \setminus SZ(\mathcal{P}_n)$. It is not guaranteed that they are entirely in $C_{\mathcal{P}_n}(G)$, *i.e.*, we cannot say if they act as encoded operators on the logical qubits. This implies that in general all these operators act nontrivially on both A and B. Consequently, we must be careful in choosing the encoded operators and the gauge group must be taken into account.

We give two slightly different methods for encoding subsystem codes. The difference between the two methods is subtle. Both methods require the gauge qubits to be initialized to zero. In the second method (see Algorithm 2) however, we can avoid the encoded operators associated to them. Under certain circumstances, we can also permit initialization to random states. In both algorithms 1 and 2 we assume the same notation as in Lemma 11.

Algorithm 1 Encoding subsystem codes–Standard form

```
Require: Stabilizer, $S = \langle z_1, \ldots, z_{n-k-r}\rangle$ and gauge
    group, $G = \langle S, x_{s+1}, z_{s+1}, \ldots, x_{s+r}, z_{s+r}, Z(\mathcal{P}_n)\rangle$ of the
    $[[n, k, r, d]]$ subsystem code.
Ensure: $[x_i, x_j] = [z_i, z_j] = 0$; $[x_i, z_j] = 2x_i z_i \delta_{ij}$
1: Form $S_A = \langle S, z_{s+1}, \ldots, z_{s+r}\rangle$, where $s = n - k - r$
2: Compute the standard form of $S_A$ as per Lemma 6
    $$S_A =_\pi \left[\begin{array}{ccc|ccc} I_{s'} & A_1 & A_2 & B & 0 & C \\ 0 & 0 & 0 & D & I_{s+r-s'} & E \end{array}\right]$$
3: Compute the encoded operators $\overline{X}_1, \ldots, \overline{X}_k$ as
    $$\left[\begin{array}{c} \overline{Z} \\ \hline \overline{X} \end{array}\right] =_\pi \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & A_2^t & 0 & I_k \\ \hline 0 & E^t & I_k & C^t & 0 & 0 \end{array}\right]$$
4: Encode using the primary generators of $S_A$ and $\overline{X}_i$ as
    encoded operators, see Lemma 10; all the other $(n - k)$
    qubits are initialized to $|0\rangle$.
```

Correctness of Algorithm 1. Since stabilizer $S_A \geq S$, the space stabilized by S_A is a subspace of the $A \otimes B$, the subspace stabilized by S. As $|S_A|/|S| = 2^r$, the dimension of the subspace stabilized by S_A is $2^{k+r}/2^r = 2^k$. Additionally, the generators $z_{s+1}, \ldots, z_{s+r}$ act trivially on A. The encoded operators as computed in the algorithm act nontrivially on A and give 2^k orthogonal states; thus we are assured that the information is encoded into A.

Let us encode the $[[9, 1, 4, 3]]$ Bacon-Shor code [3] using the method just proposed. The stabilizer and the gauge group are given[5] by

$$S = \left[\begin{array}{ccc|ccc|ccc} X & X & X & & & & X & X & X \\ & & & X & X & X & X & X & X \\ Z & & Z & Z & & Z & Z & & Z \\ & Z & Z & & Z & Z & & Z & Z \end{array}\right],$$

$$G = \left[\begin{array}{ccc|ccc|ccc} X & X & X & & & & X & X & X \\ & & & X & X & X & X & X & X \\ Z & & Z & Z & & Z & Z & & Z \\ & Z & Z & & Z & Z & & Z & Z \\ \hline & X & & & & & & X & \\ & & X & & & & & & X \\ & & & & X & & & X & \\ & & & & & X & & & X \\ \hline Z & & Z & & & & & & \\ & & & Z & & Z & & & \\ & Z & Z & & & & & & \\ & & & & Z & Z & & & \end{array}\right] = \left[\begin{array}{c} S \\ \hline G_x \\ \hline G_z \end{array}\right].$$

Let us form S_A by augmenting S with G_z. Then

$$S_A = \left[\begin{array}{ccc|ccc|ccc} X & X & X & & & & X & X & X \\ & & & X & X & X & X & X & X \\ Z & & Z & Z & & Z & Z & & Z \\ & Z & Z & & Z & Z & & Z & Z \\ \hline Z & & Z & & & & & & \\ & & & Z & & Z & & & \\ & Z & Z & & & & & & \\ & & & & Z & Z & & & \end{array}\right].$$

The encoded X and Z operators are $X_7X_8X_9$ and $Z_1Z_4Z_7$, respectively. After putting S_A in the standard form, and encoder for this code is given in Figure 5.

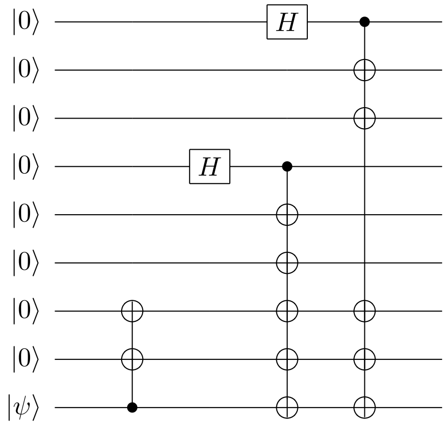


Fig. 5. Encoder for the $[[9, 1, 4, 3]]$ code. This is also an encoder for the $[[9, 1, 3]]$ code

If on the other hand we had formed S_A by adding G_x instead, then S_A would have been

$$S_A = \left[\begin{array}{ccc|ccc|ccc} X & & & & & & X & & \\ & X & & & & & & X & \\ & & X & & & & & & X \\ & & & X & & & X & & \\ & & & & X & & & X & \\ & & & & & X & & & X \\ Z & & Z & Z & & Z & Z & & Z \\ & Z & Z & & Z & Z & & Z & Z \end{array}\right].$$

[5] We do not show the identity.

The encoded operators remain the same. In this case the encoding circuit is given in Figure 6.

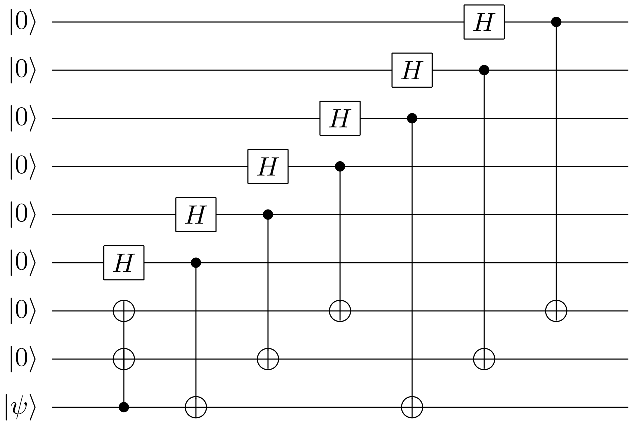


Fig. 6. Encoder for the $[[9,1,4,3]]$ code with fewer CNOT gates.

The circuit in Figure 6 has fewer CNOT gates, though the number of single qubit gates has increased. Since we expect the implementation of the CNOT gate to be more complex than the H gate, this might be a better choice. If on the other hand we interchanged the encoded X and Z operators, we could end up with a simpler circuit, see Figure 7, equivalent to the one proposed by Shor for the $[[9,1,3]]$ code.

In any case, this demonstrates that by exploiting the gauge qubits one can find ways to reduce the complexity of encoding circuit.

The gauge qubits provide a great degree of freedom in encoding. We consider the following variant on standard form encoding, where we try to minimize the the number of primary generators. This is not guaranteed to reduce the overall complexity, since that is determined by both the primary generators and the encoded operators. Fewer primary generators might usually imply encoded operators with larger complexity. In fact we have already seen, that in the case of $[[9,1,4,3]]_2$ code that a larger number of primary generators does not necessarily imply higher complexity. However, it has the potential for lower complexity.

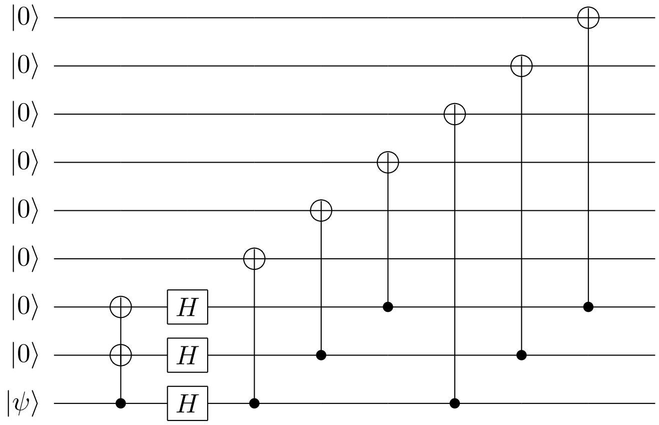


Fig. 7. Encoder for an equivalent $[[9,1,4,3]]$ code (logical X and Z interchanged).

Algorithm 2 Encoding subsystem codes–Standard form

```
Require: Stabilizer, $S = \langle z_1,\ldots,z_{n-k-r}\rangle$ and gauge group, $G = \langle S, x_{s+1}, z_{s+1},\ldots,x_{s+r},z_{s+r}, Z(\mathcal{P}_n)\rangle$ of the $[[n,k,r,d]]$ subsystem code.
Ensure: $[x_i,x_j] = [z_i,z_j] = 0$; $[x_i,z_j] = 2x_iz_i\delta_{ij}$
1: Compute the standard form of $S$ as per Lemma 6
     $S =_{\pi_1} \left[\begin{array}{ccc|ccc} I_{s'} & A_1 & A_2 & B & 0 & C \\ 0 & 0 & 0 & D & I_{s-s'} & E \end{array}\right]$
2: Form $S_A = \langle S, z_{s+1},\ldots,z_{s+r}\rangle$, where $s = n-k-r$
3: Compute the standard form of $S_A$ as per Lemma 6
     $S_A =_{\pi_2} \left[\begin{array}{ccc|ccc} I_l & F_1 & F_2 & G_1 & 0 & G_2 \\ 0 & 0 & 0 & D' & I_{s+r-l} & H \end{array}\right]$
4: Compute the encoded operators $\overline{X}_1,\ldots,\overline{X}_k$ as
     $\left[\begin{array}{c} \overline{Z} \\ \overline{X} \end{array}\right] =_{\pi_2} \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & F_2^t & 0 & I_k \\ \hline 0 & H^t & I_k & G_2^t & 0 & 0 \end{array}\right]$
5: Encode using the primary generators of $S$ and $\overline{X}_i$ as encoded operators, accounting for $\pi_1$ and $\pi_2$, see Lemma 10; all the other $(n-k)$ qubits are initialized to $|0\rangle$.
```

The main difference in the second method comes in lines 1 and 5. We encode using the primary generators of the stabilizer of the subsystem code instead of the augmented stabilizer. The encoded operators however remain the same as before.

Correctness of Algorithm 2. The correctness of this method lies in the observation we made earlier (see discussion following Definition 12), that any logical all zero state of the stabilizer code is also a logical all zero of the subsystem code and the fact that both share the encoded operators on the encoded qubits.

Remark 13. *The permutation π_2 in Algorithm 2 can be restricted to the last $n-s'$ columns, since while adjoining the additional r generators to S, we could take it to be in the standard form.*

The encoded operators are given modulo the elements of the gauge group as in Algorithm 1, which implies that the their action might be nontrivial on the gauge qubits. The benefit of the second method is when S and S_A have different number of primary generators. The following aspects of both the methods are worth highlighting.

1) The gauge qubits must be initialized to $|0\rangle$ in both methods.
2) In Algorithm 1, the number of primary generators of S and S_A can be different leading to a potential increase in complexity compared to encoding with S.
3) In both methods, the encoded operators as computed are modulo S_A. Consequently, the encoded operators might act nontrivially on the gauge qubits.

VI. Encoding Subsystem Codes by Conjugation Method

The other benefit of subsystem codes is the random initialization of the gauge qubits. We now give circuits where we can encode the subsystem codes to realize this benefit. But instead of using the standard form method we will use the conjugation method proposed by Grassl *et al.*, [16] for stabilizer codes. After briefly reviewing this method we shall show how it can be modified for encoding subsystem codes.

The conjugation encoding method can be understood as follows. It is based on the idea that the Clifford group acts transitively on the Pauli error group. Therefore, we can transform the stabilizer of an arbitrary $[[n, k, d]]$ code to the trivial stabilizer given by $\langle Z_1, \ldots, Z_{n-k}\rangle$. Additionally, we can also transform the encoded operators $\overline{X}_i$, $\overline{Z}_i$ to X_{n-k+i}, Z_{n-k+i} for $1 \leq i \leq k$. Put differently, we transform the stabilizer matrix of any $[[n, k, d]]$ stabilizer code into the matrix $(00|I_{n-k}0)$. The associated encoded $\overline{X}$ and $\overline{Z}$ operators are given by $(0I_k|00)$ and $(00|0I_k)$ respectively. For a code with this stabilizer matrix the encoding is trivial. We simply map $|\psi\rangle$ to $|0\rangle^{\otimes^{n-k}}|\psi\rangle$. Here we give a sketch of the method for the binary case, the reader can refer to [16] for details.

Assume that the stabilizer matrix is given by S. Then we shall transform it into $(00|I_{n-k}0)$ using the following sequence of operations.

$$(X|Z) \mapsto (I_{n-k}0|0) \mapsto (00|I_{n-k}0). \qquad (9)$$

This can be accomplished through the action of $H = \left[\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right]$, $P = \left[\begin{smallmatrix} 1 & 0 \\ 0 & i \end{smallmatrix}\right]$ and CNOT gates on the Pauli group under conjugation. The H gate acting on the ith qubit on $(a|b) = (a_1, \ldots, a_n|b_1, \ldots, b_n)$ transforms it as

$$(a|b) \overset{H_i}{\mapsto} (a_1, \ldots, \mathbf{b_i}, \ldots, a_n|b_1, \ldots, \mathbf{a_i}, \ldots, b_n). \qquad (10)$$

These modified entries have been highlighted for convenience. The phase gate P on the ith qubit transforms $(a_1, \ldots, a_n|b_1, \ldots, b_n)$ as

$$(a|b) \overset{P_i}{\mapsto} (a_1, \ldots, \mathbf{a_i}, \ldots, a_n|b_1, \ldots, \mathbf{a_i} + \mathbf{b_i}, \ldots, b_n). \qquad (11)$$

We denote the CNOT gate with the control on the ith qubit and the target on the jth qubit by CNOT^{i,j}. The action of the CNOT^{i,j} gate on $(a_1, \ldots, a_n|b_1, \ldots, b_n)$ is to transform it to

$$(a_1, \ldots, a_{j-1}, \mathbf{a_j} + \mathbf{a_i}, \ldots, a_n|b_1, \ldots, b_{i-1}, \mathbf{b_i} + \mathbf{b_j}, \ldots, b_n).$$

Note that the jth entry is changed in the X part while the ith entry is changed in the Z part. For example, consider

$$(1,0,0,1,0|0,1,1,0,0) \overset{\text{CNOT}^{1,4}}{\mapsto} (1,0,0,\mathbf{0},0|0,1,1,0,0),$$
$$(1,0,0,1,0|0,1,1,1,0) \overset{\text{CNOT}^{1,4}}{\mapsto} (1,0,0,\mathbf{0},0|\mathbf{1},1,1,1,0).$$

Based on the action of these three gates we have the following lemmas to transform error operators.

Lemma 14. *Assume that we have a error operator of the form $(a_1, \ldots, a_n|b_1, \ldots, b_n)$. Then we apply the following gates on the ith qubit to transform the stabilizer, transforming (a_i, b_i) to (α, β) as per the following table.*

(a_i, b_i)	*Gate*	(α, β)
(0,0)	I	*(0,0)*
(0,1)	H	*(1,0)*
(1,0)	I	*(1,0)*
(1,1)	P	*(1,0)*

Let $\bar{x}$ denote $1 + x \bmod 2$, then the transformation to $(a'_1, \ldots, a'_n|0, \ldots, 0)$ is achieved by

$$\bigotimes_{i=1}^{n} H^{\bar{a}_i b_i} P^{a_i b_i}.$$

For example, consider the following generator $(1,0,0,1,0|0,1,1,1,0)$. This can be transformed to $(1,1,1,1,0|0,0,0,0,0)$ by the application of $I \otimes H \otimes H \otimes P \otimes I$.

Lemma 15. *Let e be an error operator of the form $(a_1, \ldots, a_i = 1, \ldots, a_n|0, \ldots, 0)$. Then e can be transformed to $(0, \ldots, 0, a_i = 1, 0, \ldots, 0|0, \ldots, 0)$ by*

$$\prod_{j=1, i\neq j}^{n} \left[\text{CNOT}^{i,j}\right]^{a_j}.$$

As an example consider $(1,1,1,1,0|0,0,0,0,0)$, this can be transformed to $(0,1,0,0,0|0,0,0,0,0)$ by

$$\text{CNOT}^{2,1} \cdot \text{CNOT}^{2,3} \cdot \text{CNOT}^{2,4}.$$

The first step involves making the Z portion of the stabilizer matrix all zeros. This is achieved by single qubit operations consisting of H and P performed on each row one by one.

Note that we must also modify the other rows of the stabilizer matrix according to the action of the gates applied.

Once we have a row of stabilizer matrix in the form $(a|0)$, where a is nonozero we can transform it to the form $(0, \ldots, 0, a_i = 1, 0, \ldots, 0|0)$ by using CNOT gates. Thus it is easy to transform $(X|Z)$ to $(I_{n-k}0|0)$ using CNOT, P and H gates. The final transformation to $(0|I_{n-k}0)$ is achieved by using H gates on the first $n-k$ qubits. At this point the stabilizer matrix has been transformed to a trivial stabilizer matrix which stabilizes the state $|0\rangle^{\otimes^{n-k}}|\psi\rangle$. The encoded operators are $(0I_k|0)$ and $(0|0I_k)$. Let T be the sequence of gates applied to transform the stabilizer matrix to the trivial stabilizer matrix. Then T applied in the reverse order to $|0\rangle^{\otimes^{n-k}}|\psi\rangle$ gives the encoding circuit for the stabilizer code.

Now we shall use the conjugation method to encode the subsystem codes. The main difference with respect to [16] is that instead of considering just the stabilizer we need to consider the entire gauge group. Let the gauge group be $G = \langle S, G_Z, G_X, Z(\mathcal{P}_n)\rangle$, where $G_Z = \langle z_{s+1}, \ldots, z_{s+r}\rangle$, and $G_X = \langle x_{s+1}, \ldots, x_{s+r}\rangle$. The idea is to transform the gauge group as follows.

$$G = \left[\begin{array}{c} S \\ \hline G_Z \\ \hline G_X \end{array}\right] \mapsto \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & I_s & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & I_r & 0 \\ \hline 0 & I_r & 0 & 0 & 0 & 0 \end{array}\right]. \qquad (12)$$

At this point the gauge group has been transformed to a group with trivial stabilizer and trivial encoded operators for the gauge qubits and the encoded qubits. The sequence of gates required to achieve this transformation in the reverse order will encode the state $|0\rangle^{\otimes^s} |\phi\rangle |\psi\rangle$. The state $|\phi\rangle$ corresponds to the gauge qubits and it can be initialized to any state, while $|\psi\rangle$ corresponds to the input.

Algorithm 3 Encoding subsystem codes–Conjugation method

```
Require: Gauge group $G$ of the $[[n,k,r,d]]$ subsystem code.
    $G = \langle S, G_Z, G_X, Z(\mathcal{P}_n)\rangle$, where $S = \{z_1, \ldots, z_{n-k-r}\}$,
    $G_Z = \{z_{s+1}, \ldots, z_{s+r}\}$, and $G_X = \{x_{s+1}, \ldots, x_{s+r}\}$.
Ensure: $[x_i, x_j] = [z_i, z_j] = 0$; $[x_i, z_j] = 2x_i z_i \delta_{ij}$
1:  Assume that $G$ is in the form $G = \left[\begin{array}{c} S \\ \hline G_Z \\ \hline G_X \end{array}\right]$.
2:  for all $i = 1$ to $s + r$ do
3:      Transform $z_i$ to $z'_i = (a_1, \ldots, a_i = 1, \ldots, a_n|0)$ using Lemma 14
4:      Transform $z'_i$ to $(0, \ldots, a_i = 1, \ldots, 0|0)$ using Lemma 15
5:      For $i \leq s$ perform Gaussian elimination on column $i$ for rows $j > i$
6:  end for
7:  Apply $H$ gate on each qubit $i = 1$ to $i = s + r$
8:  for all $i = s + 1$ to $s + r$ do
9:      Transform $x_i$ to $x'_i = (a_1, \ldots, a_n|0, \ldots, 0)$ using Lemma 14
10:     Transform $x'_i$ to $(0, \ldots, a_i = 1, \ldots, 0|0)$ using Lemma 15
11:     Perform Gaussian elimination on column $i$ for rows $j > i$
12: end for
```

In the above algorithm, we assume that whenever a row of G is transformed according to Lemma 14 or 15, all the other rows are also transformed according to the transformation applied. The lines 8–12 are essentially responsible for the tolerance to initialization errors on the gauge qubits.

Correctness of Algorithm 3. The correctness of the algorithm is straightforward. As G has full rank of $n - k + r$, for each row of G, we will be able to find some nonzero pair (a, b) so that the transformation of S and G_Z to $(I_{s+r}0|0)$ (lines 2–6) can be achieved. After line 7, when S and G_Z are in the form $(0|I_{s+r}0)$, the rows in G_X are in the form

$$\left[\begin{array}{ccc|ccc} 0 & A & B & 0 & C & D \end{array}\right]. \tag{13}$$

The first $n-k-r$ columns of the (transformed) G_X are all zero because they must commute with $(0|I_s0)$, the elements of the transformed stabilizer, while the remaining zero columns are due to Gaussian elimination. The submatrix A must have rank r, otherwise at this point one of the rows of G_X commutes with all the rows of G_Z and the condition that there are r hyperbolic pairs is violated. In fact we must have $A = I_r$. Therefore it is possible to transform equation (13) to the form $(0I_r0|0)$. Thus Algorithm 3 transforms G to the form given in equation (12). The encoded operators for this gauge group are clearly $(0I_k|0)$ and $(0|0I_k)$. The transformations in reverse order encode the subsystem code.We conclude with a simple example that illustrates the process.

Example. Consider the following $[[4,1,1,2]]$ code. Let the gauge group G, stabilizer S be given as

$$S = \left[\begin{array}{cccc} X & X & X & X \\ Z & Z & Z & Z \end{array}\right] = \left[\begin{array}{c} z_1 \\ z_2 \end{array}\right],$$

$$G = \left[\begin{array}{cccc} X & X & X & X \\ Z & Z & Z & Z \\ \hline I & I & Z & Z \\ I & X & I & X \end{array}\right] = \left[\begin{array}{c} z_1 \\ z_2 \\ \hline x_3 \\ z_3 \end{array}\right].$$

In matrix form G can be written as

$$G = \left[\begin{array}{cccc|cccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right].$$

The transformations consisting of $T_1 = \text{CNOT}^{1,2}\text{CNOT}^{1,3}\text{CNOT}^{1,4}$ followed by $T_2 = I \otimes H \otimes H \otimes H$ maps G to

$$\overset{T_1}{\mapsto} \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right]$$

$$\overset{T_2}{\mapsto} \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right].$$

Now transform the second row using $T_3 = \text{CNOT}^{2,3}\text{CNOT}^{2,4}$. Then transform using $T_4 = \text{CNOT}^{4,3}$. We get

$$\overset{T_3}{\mapsto} \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

$$\overset{T_4}{\mapsto} \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

Applying $T_5 = H \otimes H \otimes I \otimes H$ gives us

$$\overset{T_5}{\mapsto} \left[\begin{array}{cccc|cccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

We could have chosen $T_5 = H \otimes H \otimes I \otimes I$, since the effect of H on the fourth qubit is trivial. The complete circuit is given in Figure 8.

The algorithm guarantees (due to lines 8–12) that just prior to encoding the gauge qubits can be identified with physical qubits. Since we do not care about the state of the gauge qubits, we can tolerate arbitrary errors on the physical qubits

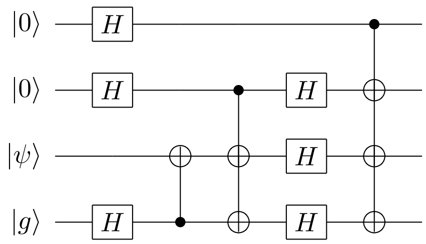


Fig. 8. Encoding $[[4, 1, 1, 2]]$ code by conjugation method.

at this point. In the present case $|g\rangle$. By switching the target and control qubits of the CNOT gates in T_3 and T_4 we can show that this circuit is equivalent to circuit shown in Figure 9.

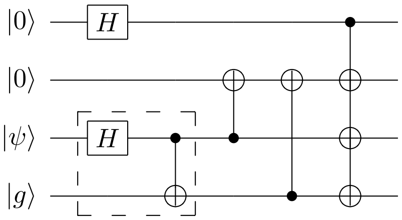


Fig. 9. Encoding $[[4, 1, 1, 2]]$ code by conjugation method.

It is instructive to compare the circuit in Figure 9 with the one given earlier in Figure 2. The dotted lines show the additional circuitry. Since the gauge qubit can be initialized to any state, we can initialize $|g\rangle$ to $|0\rangle$, which then gives the following logical states for the code.

$$\left|\bar{0}\right\rangle = |0000\rangle + |1111\rangle + |0011\rangle + |1100\rangle\,, \quad (14)$$
$$\left|\bar{1}\right\rangle = |0000\rangle + |1111\rangle - |0011\rangle - |1100\rangle\,. \quad (15)$$

It will be observed that $IIXX$ acts as the logical Z operator while $IZIZ$ acts as the logical X operator. We could flip these logical operators by absorbing the H gate into $|\psi\rangle$. If we additionally initialize $|g\rangle$ to $|0\rangle$, we will see that the two CNOT gates on the second qubit can be removed. The circuit then simplifies to the circuit shown in Figure 10.

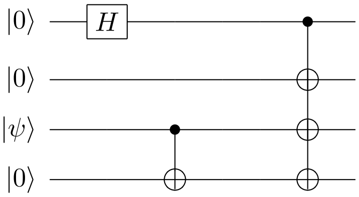


Fig. 10. Encoding $[[4, 1, 1, 2]]$ code by conjugation method – optimized.

This is precisely, the same circuit that we had arrived earlier in Figure 3 using the standard form method. The preceding example provides additional evidence in the direction that it is better to initialize the gauge qubits to zero and avoid the encoding operators on them.

Two important optimizations are possible in Algorithm 3. Firstly, we could choose to initialize the gauge qubits to all zero and then we could dispense with the lines 8–12. Secondly, we could also dispense with line 4 for $s + 1 \leq i \leq s + r$ when the gauge qubits are initialized to all zero. The first optimization trades off random initialization with all zero initialization. The second one will lead to a further reduction in CNOT gates.

VII. Summary

We now briefly summarize the main results of this paper. The recent activity in subsystem codes was motivated to a great extent by the promise of efficient error recovery schemes. A largely neglected aspect in the study of subsystem codes was that of encoding schemes for these codes. In this paper we have argued that subsystem codes afford benefits in this area as well that are worth studying. Given a subsystem code, one is concerned with how to encode the bare quantum information. We showed here that there are some benefits to be gained, namely, that some of the qubits are resistant to noise even in an unencoded stage. We then showed that this tolerance can be traded for reduced complexity.

Specifically, we showed first that subsystem codes can be encoded using the techniques used for stabilizer codes. In the process, we also showed how to encode Clifford codes, a class of codes that generalize the stabilizer codes and which are useful in the construction of subsystem codes. We then focussed on giving explicit encoding circuits for the binary subsystem codes. In particular, we have considered two methods for encoding stabilizer codes—the standard form method and the conjugation method. Both these methods can be easily generalized to the nonbinary codes.

We also showed the gauge degrees of freedom can be exploited to tolerate initialization errors on some qubits or a reduced complexity of encoding. While the standard form method explored here required us to initialize the gauge qubits to zero, it admits two variants and seems to have the potential for lower complexity; the exact gains being determined by the actual codes under consideration. The conjugation method allows us to initialize the gauge qubits to any state. The disadvantage seems to be the increased complexity of encoding. It must be emphasized that the standard form method is equivalent to the conjugation method and it is certainly possible to use this method to encode subsystem codes so that the gauge qubits can be initialized to arbitrary states. However, it appears to be a little more cumbersome and for this reason we have not investigated this possibility in this paper.

Stabilizer codes can also be encoded using teleportation. We expect that gauge qubits can be exploited even in this method to reduce its complexity. It would be interesting to investigate fault tolerant encoding schemes for subsystem codes exploiting the gauge qubits.

Acknowledgment

We thank Markus Grassl for very helpful discussions. We also thank the referees for their comments which helped improve the presentation of the paper. This research was supported by NSF CAREER award CCF 0347310 and NSF grant CCF 0622201. P.S. was also supported by grants from NSERC, MITACS and CIFAR.

References

[1] P. Aliferis and A. W. Cross. Sub-system fault tolerance with the bacon-shor code. quant-ph/0610063, 2006.

[2] S. A. Aly, A. Klappenecker, and P. K. Sarvepalli. Subsystem codes. In *Forty-Fourth Annual Allerton Conference on Communication, Control, and Computing, Illinois, USA*, 2006.

[3] D. Bacon. Operator quantum error correcting subsystems for self-correcting quantum memories. *Phys. Rev. A*, 73(012340), 2006.

[4] C.H. Bennett, D.P. DiVincenzo, J.A. Smolin, and W.K. Wootters. Mixed state entanglement and quantum error correction. *Physical Review A*, 54:3824–3851, 1996.

[5] A.R. Calderbank, E.M. Rains, P.W. Shor, and N.J.A. Sloane. Quantum error correction via codes over GF(4). *IEEE Trans. Inform. Theory*, 44:1369–1387, 1998.

[6] H. Chen. Some good quantum error-correcting codes from algebraic-geometric codes. *IEEE Trans. Inform. Theory*, 47:2059–2061, 2001.

[7] H. Chen, S. Ling, and C. Xing. Asymptotically good quantum codes exceeding the Ashikhmin-Litsyn-Tsfasman bound. *IEEE Trans. Inform. Theory*, 47:2055–2058, 2001.

[8] R. Cleve and D. Gottesman. Efficient computations of encodings for quantum error correction. *Phys. Rev. A*, 56(1):76–82, 1997.

[9] R. Cleve, D. Gottesman, and H.-K. Lo. How to share a quantum secret. *Phys. Rev. Lett.*, 83(3):648–651, 1999.

[10] K. Feng. Quantum error-correcting codes. In *Coding Theory and Cryptology*, pages 91–142. World Scientific, 2002.

[11] K. Feng, S. Ling, and C. Xing. Asymptotic bounds on quantum codes from algebraic geometric codes. *IEEE Trans. Inform. Theory*, 52(3):986–991, 2006.

[12] D. Gottesman. A class of quantum error-correcting codes saturating the quantum Hamming bound. *Phys. Rev. A*, 54:1862–1868, 1996.

[13] D. Gottesman. Stabilizer codes and quantum error correction. Caltech Ph. D. Thesis, eprint: quant-ph/9705052, 1997.

[14] D. Gottesman. Theory of quantum secret sharing. *Phys. Rev. A*, 61(042311), 2000.

[15] D. Gottesman. An introduction to quantum error correction and fault-tolerant quantum computation. arXiv:0904.2557, 2009.

[16] M. Grassl, M. Rötteler, and T. Beth. Efficient quantum circuits for non-qubit quantum error-correcting codes. *Internat. J. Found. Comput. Sci.*, 14(5):757–775, 2003.

[17] G. James and M. Liebeck. *Representations and Characters of Groups*. Cambridge University Press, Cambridge, 2001.

[18] K. Julia. Approaches to quantum error correction. quant-ph/0612185, 2006.

[19] A. Ketkar, A. Klappenecker, S. Kumar, and P.K. Sarvepalli. Nonbinary stabilizer codes over finite fields. *IEEE Trans. Inform. Theory*, 52(11):4892–4914, 2006.

[20] A. Klappenecker and M. Rötteler. Beyond stabilizer codes II: Clifford codes. *IEEE Trans. Inform. Theory*, 48(8):2396–2399, 2002.

[21] A. Klappenecker and P. K. Sarvepalli. Clifford code constructions of operator quantum error-correcting codes. *IEEE Trans. Inform. Theory*, 54(12):5760–5765, 2008.

[22] E. Knill. Group representations, error bases and quantum codes. Los Alamos National Laboratory Report LAUR-96-2807, 1996.

[23] E. Knill. Non-binary unitary error bases and quantum codes. Los Alamos National Laboratory Report LAUR-96-2717, 1996.

[24] E. Knill. On protected realizations of quantum information. Eprint: quant-ph/0603252, 2006.

[25] E. Knill and R. Laflamme. A theory of quantum error–correcting codes. *Physical Review A*, 55(2):900–911, 1997.

[26] D. Kretschmann, D. W. Kribs, and R. W. Spekkens. Complementarity of private and correctable subsystems in quantum cryptography and error correction. *Phys. Rev. A*, 78:032330, 2008.

[27] D. W. Kribs. A brief introduction to operator quantum error correction. *Contemporary Mathematics, American Mathematical Society*, 414:27–34, 2005. Eprint: math/0506491.

[28] D. W. Kribs, R. Laflamme, and D. Poulin. Unified and generalized approach to quantum error correction. *Phys. Rev. Lett.*, 94(180501), 2005.

[29] D. W. Kribs, R. Laflamme, D. Poulin, and M. Lesosky. Operator quantum error correction. Eprint: quant-ph/0504189, 2005.

[30] W.J. Martin. A physics-free introduction to quantum error correcting codes. *Util. Math.*, pages 133–158, 2004.

[31] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, 2000.

[32] D. Poulin. Stabilizer formalism for operator quantum error correction. *Phys. Rev. Lett.*, 95(230504), 2005.

[33] D. Poulin. Operator quantum error correction: An overview. [Online] http://www.physique.usherbrooke.ca/d̃poulin/ Documents/OQEC.pdf, 2006.

[34] P. K. Sarvepalli and A. Klappenecker. Encoding subsystem codes with and without noisy gauge qubits. In *Proc. of the Third International Conference on Quantum, Nano and Micro Technologies*, pages 48–53, 2009.

[35] P. Shor and J. Preskill. Simple proof of security of the BB84 quantum key distribution protocol. *Phys. Rev. Lett.*, 85(441), 2000.

Security and User Aspects in the Design of the Future Trusted Ambient Networked Systems

Seppo Heikkinen[1], Kari Heikkinen[2], Sari Kinnari[1]

[1]Department of Communications Engineering
Tampere University of Technology
Tampere, Finland
firstname.lastname@tut.fi

[2]Communications Software Laboratory
Lappeenranta University of Technology
Lappeenranta, Finland
firstname.lastname@lut.fi

***Abstract*—Research visions of ambient computing promise seamless co-existence of technology and user in such a way that the environment adapts to user context. This adaptivity also means more extensive information disclosure, hence the security concerns become paramount. While new architectures should be able to provide security as a basic feature, they also need to take into account the way users behave and experience the system, as users are not likely to be interested in technical details and configurations, but instead in the added value they can get. Thus, if the users find the system too complex to use, they might find it hard to trust and not adopt it. Therefore, usability and user experience issues have to be considered tightly along with security and they need to be in the design process right from the start. In this article we discuss how security and user design aspects within the ubiquitous future environment can be used to enhance both the security and user experience in the creation of the trusted communication services.**

Keywords-Ambient Networks, design, security, trust, user experience

I. INTRODUCTION

The future holds much promise for the ordinary user of the communication systems. Connectivity is available everywhere and the ambient intelligence around the user takes automatically care of the complexities of the technology and concentrates on bringing services to the user at the right moment and at the right place. Thus, the user should not need to ask how, why, what, when, and where as we take a plunge towards ubiquitous media society.

While this vision certainly sounds attractive, the diversity of the environment sets challenging requirements for providing a concise user experience and enabling secure and flexible interworking between the available heterogeneous networks and composed services. Security is imperative to make the users trust and use the systems, but the design has to take into account the user experience factors as the complex configuration of security measures too often leads to a situation, where these measures are not used.

In this article we extend the work presented in [1] by further elaborating the user perception of security and user experience for the successful design of future networked systems. We discuss the presented ambient visions and base many of the technical concepts on the findings of Ambient Networks (AN), a partly EU funded project [2], in which one of the authors participated. The motivation is not to list all the possible results of the project, but to concentrate on the relevant security topics discussed within the project and show how they can be used to enhance the security of the ubiquitous environment [3]. As AN mostly concentrated on the network level with an objective of creating a scalable and affordable mobile communication system for heterogeneous environments, we also bring the user and user experience factors into the picture in order to show that it can be challenging for the technical solutions to respond to the decisions made by the user. Thus, we are trying to determine, whether there is in this setting any common ground of mutual benefits between these different viewpoints, which often have contradictory goals.

The article is organised as follows. In the next section we discuss the evolving ubiquitous environment. In the third section we provide a short introduction to user-centric design methodology. The fourth section presents different aspects of trust within the context of our work. The fifth section considers various technical guidelines and principles that the future network design should take into account in order to ensure the security of the systems. The sixth section considers both security and usability factors and the benefits of their combination from the user experience point of view. Additional discussion is provided in section seven. The final section concludes the paper.

II. EVOLUTION TOWARD AMBIENT ENVIRONMENT

Different kinds of terms, such as ambient intelligence, ambient networking and ubiquitous computing, have been introduced to portray the visions of enhanced interaction between the users and the surrounding technology. One vision lists the following as key requirements [4]:

- Unobtrusive hardware
- Seamless communication
- Dynamic and distributed device networks
- Natural feeling human interfaces
- Dependability and security

We do not claim this to be a conclusive list nor does the transition to this kind of system take place overnight. We would like to, however, emphasise the dynamic interaction aspects (both with technology and other users) and concentrate on networking and users with security viewpoint. It could be further noted that it can be claimed that ubiquitous computing already is here, even though not in a very seamless nor unobtrusive fashion, whereas the "clean" ambient vision is something that is always "just around the corner" [5]. This can be actually seen in the various Future Internet research activities, which basically try to address similar issues. However, on less physical scale, this mixing of technology and social world is, in fact, already quite prominent in the proliferation of social online communities, where, e.g., certain user threats have already become an issue.

A. Network level aspects

Forward looking projects, such as Ambient Networks, envisage a drastic change in the future landscape of networking as the user is put in the focus [6]. The availability and internetworking of heterogeneous networks provide the possibility of getting seamless connectivity and services in a ubiquitous manner. This ubiquity sets requirements for the terminal devices in terms of adaptability and usability as people also have the possibility to use different devices within a session, i.e., the users are less device dependant. Also, one should not forget that in this kind of versatile environment the security will play even more important part as the mobile users no longer clearly separate the time they are on- or off-line and possibilities to interact with various previously unknown parties are vastly different.

The user context affects the available services as the surrounding networking environment adapts to the needs of the user, which could be related, for instance, to the offered prices and quality. Various pieces of information are made available to the networks in order to provide a concise user experience, thus leading to privacy issues. This also brings user and network levels closer to each other as service specific network overlays are introduced and cross layer principles are applied for enhanced performance.

In traditional use scenarios the users have placed their trust on the operators, either consciously or subconsciously. It has been rather clear that the big telecom operators provide the communication services and the people have static relationships with them, be it in the form of post- or pre-payment. In the future this will change as there will be more players entering the market. In essence, everybody could be an operator providing access through their own networks as the technical development enables even a single node, i.e., a networked device, to provide access services in automated fashion. Even though some may have idealistic views about offering services to anybody for free, to most there still will be clear motivation to get compensation for the provision of their resources. This calls for solutions to ensure that every party gets what they have agreed to. New business models and roles will emerge, and the value chains transform into more complex value nets. User identity will be a valuable commodity.

Single nodes will exhibit more intelligence and can provide access services to other, perhaps slightly more limited devices. Thus, everything can be considered to be a network. Hence, they interwork with other networks and compose into even larger entities with common control plane, which hides the differences resulting from the specific technological domains and allows the controlled sharing of resources [7].

B. User and service level threats

From the user perspective one of the major issues in the ambient environment is the user privacy. There will be plenty of information available about the user as information is mediated and recorded, and the lifespan of information availability is vastly different. Hence, it is easier to target attacks against a particular user. Information availability is already evident in the emergence of social networking and the way people freely give out information about themselves and the people they know, providing avenues for identity theft. Think, for example, the amount of information people publish about themselves in services such as Facebook with no real guarantee about the privacy of the data [8]. The emergence of virtual worlds and online games and their accompanying side economies provide yet additional ways of cheating the user [9]. One can argue, though, that the strictest privacy would mean zero personal information transfer; i.e., all personal data would lie in personal trusted device(s) (PTD), and no data would be collected, e.g., by the operator. Such devices naturally would make attractive targets of trickery, thus they require strong security solutions.

In a sense these social networking sites provide an application framework, which form a limited overlay network with their own semantic properties. While they currently work on application level, the work done on developing service specific overlays for network level will reduce the gap [10]. Thus, it becomes increasingly more important who is controlling the overlay and how the collected information is used. When the borders become blurred, it can be challenging for the user to know, which action has what sort of privacy sensitive consequences. Especially if the user is presented with opt-out policy as default action, i.e., in order to restrict the information disclosure the user has to actively know how to configure the system right from the start.

The information about the users can leak in various other ways, as well. The existence of caches and archival services ensure that the data is still available, even though the person may think that it has been removed [11]. The availability of context information, for networks and users, provide new interesting possibilities to spy on people and launch personalised attacks, e.g., in the form of phishing involving social engineering techniques. The availability of accurate personal information can also be used to falsely build a context of trust and then this trust can be abused or various other kinds of identity thefts can be done.

An additional disclosure threat is that when people are no longer so location-dependant in their service usage and use the services casually in public places, it provides more opportunities for simple shoulder surfing and eavesdropping.

Also, using a multitude of social networking services means that the users are at the mercy of the security of these services. Lately there has been news about incidents, where the user database of the services has been acquired through vulnerabilities in their software. Thus, even though the users might have conducted proper password policies, their credentials can still leak out. This is even more disastrous in cases, where people use the same password on multiple sites as often seems to be the case. In a way this is quite understandable, because the burden of remembering numerous passwords is getting higher as people use more and more of these services. In similar sense, the systems offering federated authentication and single sign-on have the risk of cascading. This sets more strict requirements for privilege granularity.

The possibility to use ubiquitous service environments may also mean that in the name of better usability, various places provide external display or input devices for mobile devices, which themselves are limited in this respect. This can pose a threat to the user, if it is not certain under which administration these external devices are. They can be compromised and steal sensitive user information or even execute unintended action on behalf of the user. For instance, there could be a scenario, where one inserts a smart card into a compromised public reader. While the user credentials may stay safe, the card can be made to create signatures on unintended data.

The future concepts also talk much about the flexibility and adaptability of the system. This can, for instance, happen through reconfigurable devices. That, however, can present additional threats to the user as already has been seen with programmable environments in mobile handsets. Even though it can be claimed that the security model of such environment controls tightly the privileges of each component, the user can still be tricked into giving additional rights by promising free SMSs, for instance [12]. Thus, one cannot be certain that the user is always capable of making the right decisions in terms of privilege granting. In fact, allowing the user to make any decisions in the system without knowledge of his mental models for security and privacy is a pitfall. Some vendors are already providing more controlled environments with requirements for vendor signed components, but they tend to result in public outcry for openness.

III. Design process and methodology

So, how does one approach the problem of designing a system that should take into account the user aspects and the aforementioned threats that emerge in the introduction of completely new way of service interaction? ISO 13407 [13](*Human-centered design processes for interactive systems)*, is a widely acknowledged international standard, established in 1999, that provides general guidance for user-centered design (UCD). ISO 13407 focuses on the descriptions of principles and activities to be used in a user-centric design. In the standard there are four particular characteristics that have to be fulfilled in the design activities in order to claim them user-centric. These four characteristics are a) *user involvement*, b) *function*, e.g., carrying out some security related task, c) *iterative manner of design* and d) *multidisciplinarity*. The standard emphasises the role of planning, and one should spend adequate amount of time in planning the study; i.e., identification of users, user demand (for particular task) and task or/and goal setting.

In the first phase context of use has to be found based on collected user, task and environment details, i.e., try to learn to understand the users. Most often in technical oriented studies these are written in a form of narrative scenarios. Naturally, the textual description utilises figures, story-like narration, sketches etc. to support the flow of scenario. The scenario is described from the user point of view and may include different varying constraints and relevant background information. By using such a description it is possible to capture more information about the user's goals and the context the user is operating in. One has to understand that different stakeholders handle the scenarios differently. As an example, an interaction designer looks different aspects while reading the scenario, as he/she looks all the transactions that take place between the human and the computer/device/UI. At the same time he/she follows the description of flow of the activity the user is supposed to be doing. In the second phase all possible requirements are collected, i.e., user, system, organisational, software requirements etc. In the third phase one has to produce appropriate concepts. In the fourth phase the evaluation is carried out to find out if the requirements are met.

Scenario analysis is also a common technique for finding and analysing the security requirements of the system to be designed. Common Criteria (CC) is another alternative for security requirement evaluation, but is has a steeper learning curve [14]. Thus, in a research project the scenario approach is often favoured as it is easier to get involvement from a larger party, even though it is not so readily quantitative. This was also the approached adopted in AN project [15]. Naturally, one also has to have an understanding of the threat model in the envisaged environment. This can further lead to risk management decisions, e.g., a certain risk is deemed so improbable that the mitigation effort to be invested is not seen feasible.

Figure 1 illustrates how security and user aspects can be processed in an iterative manner. The figure is a modified view of the UCD design process. The starting point of the spiral is in the center and curves firstly towards understanding the users. In the first iteration the current knowledge and state-of-the-art understanding are collected so that the awareness of user behaviour, user perception, user motivation, and user attitudes can be obtained. After obtaining that information the conceptualization begins, in which user requirements, software requirements (front-end for the user and back-end for the system) and security requirements are taken into account. In first iteration, low level fidelities of created concepts are available in user studies, which could involve, e.g., simple paper prototypes. These user studies also are used for evaluating whether requirements are met. These first user studies would be followed with some constructive research, e.g., creation of algorithm or mechanism so that a concept can be further prototyped in the second iteration.

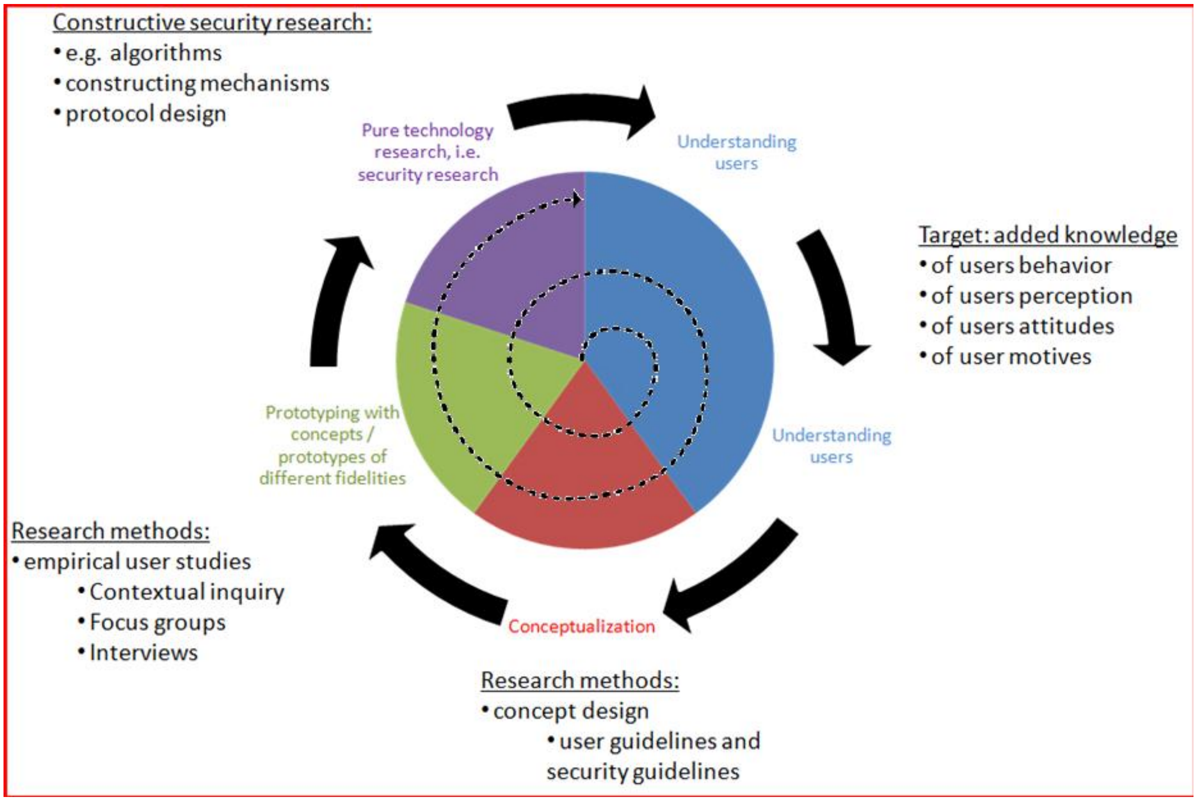


Figure 1. Process for guideline integration

In the second iteration we have a concept design of one or several context(s) of use. We are able to collect user experience within selected contexts by, e.g., observing users. User studies carried out in the first iteration provide us better understanding of user behaviour, user perception, user motivation and user attitudes towards security. In the second round, user studies would have prototypes of higher fidelity in use. Again in constructive security research, a more complex and detailed security-related research is carried out. In theory, the spiral is never-ending, and all the time the accuracy of understanding of users will increase.

The following sections V and VI are linked to the Figure 1. Section V along with sections I and II provide the basis for conceptualization with respect to the security guidelines and security mechanisms. Section VI is focused on user guidelines that are present in conceptualization, user studies and in understanding users.

IV. Aspects of Trust

As ambient computing is about interaction, there generally has to be some sort of trust relationship between the communicating parties. Thus, when doing the initial environment analysis, one needs to understand in what ways trust enters the picture. In technical sense, one can say that trust relationship entails establishing the identity and certain characteristics (such as expected behaviour) of an entity. Here we briefly list different categories of security relations from technical viewpoint. However, one should also note that the trust can be very subjective matter from a user point of view. Therefore, one also needs to analyse the user perception of trust within the evolved landscape.

A. Technical trust

There are several ways of establishing trust relationships, depending on the use scenario and the requirements set by the policies. The simplest case, of course, is to have no protection at all and just blindly trust that nothing goes wrong, i.e., relying on that fame and other external factors, like fear of legal actions, will provide enough protection against misuse. While this is quite common approach nowadays in Internet, this clearly is not recommended in the future setting of potentially unknown operators.

Direct trust on the other hand is based on some common knowledge that has been agreed beforehand. It can be, for instance, a shared secret as is done in the current Subscriber Identity Module (SIM) based solutions. While this can be used to secure mutual connections, it requires some form of pre-configuration. Thus, it is not suitable for most dynamic environments. However, direct trust can be delegated leading to a brokered trust setting, where the trustworthiness of an entity is vouched by an entity one is willing to trust. Through this kind of transitivity the trust relationships can be extended more easily, although in this kind of setting one should talk about liability instead of trust, which often has rather unambiguous meaning. Especially in cases involving compensation and monetary exchange, there is incentive to accept a potential risk only to a certain amount. For instance, a visited operator provides service under the assumption that the home operator accepts liability for the roaming user, hence ensuring the compensation for the visited operator even though the real identity of the roaming user may remain unknown.

The last trust category is based on the opportunistic approach. While it is close to the blind trust case as it takes a leap of faith in the beginning of the communication, it provides an assurance that the party of the initial communication does not change. In other words, you may not know who the communication partner is, but you know that it has remained the same all through the session or that it is the same one with whom you conversed previously. While not suited to every case, it can be a flexible and simple way of providing security in the absence of security infrastructures.

One can also approach trust from the reputation perspective. In other words, the historical behaviour of the entity affects how it is viewed. This can be evaluated with various kinds of mathematical trust metrics. Naturally, it is easy to argue that the past behaviour is not a guarantee of expected behaviour (much like the commonly used phrase in the stock market) and having multiple faked identities allows you increasing your reputation in a certain community (i.e., Sybil kind of attack). Basically, however, the trust based on reputation reduces to the categories discussed above.

B. User perception of trust

The previous technical discussion about trust is somewhat straightforward with quantitative properties suitable for engineers and the like. However, when one talks about trust between persons, there is always a certain amount of uncertainty and it is very subjective experience dependant on the context. Thus, it can be said to be an attitude, based on beliefs and feelings, and implying expectations and dispositions [16]. One can also see it as a process that takes time to develop and shapes the interactions people have [17]. With new things, the reputation and recommendations can form the basis of the initial attitude, but it also depends on the risk-taking attitude of the individual, given the potential benefits. Thus, in the advent of ambient computing environment, user has to trust the system in order to agree to disclose information about themselves, i.e., adjust their privacy settings accordingly. However, the trust evaluation made by a person can be affected and it is not always a rational thing. For instance, the mere look and feel of the system can heavily impact how trustworthy the user sees it [18]. While challenging topic, the design process should also take into account the user perception of trust. Additionally, it is worthwhile to remember that the user can be actively influenced and the user trust abused, e.g., by social engineering means [19].

V. Technical principles and mechanisms

When one starts conceptualizing the initial scenario ideas, one also needs to start considering the security guidelines you wish to follow within the design. This should then lead to some ideas of the actual building blocks used to ensure that the guidelines are followed. Naturally, this entails the actual research to come up with the suitable solutions.

Thus, next we discuss some of the technical aspects of the ambient design in order to ensure that trust relationships can be created. Examples of mechanisms are given, but it is more important to pay attention to the design principles, which should guide the design decisions made early on.

A. Technical design guidelines

In building future secure networks, several general technical design guidelines need to be followed. The list is not focused on any given technology, but rather on the context(s) of the future ambient networks. Many of them (naming, default security, authorisation) already appear in the AN security architecture principles [3]. The designers should keep in mind the classical general principles, as well [20]. For instance, one should honour the defence in depth thinking and not rely solely on one defence mechanism. The list includes:

- Security in design right from the start
- Ease of configuration
- Security by default
- Secure naming
- Privileges and delegation
- Decoupling authentication and authorisation
- Liability brokers

The first and foremost point to consider is the design process itself and how security is brought into it. Quite often security is added as an afterthought and this has a tendency to lead to patched approach, which will cause additional vulnerabilities and degrade usability [21]. Hence, the design process needs cooperation of all the parties right from the start (including both security and usability people). It is also important that they understand each other, i.e., speak the same "language". Otherwise, the parts of the solution might not support each other and instead end up confusing the user even more.

All the more confusing to the user is the complex configuration of security measures [22]. The users have a tendency to think in service centric terms, i.e., they are interested in the added value that the service will bring to them, and not in the details of configuration. For instance, a person might buy a WLAN access point, plug it in, notice it works, and then happily start using it. However, the user easily forgets that there is no security configured as the user would have to get involved with the complexities of the configuration settings. Thus, there is a need for making the configuration as easy as possible in terms the user understands, for instance, by using templates to abstract away the details and mechanisms to support auto-configuration. Currently dominant dynamic address configuration method, Dynamic Host Configuration Protocol (DHCP), is a good example of a mechanism that requires little user involvement. Additional specifications were needed to add security features, but due additional manual configuration requirements these features are hardly ever deployed.

While design effort should go into making the security configuration as easy as possible, it is even more important that there is always some security present. In other words, the design of future ambient networks should follow the security by default principle. It means that there always is some level of security available and it is not something that can be turned off at the time of deployment with an excuse of increasing performance or usability. While this approach does not protect against all the possible threats, it is better to have at least some security than nothing at all. In other words, one should consider opportunistic approach to trust, as discussed earlier.

If security provision is desired between the different communicating parties, naming these parties in a secure fashion would also be desired. This way it is possible to refer to these entities without having to worry about the possibility of spoofing, which can become evident, for instance, in the case of three party protocols. Currently, for example, in a typical cellular setting the user and the home network know each other, but the user has no real knowledge about the identity of the access network. Therefore, it should be required that the protocol design can explicitly identify every party involved in the transaction.

When you are able to name the parties, you can also assign privileges to them. One should not just adopt the approach, where you authenticate an entity and then give implicitly all the possible rights. Instead, one should honour the least privilege principle, which dictates that you only give the rights needed in the current context. This way you minimise the actions that might lead to exploits. In addition, one needs to make sure that authentic data cannot be used in unauthorised context. Privileges also enable one to use delegation mechanisms to outsource the execution of specific tasks to others in order to gain performance benefits. One can, e.g., delegate certain signalling tasks to core network or proxy elements rather than expecting always the end device to do them.

When an entity has a privilege, it is authorised to execute a specific action. However, it is important to remember it should be possible to decouple authorisation from authentication. In other words, it is not always necessary to actually know who the entity using the service is, as long as it has legitimate authorisation for its actions. This helps to alleviate the privacy concerns and the service providers still can be sure that the users are legitimate ones and there is a party, which can be held liable for the actions.

Such liability needs to be established with the help of trusted third parties (TTP). They are needed to broker between the previously unknown parties, because the transactions having real world effects, such as those related to money, need the level of assurance and scalability, which can only be offered by well established institutions that provide financial liability for the interaction. While the old incumbent operators could assume this role, it is also a new business opportunity for the potential new identity providers.

B. Security design building blocks

In building future secure networks, several building blocks need to be implemented to adhere to the above mentioned principles. The list is by no means exhaustive, but rather provides examples of the building blocks suggested for implementing the AN security architecture [3]. They were chosen here for the sake of their fundamental nature as essentials for realising the ambient visions. The list includes:

- Cryptographic identifiers
- Secure network attachment
- Authorisation tokens
- Dynamic roaming agreements
- Non-repudiative service usage

For implementing secure naming one can use cryptographic identifiers. In other words, every entity is assigned an identifier, for which it can provide proof of ownership. That is, it is not probable (in mathematical terms) that anybody else could use the same identifier. Basically, this is a representation of a public key pair. Authentication of the identifiers does not necessarily require existence of any global infrastructure, such as Public Key Infrastructure (PKI), but can take the benefit of local decisions, e.g., key continuity. Thus, there is no need for the user to worry about the complexities involved with PKI [23]. Also, the identifiers can be either short or long lived. When the identifier is only used for a short period of time and it is discarded after use, the privacy of the user can be better preserved. Note that the employment of identifiers on several different levels also demands user centric identity management solutions.

By relying on the "self-certifying" nature of these identifiers, it is possible to provide a default level of security. This relies on the aforementioned concept of opportunistic trust, which is based on the sameness property of the identifiers. In other words, there might not be assurance about the real identity, but the invariability of the identity can be guaranteed. Usually this approach works in environments, where the attacks are more likely to be passive in nature, such as snooping of information. Thus, attacks like man in the middle can still be a concern. However, this allows adhering to the “better than nothing” security principle and one notable example of this is the success of Secure Shell (SSH). Introduction of TTP can be used to further enhance the level of security (see below), if the use case has more stringent requirements.

As the ambient vision states that there will be a multitude of different kind of access networks, there will also be a need for secure way of attaching to them. This can lead to a configuration nightmare. Instead of having many different mechanisms, one should consider providing a common approach, which can be adapted to various interworking layers. This is done with the help of network attachment protocol [24], which in its origin resembles Host Identity Protocol (HIP) [25]. This procedure takes advantage of the well studied security properties of HIP and provides the means for the parties to exchange their identity information and establish keying material, which can be used to secure any subsequent communication as is done, e.g., in the typical use case of HIP (see [26]). Additionally, a conceptual identity layer is created, which can be used for directing traffic between the entities, thus allowing decoupling the locator and identity information for the benefit of better and secure mobility. An important point is also the consideration for Denial of Service (DoS) mitigation through the use of an adaptive puzzle scheme as DoS is currently one of the major threats to the modern data networks. The protocol is run with the help of a four way handshake and it is possible to include additional information into the signalling messages. This could be, for instance, dynamic configuration information to replace DHCP [27]. Subsequently, additional information elements can be exchanged in secure fashion. Thus, by using just basic opportunistic mode the procedure can provide zero-configuration capability.

While the above mentioned procedure can provide the identifiers of the involved parties, it should be further enhanced with the possibility of including authorisation statements, which dictate the rights of the entities and are securely bound to their identifiers. Such statements could be made with the help of X.509 certificates or Security Assertion Markup Language (SAML), but a more flexible (and concise) approach for this environment can be achieved through Simple Public Key Infrastructure (SPKI) certificates [28]. After all, on network level one also needs to consider packet fragmentation issues. The use of such assertions naturally requires that the parties have a common understanding about the trust levels associated with the entities, who have issued the statements. They could be individual delegations or statements issued by the liability brokers. Thus, TTP can, e.g., assign an authorisation to an ephemeral identifier of the user, underlining the fact that the authenticity of the user is not as important as the accompanying token, which ensures the right to perform the action. In other words, there is decoupling between the authorisation and the authentication of the real identity.

It was already mentioned that the operator landscape can change. Hence, it no longer can be expected that the static roaming agreements can cover all the internetworking between the future operator entities as the relationships are more dynamic in nature and perhaps only contain one transaction. This requires measures for establishing dynamic roaming agreements, which also subsequently affect the trust evaluations of the individual subscribers. The operator entities engage in a similar association creation procedure as is done in the network attachment phase. However, this also includes offer and counter-offer steps, which could additionally include an external entity for brokering the agreement or it could be handled by a federation of brokers. The framework for dynamic roaming agreements is depicted in Fig. 2 [29] . In a sense, such setting is currently employed between current incumbent operators, which exchange traffic through closed networks, such as GPRS Roaming Exchange (GRX) networks, although in this setting no direct authorisations for actions are provided by the GRX operators, but instead carrier services with certain security and quality parameters are offered.

While the involved parties, such as operators, can establish agreements concerning their interaction and the actions of their roaming users, there is still need to make sure that the agreements are honoured. Nowadays, in a typical setting the accounting of a visited network is based on the declaration of the visited party. While overly large figures can be spotted, the dynamic environment requires more stringent measures to ensure that the agreed services are received at agreed terms. Thus, there is need for protocols that ensure non-repudiation, so that the user can be sure that he gets the service he is paying for, and the service provider can be sure that it can get the compensation for the provided service. This can be realised with the help of signed hash chains, which can be used as micropayments to represent a piecemeal commitment to the service usage [30]. In other words, if no service is received, no new hash chain values are provided. Similarly, if no hash chain values are received, no service is given. At a later stage the user cannot repudiate the use of the chain values, because they are signed with his identity or that of his operator during the initial service negotiation phase. In practise this requires involvement of TTP, which will ensure the liability of the user, i.e., the service provider knows the brokering party. Thus, the service does not need to learn the “real” identity of the user as long as the presented identity (possibly very short lived one) is asserted by TTP.

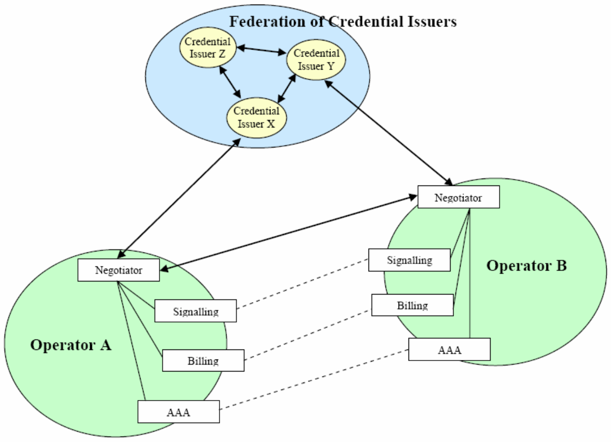


Figure 2. Framework for dynamic roaming agreement [29]

In the Table 1 we have listed some of the presented guidelines and the suggested mechanisms for implementing them. As can be seen the cryptographic identities play an important part in many of them and should be considered to be one of the key building blocks for ensuring the security of the future networks. Naturally, important principles such as security in design right from the start need to be considered more broadly than just in terms of certain mechanisms.

TABLE 1. CORRESPONDENCE OF GUIDELINES AND MECHANISMS

Guideline	Mechanism
Security by default	Secure network attachment
Ease of configuration	
Secure naming	Crypto ids
Privileges and delegation	Crypto ids, authorisation tokens (e.g. SPKI)
Decoupling of authentication and authorisation	
Liability brokers (TTP)	Authorisation tokens, dynamic roaming agreements, non-repudiative service usage

VI. User experience for ambient systems

As indicated above, the design is not just about solving technical obstacles. The future networked landscape will have several emerging trends that will affect how users will interact with the ambient networks. Let us consider the fact that the fundamentals of an ambient network are built on the promises of i) *intelligence* (algorithms, learning capability), ii) *natural interaction* (e.g., multi-modal interfaces) and iii) *ubiquity* (provided by the communication technology). This section focuses on intelligence and natural interaction, which affect the level of obtrusiveness the user can experience.

Riva has introduced several psychological principles for designing ambient spaces [31]. These principles can be applied to any ambient "front-end", i.e., the environment in which the user interacts.

- The environment has to identify what the user is aiming to do. Literally this means that a lot of data has to be collected in order to identify the user objective. If a situated and context-aware profile were available, the environment could respond either proactively or be triggered based on some not (necessarily) known event.
- The environment has to be able to identify the equipments (e.g., mobile phone) the user needs to carry out the objectives. These equipments include both physical and social tools.
- The environment has to be able to understand the current path of user thinking (and future behavioural patterns). This piece of information helps to make decisions, e.g., when a particular task will end. Different sensors will become valuable assistants as information collectors.
- The environment should interrupt the user as little as possible. Most of the actions should be carried out automatically. The intervention should occur only as last resort (i.e., the user has to be helped out). However, the environment should also be transparent to the user. That means that the user is aware of its actions and does not need to "worry" whether things have been appropriately done.
- The environment should be able to utilise situated contextual benefits and restrictions of it in a transparent manner.
- The environment should also support social behaviour of the user; identifying the roles and social networks in a manner supporting normal activity of a given user.

Even though the environment mostly carries out the tasks based on, e.g., situated and context-aware profile of the user without explicit orders from the user, the user sometimes has to interact with the environment explicitly. The key elements here are natural interaction and multimodality. These multimodal interaction models include things like

- Speech recognition and spoken interaction (or sounds/voices in general)
- Physical interaction (e.g., touch-based)
- Adaptive graphical interfaces (e.g., appropriate for public spaces)
- Gesture and gaze interaction
- Haptical interaction
- Space and virtual reality -oriented interaction.

In a sense ambient systems are challenging for user-centric design as you cannot summon experiences from other researchers. Thus, for pioneers it might be partially guesswork. If the technology goes into background and you have to rely its black-box way of operation in which you have to start trusting to the system so that it operates as it was designed and as you were told. Thus, the environmental characteristics in the context of use for understanding the users are different.

A. Introducing user experience

According to [32], experience can be put into four realms based on the level of participation of a user and his/her connection to experience itself. The first realm is pure entertainment, in which users are passive viewers (e.g., opera). In the second realm, the user is actively absorbing information from the environment (e.g., classroom with active learning settings). In the third realm, experience is summoned in immersed manner (e.g., flight simulator). In the fourth realm, the immersion is obtained in a passive environment (e.g., going to a medieval castle). The user perception in all these realms is different; thus user experience should surpass the expectations. As new technology is often viewed sceptically, surpassing the expectations should not be a big hurdle.

There is no single definition of user experience (UX). COST 294 action (MAUSE, towards to Maturation of information technology Usability Evaluation) tries to build a holistic view on the UX. In their deliverables ([33],[34]), user experience is viewed from many angles; for their purposes user experience terminology is put into statements that deal with fundamental assumptions underlying UX (principles), positioning of UX relative to other domains (policy), and action plans for improving the design and evaluation of UX (plans) [33]. By their terminology, e.g. trust is seen as one attribute of structuring user experience. The structuring of UX itself is part of the principles. As said, UX is a broadly defined term, including attainment of behavioural goals, satisfaction of non-instrumental (or hedonic) needs, and acquisition of positive feeling and well-being. Neither a universal definition of UX nor a cohesive theory of experience yet exists how to practically design for and evaluate UX [34].

In [34], UX is differentiated from usability because i) UX aims to follow holistic approach, ii) UX is subjective and iii) UX aims towards positive experiences.

i. Usability strongly focuses on tasks and user accomplishment. UX holistic approach aims for a balance between pragmatic aspects (i.e., usability) and other non-task related aspects (hedonic), such as beauty or self-expression.

ii. As conceptual origins of usability are in cognitive psychology, work psychology and human factors, usability is more of an objective expert-oriented approach. In contrast, UX is subjective and is not based on task success or results of usability studies. UX is explicitly interested in how users experience and judge, e.g., technology products they use. Thus, the perception of a user plays a much bigger role.

Usability focuses on negative aspects of the studies; most often problems, errors etc. are investigated through usability studies. However, UX tries to build positive outcomes of the use or possession of technology, e.g., positive emotions such as joy, pride, and excitement.

User experience is mostly collected by observing / interacting with people. This can be done in laboratory settings which might distort the results, as some people might not behave naturally when observed. Observation can also be done in field, e.g., travelling in buses for few weeks and just observing how users use their mobile phones. One can get the general trend and maybe the frequency of using mobile device, but not necessarily the details. Of course, in a research setting, a researcher has to use multiple research methods, both qualitative (e.g. interviews) and quantitative (e.g. surveys).

B. *Designing good user experience*

Jameson has emphasised the following goals for enabling enhanced user experience especially in the context of user adaptive systems [35]. As such they act as guidelines and design restrictions in ambient intelligence environments. These elements include

- Predictability
- Visibility
- Manageability
- Non-disturbance
- Privacy and feeling of being secure
- Depth and severity of the experience

Predictability and visibility relate to the working of a system according to the user expectations. Thus, if the observed behaviour is in contradiction to what the user expected, the user is bound to get confused. In the case of a mental model of security this can be quite dangerous, because the user may end up compromising his security without really realising it.

Manageability or controllability refers to the amount of control the user has over the system. The system could, for instance, ask the user to decide whether to accept certain connection attempts. However, this can be a challenging topic when weighted against the unobtrusiveness.

The system should not bother the user unnecessarily. Otherwise the user may find the system obtrusive and burdensome to use. It can also overload the user with too much information, thus the user no longer evaluates information carefully. The discussion in the next subsection about SSL with too many warning messages is a good example of this.

User adaptivity generally requires storing information about the user and his actions. Some of this could be even considered to be very sensitive information. Thus, the potential information disclosure to unauthorised parties can have severe consequences. Some users might even get a "big brother is watching" feeling and turn off any functionality that otherwise might enhance their user experience.

Breadth of experience can be seen as a challenge of filtering too much information and hence limiting the user decisions. In other words, the user "learns" less, when all the decisions are made for him by the system.

This last (depth of experience) is important in order to get main stream experience correctly. However, it is difficult to get those extreme set of experiences of first-timers and those who like to do it "my way". It might be reasonable to downplay the benefits of technology, so that it actually surpasses the user expectations, and thus user perception is positive, which provides better overall for, e.g., a task that is not previously seen important or cumbersome to carry out.

C. *User experience in security*

As the discussion above has indicated, the security should be built-in, not an add-on feature. Security as a theme focuses on the risks and uncertainty. These are extremely difficult concepts for the people to evaluate, argues West in [36]. Furthermore, he argues that it is more important to understand the basic principles of human behaviour (as also the previous section indicates). He also lists a comprehensive list (see Table 2) of predictable and exploitable characteristics of our decision-making.

TABLE 2. USER CHARACTERISTICS IN SECURITY THEME [36]

Characteristic	Comment and effect
Users do not think that they are at risk	The users most often think that they are better than others, and thus either do not use security features or proceed with more risky behaviour.
Users aren't stupid, they are unmotivated	Human beings (as a species) tend to favour quick decisions based on learned rules and heuristics. Security can be seen as overly exhaustive action.
Safety is an abstract concept	The less concrete the threat is, the less willingness there is to carry out security instructions.
Feedback and learning from security-related decisions	Behaviour is shaped by positive or negative reinforcements. In security domain, most often the reinforcements are negative.
Evaluating the security vs. cost trade-off	Gains are often abstract and the negative consequences stochastic, the cost is real and immediate.
Making trade-offs between risk, losses, and gains	If security gains are intangible, with well-known costs, and while negative consequences involve probabilities, it is possible to try to make security more "profitable" for the user.
Users are more likely to gamble for a loss than accept a guaranteed loss	People react differently on whether they think they are gaining or losing something (in concrete value).
Security is a secondary task	People tend to focus on the immediate task. As such, security decisions need to be carried out most often in the middle of some other (more relevant) task.

Losses perceived disproportionately to gains	People do not perceive gains and losses equally. So the user has to perceive gain visibly better than a loss.

West also lists several approaches that could help the security designer to improve human compliance (to security) and decision making. These approaches include

- Rewarding pro-security behaviour (e.g., immediate feedback given to the user)
- Improving the user awareness of risk
- Catching security policy violators (non-repudiation / deterrence)
- Reducing the cost (for the user) of implementing the security (e.g., sufficient always-on security by default)

With respect to the interaction with the user it is important to also consider the amount of information provided to the user, i.e., how obtrusive the systems can be. If the user is overloaded with information it might lead to cases, where the user no longer evaluates the information but just concentrates on absorbing or merely ignoring. Nowadays, this is quite evident with the use of SSL warning messages: users simply click ok, because they have seen similar windows so many times or actually do not even have any idea what the warning means. Similar things can be faced if poorly functioning heuristic systems are used to evaluate potential threats to the user and too many false positives "condition" the user to ignore the warning messages, just like crying "wolf" too many times [37].

In Table 3 we have summarised the relations between the presented security and user guidelines in order to show that even though the concepts can be claimed to be residing on different levels, correct security design decisions taken already at lower levels can benefit the user experience and increase the overall effective security. We have further developed a hypothetical example scenario to illustrate the applicability of security and user guidelines presented in this paper. The scenario is, as scenarios are, narrative and focuses on user experience. Beneath it the technology research has to be read partially between the lines and as such leaves lot of room for different implementation options for the actual developer. Similar scenarios are easy to create in a hypothetical manner, but for real-world case one needs to empower and engage real users to do security related tasks in order to get relevant and accurate information from the users.

<u>Hypothetical scenario and example case:</u>

It is August 25th, year 2012,, and a time for the annual company party at the AmbVision Ltd headquarters. Matt Ellis, one of company's security staff, who was given the task of organizing the event for this year proudly waits for employees to arrive. In his left hand he has a company wrist clock awarded for dedicated work for the company. The wrist clock also has a security functionality and capability to communicate with the company's information system in a secure manner. It also contains wireless tag reader so that visitor tags could be read while they arrive to the AmbVision lobby and at the same time for a security check. It is still an hour and half for the company CEO speech and the official kick-off for the party, and the employers start to arrive. Some of the employees have dedicated tasks to carry out in order to make the party successful. Their tasks will be transferred to their wrist computers at the security check point.

Maria Smith is a new employee and for her this will be the first company party event. She has been working for two and half months and is really waiting for this party. However, today his boyfriend Frank Sonay came to a surprise visit and wants to come with her. After all, she has attended his parties, too. She knows that the security procedures are strict but she borrows the wrist computer of a fellow employee who happens to be in a hospital due to a traffic accident. Maria is able to delegate the watch to the identity of her boyfriend, but only with a limited profile with no access to the services of AmbVision. Frank acknowledges the watch by tapping it with his own company issued phone, which ensures the pairing of identities.

Over 100 employees have already arrived and Matt feels that nothing can go wrong today. He has made all the necessary security checks and even stricter security policies for communication. He puts the security lens on top of his eyeglasses and views the security logs, i.e., hardware reports, communication logs, network traffic graphs, and user profile data. So far no major deviations and everything is under control. Maria and Frank arrive at the security check point. Their wrist computers are scanned and the system informs the guard that the current user identity associated with the watch cannot be identified as AmbVision employee nor does it have the correct authorisation.. According to the policy, the guard is supposed to send a dedicated message to the information system which will control the further activities. Incidentally, Frank happens to be using his own employee identity and his company is also doing mutual projects with AmbVision. This same information is relayed (with information exchange to registry of Frank's company, which tells who this unidentified person is) to Matt who feels the wrist computer to tremble and sees the message and appropriate information. Matt browses the event data file and changes Maria's task in her computer so that her job is to clean the mess in a meeting room in 2nd floor in the corner of the building, far away from the CEO speech room. Maria's wrist computer begins to tremble and she reads the message and acknowledges that the task could be done faster if two persons would do the cleaning and thus Frank comes with her. They arrive at the room and Matt is already "cleaning" the room with two security staff members dressed as employees. Maria and Frank arrive and see that the mess is really big as five persons are needed to clean it. Matt asks Maria, why she has come with the boyfriend to a party. Maria starts to explain and gives her sincere apology. Matt tells her that everything is fine, he just has to change Frank's wrist computer for such ones that are meant for visitors and welcomes Frank to the party. However, Maria will lose five company points on her security portfolio.

TABLE 3. CORRESPONDENCE OF SECURITY GUIDELINES AGAINST USABILITY

Security guideline	User guideline	Usability goal	Rationale	Scenario example implications
Security by default	Reduce cost of implementing security	Unobtrusiveness Predictability	No extra mental burden is put to the user as an expected default level of security is always present.	Matt is the person in charge in selecting appropriate security policy for the event. The system has pre-set policies (so that additional policies do not need new implementation) and it is enough for Matt to select and thus also see what that chosen policy actually means on the individual, group, etc. level. in relation to the standard policy. The communication between the watches and the company systems is protected by default without the user having to configure anything. Naturally, the administrative systems are aware of the legitimate end-devices.
Ease of configuration	Reduce cost of implementing security Improving user awareness	Unobtrusiveness Visibility Controllability	User is not needlessly interrupted with secondary tasks, but still has a sense of being in control for added security.	Users do not have to go through complex configuration procedures, e.g., tapping devices together might be sufficient procedure for acknowledgement ..
Secure naming	Improving user awareness Catching policy violators	Privacy Controllability Visibility	Assurance about the communicating parties and invariability of them either with short or long term identities.	The watches carry the identity of a watch and that of its current wearer. The systems they interact with can be identified as legitimate ones. Security logs can be later on audited and one can also see who has accessed the logs. The trustworthiness of the system can be measured so that users understands/sees how it safeguards, e.g., their privacy. Also, Frank was able to control, which identity he wanted to use with the watch..
Decoupling authentication and authorisation	Improving user awareness	Privacy Visibility	Only authorisation is explicitly linked to the execution of the defined actions.	While the company watch might provide authorisation to access the event, the wearer identity does not hold such assertion, which normally could be assigned to companions, as Matt later does. The user mental model is directed toward the action, which requires authorisation instead of a person (like someone appearing with a trusted person). This may also ease the job of log administration as data protection laws may have more restrictions on the handling of data containing personal, i.e., identity information.
Privileges & delegation	Reward pro-security behaviour	Controllability	Efficient execution of tasks and assigning privileges as needed for controlling the disclosure of information with timely feedback.	The policy of the watch allows delegation of it to other people, but with limited rights. Users are always also told about the decisions and why these decisions have been made. Users are aware of control as e.g. security checkpoint in the scenario implicates. Also the security guard did not interfere for stopping the visitor as the system provided enough information for evaluation the case so that more appropriate solution could be carried out. It might have been the case that another visitor could be handled differently
Liability brokers	Catch policy violators	Unobtrusiveness Privacy	Outsourcing the trust evaluation and reliance on external mechanisms (such as litigation)	Frank is considered semitrusted as the identity system of the partner company can vouch for his identity without Frank having to actively do anything. The security logs can catch/find policy violators or possibly organisations that are liable for arranging the violating privileges.

D. Towards trusted user experience

As we are heading towards future ambient networked systems, the user should not need to ask how, why, what, when and where. However, user demand and requirements vary highly depending on the context and situation. The technology might not be mature enough yet, as fulfilling user demand and user requirements in different situated contexts faces an increasing level of uncertainty. The user demand is quite often described as a higher-level demand that can be constructed in a given situated context from the identifiable attributes. The user requirements, on the other hand, are often seen as a critical issue of using technology (e.g., so is the case in requirement engineering). Furthermore, as ambient systems naturally operate in the background, trust will become major issue in accepting new systems and environments into use.

Hoffman has created a trust model with related metrics for distributed information systems [38]. Trust model has generic model parameters and subcomponents such as security, usability, privacy, reliability and availability, audit and verification mechanisms, and user expectation. In creating user experience, usability subcomponent has as general model parameters perception issues, motor accessibility, and interaction design issues. User expectation subcomponent has product reputation, prior user experience, knowledge of technology, and use of trusted agents.

The perception issues can also be directly linked to security characteristics, e.g., perception of controllability and observability. Motor accessibility is a personal feature and thus the interaction design issues should deal with the special target groups. Product reputation can be a powerful tool as the user can feel more trustworthy towards known brands. Prior user experience could become the major element in trust provision. In general, most of users do not want to learn new things, especially if they sound too complex or look cumbersome to manage. Thus, the technology should have high enough accurate cognition of the experience and capability of the user. Observability is also a direct perception issue, and the user should have that particular capability. The interaction should provide the perception of controllability, feeling secured in private and trusted manner [39].

It is also worthwhile to note that the attitudes of the people towards the technology and its acceptance changes over time. As noted in [40], even privacy disruptive technologies such as camera phones can become socially acceptable in a relatively short timeframe. Thus, user suspicions towards the technology have faded. It is more a question of how technology is used, i.e., in an appropriate way, and whether the users are aware of the existence of such technology. As stated in [40], the designers should try to predict and influence these adoption patterns.

Trust categories, i.e., technical and human trust, introduced in section III can be seen analogous to usability and user experience (see Table 4). Technical trust definitions, attributes of trust (e.g., level of trust, origin of trust) and carrying out trust related functionalities are very much similar to usability as both have an objective and fact-based (measurable) approach. Thus, linking them together in a conceptual level is very straightforward. On the contrary, as user experience and user perception of trust are both subjective, linking them is not as straightforward. Such kind of trust cannot be modeled based on technical modeling fundamentals such as system architectures, software architectures or messages and interfaces between different nodes or/and components. However, we have tried to identify security related user aspects brought forward in this article that are important for building holistic user experience.

TABLE 4. MAPPING OF USABILITY AND USER EXPERIENCE TO TRUST

Attribute	Characteristics	Mapping to trust
Usability	*Pragmatist view:* Usability is likely the most important user requirement as it has a heavy impact on the acceptance of technology. ISO 9126 metrics can be mirrored through user experience lens. The methods are carried out in objective manner to address the required tasks and accomplishments of the user, e.g., task of changing password based on metrics such as task success.	Technical trust • Direct trust: Common shared secret, i.e., preconfiguration • Opportunistic trust: The sameness property, e.g., key continuity • Blind trust: Metrics of uncertainty
User experience	*Holistic view:* Integration of task related issues and non-task related issues such as challenges (e.g., in a fashion of games) and stimulation in order to give more joy and excitement of performing “mandatory” security functionalities, i.e., user overall experience is taken into account. The final set of functions should be based on subjective design, implementation and evaluation.	User perception of trust • Being in control: User is empowered to manage and audit the decisions taken by the system • Feeling safe: Physical security • Privacy: How and what information is disclosed • Reputation: Expectations and brand trust • Level of comfort: User is not cognitively overloaded • Assurance: System functions as expected

VII. DISCUSSION

As we have shown, users are facing risks and uncertainties in the evolved networked service landscape due to the user mindset, information leakage, and shortcomings of the platforms. Users are not generally interested in technical details such as configuring security; they only want to get their own tasks done. This can become evident in a case, where the user has the option of choosing either secure or insecure service and for some reason the secure version does not work. Thus, if DoS is launched against the secured service, the user is tempted to use the available insecure version instead [41]. Nevertheless, the end users are increasingly facing the fact that they are expected to become their own systems' administrators, at least within their home networks. The security systems provided by user’s work organisation do not cover leisure time and private mobile devices.

Thus, we underline the importance of keeping the security in the design process right from the start. So, the user can always enjoy default security without having to concern him or her with configuration issues as it is evident that users prefer unobtrusive systems, which do not require them to understand the mental models behind the security mechanisms. The importance of such proactive design choices is also underlined in [42], which proposes research priorities for future mobile telecommunications.

Additionally, secure identification (be it short or long term) along with proper privileges need to be applied to control the information disclosure. Also, many other mechanisms can be based on the existence of secure naming as a building block and proper identity management can be used to alleviate the previously mentioned shortcomings of purely password based systems. It should be noted, however, that the user mindset is a challenging topic for solely technically oriented design, thus, the lessons learnt from the user experience design can pave the way for a more holistic approach.

As known already for decades there has been confrontation between security and usability. Many data security techniques originate from military world, where those who need to use a system, are educated to use it and the rest are kept in dark. In the modern world we need to recognize both the heterogeneity of networks and the heterogeneity of users. Trying to add usability on top of an already designed and implemented service or a product can lead to serious problems. Another fact is also that security mechanisms are designed, implemented, applied, and breached by people. Thus, the user-centered design is essential for all security related systems. It has been argued that hackers pay more attention to the human link in the security chain than security designers do [43].

Designing secure architectures that should both be visible to users and hide security implementation, e.g., protocols used, is challenging. Reducing the user's burden of complex configurations is possible, but it requires rethinking of design methods and phases. Usability studies reveal critical errors and give feedback for iteration. Although the single product development of networking devices has strived for both a satisfying user experience and security, as in [44], generally the architecture design takes purely a technical approach and lacks the support for usability aspects.

Considering the tradeoffs between invisible and transparent security is unavoidable procedure when designing secure systems, but letting the user decide about the critical security features is simply bad design. There are numerous examples of situations where the problems of complex networking security have been shifted to user interface level. Many applications even offer users possibilities to bypass security elements. Relying solely on user's skills to make decisions or education as a solution to security problems is doomed to fail. Gutmann [37] has pointed out the need of considering theoretical vs. effective security: if security measures are misused, turned off, or bypassed, the system offers very little effective security. Thus, models with "always on" security should be applied, e.g., with technologies presented earlier. Also, as mentioned previously, predictability is an important property in user experience; therefore consistent solutions are needed, such as those providing secure attachment procedures across different networks.

There have been success stories of designing usable security; instead of forcing the user through 38 steps of WLAN configuration with decisions and actions, by designing a user interface there are only 4 steps to go [45]. Innovative design solutions and disruptive thinking, which take a holistic approach to the whole problem rather than concentrating on one specific problem field at a time, will be needed. Similar holistic approach can also be used when applying cross-layer thinking to reduce the performance effect of multiple overlapping security mechanisms on several protocol layers [46].

Although global PKI is still considered too complex and out of reach for typical end users, work for thinking locally has resulted in usable and secure wireless network [23]. The use of cryptographic identifiers on local scale can further benefit such systems. However, the design decisions do not have to be anything huge and unprecedented. They include small steps keeping the user in mind and also testing early prototypes. Writing lists of anti-requirements (things that your design should not allow the user to do) and simple "default-action"-tests given in [37] reveal the security level of the system.

Changes are required also within usability testing itself: e.g., better use of data logs of the systems, reformulating activities that we are observing in the field studies, and reconsidering the methods and topics of the interviews [47]. Designing tests for security systems that also take into account the usability and user experience factors differs from designing ordinary customer products or services. Thus, the development teams of security systems or architectures should always include also persons with expertise in usability and understanding of user experience.

VIII. Conclusion

In this article we have presented and elaborated some of the results found within the Ambient Networks project and related work. While they cannot be said to be conclusive, they still provide guidelines and solution concepts, such as secure naming, which can be used to raise the security of the future ubiquitous systems to a level, where there is always a baseline of security present.

Even though networks can be seen to be technical concepts, the holistic design processes have to also remember the existence of the user. The user experience factors on the chosen solutions can dictate whether the system will ever be deployed or used. Security and usability have to go hand in hand and be in the design process right from the start in order to ensure secure user experience. It is not enough that the designer thinks that the user is safe; the user also has to have the feeling of being secure. If the user finds the system obtrusive or too complex to understand, it is likely that there is little trust towards the system, hence hindering the adoption of the system.

Integration of user experience into security design is on its early stages and is not very well studied so far, even though usability and security have been the subject of many studies. In this article we have taken initial steps to introduce a more holistic view towards designing a trusted user experience, so that one can take into account the behaviour of the user and how the user perceives trust in an ambient environment.

Thus, we need to learn more about the users and how they process security related issues. The guidelines presented in this paper provide a feasible plan going forward but the real measure can only be taken when we can proudly say that we are able to provide a secure user experience and the user can agree to that.

[1] S. Heikkinen, K. Heikkinen, S.Kinnari, "Security and User Guidelines for the Design of the Future Networked Systems". Proceedings of the *Third International Conference on Digital Society (ICDS 2009)*, Feb 2009.

[2] M. Johnsson (Ed.), "AN System Description", Ambient Networks project deliverable D18-A.4, Feb 2008.

[3] F. Kohlmayer (Ed.), "Ambient Networks Security Architecture", *Ambient Networks project deliverable D7-2*, Dec 2005.

[4] K. Ducatel (Ed.), "Scenarios for ambient intelligence in 2010", European Commission IST Advisory Group report, Feb 2001.

[5] G. Bell, P. Dourish, "Yesterday's tomorrows: notes on ubiquitous computing's dominant vision", *Personal and Ubiquitous Computing*, Vol 11, Issue 2, Jan 2007.

[6] N. Niebert et al, "Ambient Networks: An Architecture for Communications Networks Beyond 3G", *IEEE Wireless Communications*, Vol .11, No. 2, Apr 2004.

[7] 3GPP. "Network Composition Feasibility Study", 3rd Generation Partnership Project Technical Report, TR22.980 V8.1.0, June 2007.

[8] Facebook. "Terms of Use", November 15, 2007. Available http://www.facebook.com/terms.php (accessed 01/2008)

[9] Y. Chen, J. Hwang, R. Song, G. Yee, L. Korba, "Online Gaming Cheating and Security Issue", Proceedings of *International Conference on Information Technology: Coding and Computing*, Apr 2005.

[10] M. Kampmann et. al., "Dynamic Adaptable Overlay Networks for Personalised Service Delivery", Proceedings of *The First Ambient Networks Workshop on Mobility, Multiaccess, and Network Management*. Oct 2007.

[11] J. Nolan, M. Levesque, "Hacking Human: Data-Archaeology and Surveillance in Social Networks", *ACM SIGGROUP Bulletin*, Vol. 25, Issue 2, Feb 2005.

[12] J. Niemelä, K. Tocheva, M. Tolvanen, "F-Secure Trojan Information Pages: Redbrowser.A", http://www.f-secure.com/v-descs/redbrowser_a.shtml, (online article, accessed 10/2007), Mar 2006.

[13] ISO/IEC, "ISO 13407:1999 Human-Centred Design Processes for Interactive Systems", International Organization for Standardization Standard, 1999.

[14] M.H.Diallo, J. Romero-Mariona, S. E. Sim, D.J. Richardon, " A Comparative Evaluation of Three Approaches to Specifying Security Requirements", Proceedings of 12th *Working Conference on Requirements Engineering*, Jun 2006.

[15] B. Busropan (Ed.), "Ambient Network Scenarios, Requirements and Draft Concepts", Ambient Networks project deliverable D1.2, Oct 2004.

[16] T. Govier, "Social trust and human communities", McGill-Queen's University Press, 1997.

[17] L. Perusco, K. Michael, "Control, trust, privacy and security: evaluating location based services", *IEEE Technology and Society Magazine*, Vol 26, Issue 1, 2007.

[18] F. N. Egger, "Trust me, I'm an online vendor: towards a model of trust for e-commerce system design", Proceedings of *Conference on Human Factors in Computing Systems*, Apr 2000.

[19] S. Heikkinen, "Social engineering in the world of emerging communication technologies", Proceedings of *Wireless World Research Forum Meeting #17*, Nov 2006.

[20] J.H. Saltzer, M.D. Schroeder, "The protection of information in computer systems", Proceedings of the *IEEE*, Vol. 63, Issue 9, Sep 1975.

[21] K. Yee, "Aligning Security and Usability", *IEEE Security & Privacy Magazine*, Vol. 2, Issue 5, Sep 2004.

[22] A. Whitten, J.D. Tygar, "Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0", Proceedings of the *8th USENIX Symposium*, Aug 1999.

[23] D. Balfanz, "In search of usable security: five lessons from the field", *Security & Privacy Magazine, IEEE*, vol. 2, 2004.

[24] T. Rinta-aho et al, "Ambient Networks Attachment", *16th IST Mobile and Wireless Communications Summit*, Jul 2007.

[25] P. Jokela (Ed.), "Host Identity Protocol", *IETF RFC 5201*, Apr 2008.

[26] S. Heikkinen, M. Priestley, J. Arkko, P. Eronen, H. Tschofenig, "Securing Network Attachment and Compensation", Proceedings of *Wireless World Research Forum Meeting #15*, Nov 2005.

[27] S. Heikkinen, H. Tschofenig, "HIP Based Approach for Configuration Provisioning", Proceedings of the *17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Sep 2006.

[28] S. Heikkinen, "Authorising HIP enabled communication", Proceedings of the *10th International Symposium on Performance Evaluation of Computer and Telecommunication Systems*, Jul 2007.

[29] M. Georgiades (Ed.), "Security Requirements, Concepts and Solutions for Secure Access and Mobility Procedures", Annex 2 of Ambient Networks project deliverable D7-2, Dec 2005.

[30] S. Heikkinen, "Non-repudiable service usage with host identities", Proceedings of the *Second International Conference on Internet Monitoring and Protection*, Jul 2007.

[31] G. Riva, "The Psychology of Ambient Intelligence: Activity, situation and presence", IOS Press, 2005.

[32] B. J. Pine, J. H. Gilmore, "The Experience Economy: Work is theater & every business a stage", HBS Press, 1999.

[33] E. Law, A. Vermeeren, M. Hassenzahl, M. Blythe (Eds.), "Towards a UX Manifesto", COST294-MAUSE affiliated workshop. Sep 2007.

[34] E. Law, E. T. Hvannberg, M. Hassenzahl (Eds.) "User experience: Towards a unified view", Proceedings of the *2nd International Workshop on User eXperience*. Oct 2006

[35] A. Jameson, "Adaptive Interfaces and Agents", Human-Computer Interaction Handbook, Erlbaum, 2003.

[36] J. West, "The Psychology of Security", *Communication of the ACM*, Vol 51, No. 4, Apr 2008.

[37] P. Gutmann, “Security Usability Fundamentals,” Available http://www.cs.auckland.ac.nz/~pgut001/pubs/usability.pdf (online article, accessed 01/2008).

[38] L.J. Hoffmann, K. Lawson-Jenkins, J. Blum, "Trust Beyond Security: An Expanded Trust Model", *IEEE Communications of the ACM* Vol 49, No. 7, July 2006.

[39] K. Heikkinen,N. Prasad, "Empowerment: Enabler for Personalized Security and Privacy",Proceedings of *IEEE Globecom Workshops* Nov 2007.

[40] G. Iachello, K. N. Truong, G. D. Abowd, G. R. Hayes, M. Stevens, "Prototyping and Sampling Experience to Evaluate Ubiquitous Computing Privacy in the Real World", Proceedings of *CHI 2006*, Apr 2006.

[41] B. Schneier, "Secret and Lies", Wiley Computer Publishing, 2000.

[42] R. Savola (Ed.). "Security, Trust, Dependability and Privacy in Wireless and Mobile Telecommunications", White paper appearing in eMobility deliverable D2.1, Nov 2008.

[43] A. Adams, M. A. Sasse, "Users are not the enemy," *Communication of ACM,* Vol. 42, Issue 12, Dec 1999.

[44] S. Elmore, S. Hamilton, S. Ivaturi, "Designing software for consumers to easily set up a secure home network,", Proceedings of *25th SIGCHI Conference on Human Factors in Computing Systems 2007,*, 2007.

[45] G. Balfanz, R. E. Durfee, D. Grinter, P. Smetters, P. Stewart, "Network-in-a-box: How to set up a secure wireless network in under a minute," Proceedings of the 13th *USENIX Security Symposium,* 2004.

[46] J. Arkko, P. Eronen, H. Tschofenig, S. Heikkinen, A. Prasad, "Quick NAP - Secure and Efficient Network Access Protocol", Proceedings of the *6th International Workshop on Applications and Services in Wireless Networks*, May 2006.

[47] D. K. Smetters, R. E. Grinter, "Moving from the design of usable security technologies to the design of useful secure applications," Proceedings of *New Security Paradigms Workshop 2002,* 2002.

Networking and Security Issues for Remote Gaming: The Approach of G@L

Christos Bouras, Vassilis Poulopoulos and Vassilis Tsogkas
Research Academic Computer Technology Institute,
N. Kazantzaki, Panepistimioupoli Patras, Greece
bouras@cti.gr, poulop@cti.gr, tsogkas@cti.gr
tel. +302610996951
fax. +302610996358

Abstract

As the evolution of computer technology introduces new advances in networks among others, online gaming becomes a new trend. Following the trends of our era, The Games At Large IST Project introduces an innovative platform for running interactive, rich content multimedia applications over a Wireless Local Area Network. The Games at Large project's vision is to provide a new system architecture for Interactive Multimedia that will enhance existing CE devices such as, Set Top Boxes (STB), Small Screen and other devices, which are lacking both the CPU power and the graphical performance to provide a rich user experience. In this study we present the controllers' sub-system of the innovative mechanism that is implemented within the context of the Games at Large project. We furthermore provide information on the encryption and security of the aforementioned communication channel.

Index Terms — *remote control channel, online gaming, remote command execution, input device capturing, asymmetric encryption, reverse channel*

1. Introduction

The future home is an always-on connected digital home. By the year 2010, there will be more than 420 million broadband households worldwide [8][16]. With the standard set for super-high speed, always-on connection, the way people view entertainment has fundamentally changed and new standards for consumption were created. Consumers no longer expect their Internet access to be only from a desktop PC - now they want it through the TV in their living room or in the palm of their hand, inside the house and on the go. The presented scenario [5] bundles video gaming capabilities into consumer electronics devices, such as Set-Top Boxes (STBs), Digital Video Recorders (DVRs), home entertainment systems, TVs, handhelds and other devices that are not considered, today, as real gaming devices since they lack the necessary CPU and GPU power. In this study we present a new system for pervasive gaming and multimedia, which is being developed under the EU FP6 project, Games At Large (G@L). This study is dedicated to the design testing concept elaboration, in order to base the approach for the development of evaluation and testing methodologies. The testing and verification process is part of the iterative, spiral-life workflow model (user-centered design and incremental improvement based on feedback from user and expert evaluation of prototypes).

The main idea of the project is that one or more powerful servers will actually execute the game on behalf of the client, which will be presented only with the screenshots of the game and not the game loader or the execution of complex graphics. On the other hand, the basic aspect of a game is the interaction with the end user (gamer). This means that apart from only presenting the game frames to the user (through a client – server architecture) the system must be able to capture any input from the input devices of the end user and transfer them to the server in order to emulate the interaction that is done on a physical level when playing a game. An important aspect of the aforementioned procedure is security when transferring the input commands from the client to the server. In particular, keyboard input, which in most cases depicts user sensitive data such as passwords or credit card numbers, must be foolproof.

In this study we present a mechanism for transferring input commands from any device, acting as the client, to execution commands at the corresponding program - game of the server. The purpose of this mechanism is to be able to control a program that runs on the centralized server from a remote operating system. This mechanism is created within the scope of the Games at Large project. Meeting the demand of highly interactive multimedia systems with low cost end devices (CE), requires a radical change in the system's architecture. The Games At Large project intends to design a platform for running interactive rich content multimedia applications. Games At Large vision is to provide a novel system architecture for Interactive Multimedia which will enhance existing CE devices such as, Set Top Boxes (STB) and other devices which are lacking both the CPU power and the graphical performance, to provide a rich user gaming experience. We thus present the general architecture of the sub-system that controls the input of the client devices and their server side execution. More specifically, we examine how, input is able to be captured by any input device on the different end devices and on different operating systems, how commands are sent over the network and finally, how commands are executed at the target software of the server. Moreover, we present the general architecture of the encryption subsystem which ensures that the input from any keyboard devices connected to the client side is encrypted before being transmitted to the server side for execution. The purpose of this mechanism is to expand the capabilities of the command transferring channel. More specifically, we examine how capturing from any input device on different end devices and on different operating systems is done, how public key encryption is applied and how commands are decrypted and executed at the target software of the server. In our work, we are considering only the confidentiality issues of the cryptographic module assuming that authenticity should be provided by the general architecture of the system or by a different module.

The rest of the manuscript is structured as follows: the next section provides information about related work. Section 3 describes the vision and goal of the Games at Large project. Section 4 describes the general architecture of the system and the architecture on each device (the end device and the server). Section 5 describes the encrypted command channel infrastructure, while section 6 presents the encryption subsystem. In section 7 we present the general client-server infrastructure. The paper concludes with general remarks and future work that will be done within the scope of the project.

2. Related Work

Computer games constitute nowadays one of the most dynamic and fastest changing technological area, both in terms of market evolution and technology development. In this area, as the computer games are evolving and online activities and gaming become parts of our lives, the need for interaction within a client – server architecture becomes very intense. The successful paradigms of online gaming such as WoW [15], Half Life [7] and Second Life [11] are only just the beginning of a new era for the online games. The idea that lies behind online gaming is that a game that can be played by multiple users should not have only a local context. The basic game software is installed on the client machine, while multiple servers are assigned with the task of interconnecting all the possible users to what is called the "world" or the scenario of the game. The Games at Large project, as described in the official website, goes one step further than the classical procedure of online gaming and the main intention is to enhance the idea of application on demand [6], in order not only to support games on demand, but also to enable devices that lack the physical power to load a game, to run games **Error! Reference source not found.**[1][4][13].

The proposed architecture resembles that of thin-client computing [17], consisting of a server and a client that communicate over a network using a remote display protocol. Graphical displays are virtualized and served across a network to a client device by the protocol, while application logic is executed on the server. By using a remote display protocol, the client transmits user input to the server, and the server returns screen up dates of the user interface of the applications from the server to the client.

Some previous works on computing platforms include STARS [9], a unified platform that focuses on tabletop gaming, and [3], where the authors explore how computer games can be designed to regain some of the social aspects of traditional gameplay.

Many of these remote display protocols can effectively be web-enable applications without application modification. Some examples of thin-client platforms include Citrix MetaFrame [18], AT&T Virtual Network Computing (VNC) [19] and Tarantella [20]. The remote server typically runs a standard server operating system and is used for executing all application logic. Because all application processing is done on the server, the client only needs to be able to display and manipulate the user interface. The client can either be a specialized hardware device or simply an application that runs on a low-end personal computer.

The objective of secure communications has been to provide privacy or secrecy, i.e., to hide the contents of a publicly exposed message from unauthorized recipients. The asymmetric encryption / decryption channel solves the major confidentiality issue of secure communications. Cryptosystems, as explained by the classic work of Simmons [12], are symmetric if either the same piece of information (key) is held in secret by both communicants, or else that each communicant holds one from a pair of related keys where either key is easily derivable from the other. These secret keys are used in the encryption process to introduce uncertainty (to the unauthorized receiver), which can be removed in the process of decryption by an authorized receiver using his copy of the key or the "inverse key." This means, of course, that if a key is compromised, further secure communications are impossible with that key. On the other hand, in asymmetric cryptographic schemes the transmitter and receiver hold different keys at least one of which it is computationally infeasible to derive from the other.

The work on public key cryptographic systems has been rather intense over the last 20 years. The main difficulty in developing secure systems based on public key cryptography is not the problem of choosing appropriately secure algorithms or implementing those algorithms [10]. Rather, it is the deployment and management of infrastructures to support the authenticity of cryptographic keys: there is a need to provide an assurance to the user about the relationship between a public key and the identity (or authority) of the holder of the corresponding private key. In a traditional Public Key Infrastructure (PKI), this assurance is delivered in the form of certificate, essentially a signature by a Certification Authority (CA) on a public key [2].

3. The Games at Large Project

Games at Large (Games@Large) being an Integrated Project (IP) intends to research, develop and implement a new architecture to provide users with a richer variety of entertainment experience in their entire houses, hotel rooms, cruise ships and Internet Cafés, incorporating unprecedented ubiquitous game-play. The project evolved from the home environment to other local Focus Areas (FA) regarding the benefits such FA may gain based on the unique technology approach of Games at Large. The Integrated Project includes activities of TV Multimedia and Gaming using Enhanced Media Extender, Local Processing and Storage Server(s), Handheld Devices and Local Wireless Network. Games at Large intends to enhance the existing Digital Living Network Alliance (DLNA) and the UPnP Forum standards by introducing the unique set of features required for running games over a local network, like all other media and content types (video, audio).

Market interest is now revolving around capitalizing on the rapid increase of always-on broadband connectivity. Broadband connection drives to a new, digital, “Future Home” as part of a communications revolution, which will affect every aspect of consumers’ lives, not the least of which is the change it brings in terms of options for enjoying entertainment. Taking into account that Movies and Music provided by outside sources were at home long before the Internet and Broadband, the challenge is to invent new content consumption patterns and new types of content and services.

Games offer a leisure time activity for every member of the household – from avid gamers to kids, as well as allowing whole families to play together. Games offer also leisure time activity for guests in hotels and visitors in Internet Cafes. Games at Large offers ubiquitous accessibility for all members of the household on all desired entertainment devices. The project focuses on new innovative ideas such as multiple-game execution on the Games Gateway and delivery of graphics-rendering meta-data over the home network via low latency, low bandwidth Pre-Rendering Protocol to achieve low-cost implementation of ubiquitous game play throughout the house, while taking advantage of existing hardware, and providing multiple members of the family with the ability to play simultaneously.

Games at Large intends to enable the Games to diversify from dedicated appliances and a single corner of the house, to any place at home such as, the TV in the living room, the handheld device or any other device with the relevant screen, controls and connectivity. The project will also provide the required infrastructure for running games on the hotel guest room TV or on small screens for people sitting in Internet Cafés, cruise ships, trains or airplanes.

The technological challenges of the the Games at Large project are:

- Distributed computing and storage
- Video/Image/Graphics delivery with very low latency through a wired/wireless home network
- Adaptation of PC screen-images to TV screen and handheld devices
- Integration of wireless users’ game control devices

- Translation of user ergonomics to different devices and form factors
- Research of new class of Media Extenders for games
- Enhancement of STBs to support video games
- Development of new methods for QoS linking Consumer prospective with system measurements
- Enhancement of relevant industry standards for time critical multimedia content while maximizing Users Experience
- Security aspects for the system's architecture as a whole and for each subsystem independently

The Games at Large project's mission is to develop a new method for ubiquitous video games through unique technology to transfer graphical data while reducing latency and ensuring QoS in a cost-effective manner. Main focus will be given on studying and supporting the use of video games within four different focus areas: User's home, Hotels, Internet Café, and Elderly Houses. A multi-layer approach will cut horizontally across the Games at Large focus areas, aiming to assess the conditions under which a Games at Large platform may frame within and improve the state of the art of each business domain, through performing the following, logically consecutive activities: collecting user requirements, researching and developing common Technologies, implementing and integrating those technologies within the required Servers and prototype CE Devices, running technology verification and Training and evaluating all results.

4. System Architecture

Figure 1 depicts the general system architecture. As it is obvious, the system consists of two different "levels". The first level includes all the possible servers that will be used for the system, while the second level includes the connection of the different end devices of the system. The server side constitutes of multiple different servers that are assigned with the task of serving the games and require a very quick and stable communication between them, which is guaranteed using a wired LAN network. The second level of the depicted architecture is the interconnection of any possible end device with the server in order to communicate and interact so as to load and play a game. Connected clients can utilize a variety of end devices, such as set top boxes, laptops, PDA's or IPTVs.

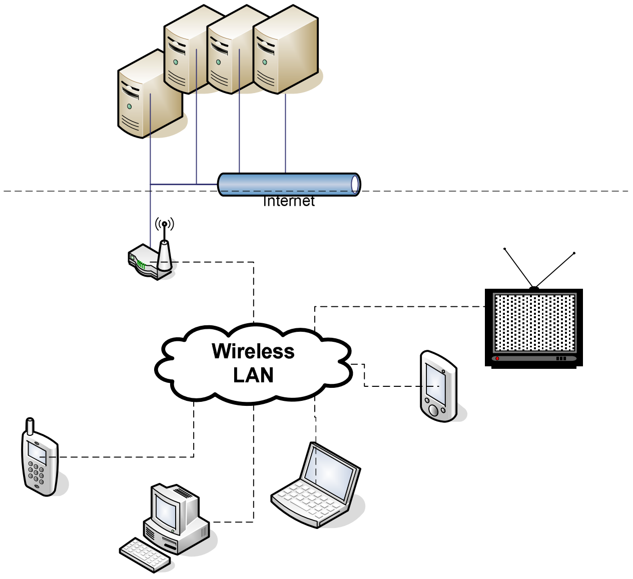


Figure. 1 General System Architecture

While on the architecture that is described all the servers can communicate to one another, the end devices can "see" only one server which is the main serving and processing server for the games. The server side of the system is assigned with various tasks, the most important of which is that of executing the game and sending the corresponding game screen to the connected clients. The Local Processing Server (LPS) coordinates the server infrastructure, and is responsible for the execution of the game graphics and the delivery of data between the clients and the server.

The clients are constantly sending feedback to the server which describes the input commands that are to be executed to the game instance. Thus, the server should be able to have at least two communication channels with each client: one for sending the game frames or 3D commands (direct channel), and one for receiving the input from the clients of the game (reverse channel). An important aspect to notice is that the channel which, if hijacked, could jeopardize the system's security, is the return channel since it contains not only the input commands that are for execution to the game instance, but also any other input from user. For instance, given the fact that the platform is targeted for commercial use, it is possible that the users will be required at some point to insert personal information, passwords, or even a credit card numbers. Hence, the encryption of the command communication channel and more specifically, the encryption of keyboard input commands is of major interest.

5. Command Channel Infrastructure

The idea that lies beneath the communication command channel architecture is depicted in the flow diagram of Figure 2. Each end device consists of many possible input devices that enable the user to interact with the device and thus enable the user to interact with the game that is played on the end device.

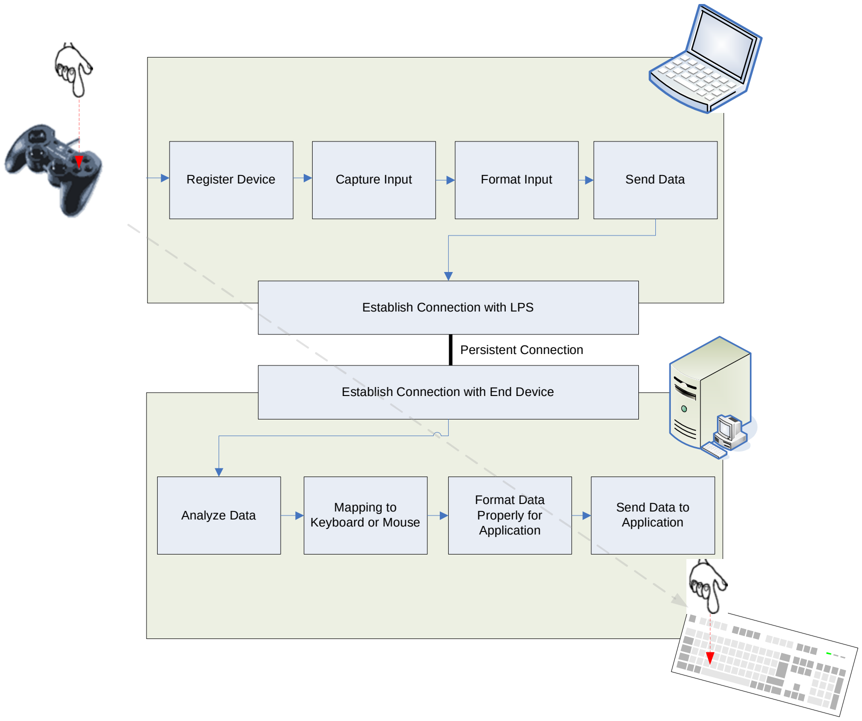

Register Device
Capture Input
Format Input
Send Data
Establish Connection with LPS
Persistent Connection
Establish Connection with End Device
Analyze Data
Mapping to Keyboard or Mouse
Format Data Properly for Application
Send Data to Application

Figure. 2: Communication Channel Architecture.

When the client program starts, it initiates the device discovery procedure, which may be offered either by a separate architectural module, e.g. the device discovery module which uses UPnP, or by a system call causing the discovery for input devices attached to the system. It is essential afterwards, that the results of the device discovery are registered in our program so that we are aware of the existing input devices marking out several other non-existing.

The next step of the procedure is to capture the input coming from the input controllers. This is achieved by recording the key codes coming from the input devices. Input devices such as mice or keyboards are interrupt-driven while with joysticks or joy pads the polling method is used for reading. The previous means that whenever an input event is caused by a keyboard or a mouse, an interrupt message is sent to the message queue of our program; then it is translated and finally recorded. However, the polling case of joysticks or joy-pads means that these devices have to be polled by a program's thread in order to sense motion or button presses. The polling period has to be small enough to capture any input, but not too small to monopolize the system's CPU. A period of 10ms seems to be in our occasion a wise trade off.

After an input key code has been captured, the transmission of it takes place. This is achieved using an already open socket connection with the server side. Data is transmitted through the socket in the form of a string with a certain communication protocol. The socket connection can either be of TCP or UDP protocol. Since UDP emphasizes on real time, low latency transmission, it is preferable for this type of communication. Even if some key codes are lost in the process of transmitting them over the network, there is no real loss since there is a flow of key codes that can overcome this possible threat. Even though, in real life, error prune networks, such as WiFi's, the TCP protocol is selected avoiding the possible game experience fall caused by lost controller's packets transmission, for wired networks UDP is preferred.

Since the key codes have arrived at the server side, they are executed at the running game instance. At this point, there needs to be a distinction between the different types of transmitted key codes. There are basically four types of possible input device’s data transmission. Commands may be coming from: (a) keyboard, (b) mouse, (c) joystick / joy-pad device or (d) any other HID input device.

In the first case, the server has to recognize the virtual key code, or the “pressed” / “released” event of a keyboard button, then do a possible mapping to some other key code, based on the game and user profile, and finally deliver it to the active application window for execution.

In the case of mouse input, the server has to recognize the virtual key code or the “pressed” / “released” event of a mouse button, recognize any mouse wheel event or any mouse movement (absolute or relative), then do a possible mapping to some other key code, based on the game and user profile, and finally deliver the key code to the active application window for execution.

In the case of joystick/joy-pad input, the server recognizes the state of the joystick/joy-pad device, maps the state to the appropriate keystrokes using the xml mapping file of the particular game-joystick/joy-pad combination. In this way, we are able to emulate the joystick/joy-pad input using pure keystrokes–mouse movements that represent the actual behavior of the input device. Finally the key code is delivered to the active application window for execution.

For any other HID input device the system treats input similarly to the joystick/joy-pad. The only prerequisite is the existence of a mapping file in order to convert the commands to keyboard and mouse instructions

6. Encryption

As already noted, the encryption procedure is only needed for the keyboard commands that the client transmits. We will now briefly describe the initialization procedure for supporting RSA public key encryption both at the client and the server of the Games At Large environment, as well as the thereafter communication between the LPS server and the connected client.

6.1 Startup phase

When both the client and the server start, some local initializations take place. Following, the client launches a connection request to the server which is advertised to the network neighborhood through the UPnP module. The server accepts the new client generating a unique RSA public-private key combination. Initially, through the persistent connection, the server transmits the modulus size in bits, the public exponent size in bits and the key pair size in bytes of the encrypted fields that follow.

The public key is described by an RSA structure and its fields are transmitted sequentially to the client though the possibly unsafe channel. The client then accepts the structure’s fields and re-generates the server’s public key. From this point on, any keyboard commands are encrypted by the client using the server’s public key and decrypted by the server providing thus the necessary security guarantee for the user-sensitive data. The aforementioned procedures are depicted in Figure 3.

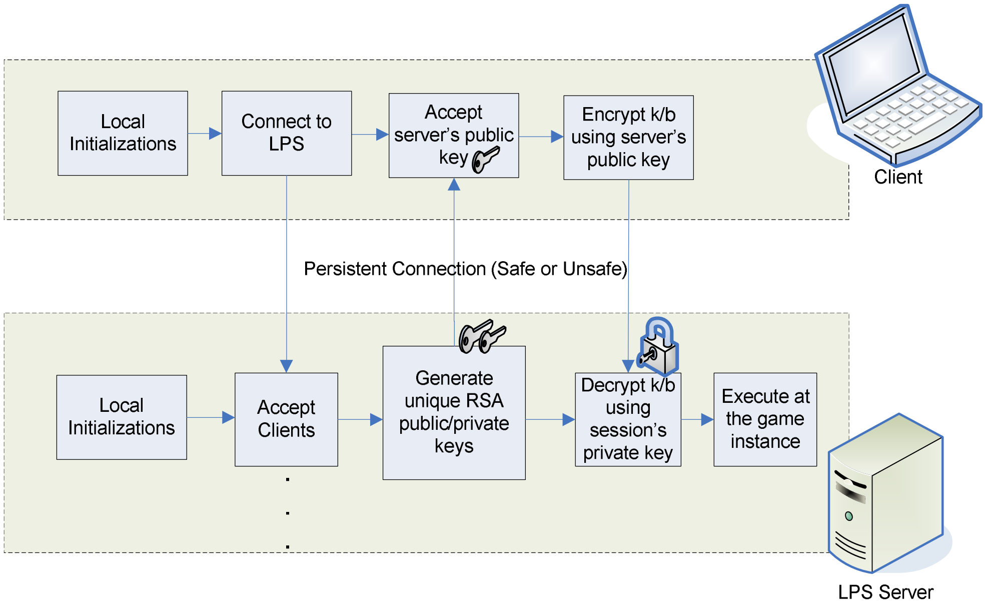

Local Initializations
Connect to LPS
Accept server's public key
Encrypt k/b using server's public key
Client
Persistent Connection (Safe or Unsafe)
Local Initializations
Accept Clients
Generate unique RSA public/private keys
Decrypt k/b using session's private key
Execute at the game instance
LPS Server

Figure 3 Initialization of the encryption module

6.2 Transfer of encrypted keyboard input

The idea that lies beneath the communication command channel architecture is depicted in Figure 4. Each end device consists of many possible input devices for interacting with the server. When the client program starts, it initiates the device discovery procedure, which may be offered either by a separate architectural module, e.g. the device discovery module which uses UPnP, or by a system call causing the discovery for input devices attached to the system. It is essential afterwards, that the results of the device discovery are registered in our program so that we are aware of the existing input devices marking out several other non-existing.

The next step of the procedure is to capture the input coming from the controllers. This is achieved by recording the key codes coming from the input devices. As already mentioned, input devices such as mice or keyboards are interrupt-driven while with joysticks or joy pads the polling method is used for reading. If the command that is to be transferred is originating from a keyboard device, the client uses the server's public key to encrypt the data after it has been suitably formatted adhering to a certain communication protocol. The encrypted message is transmitted to the server using an already open socket connection.

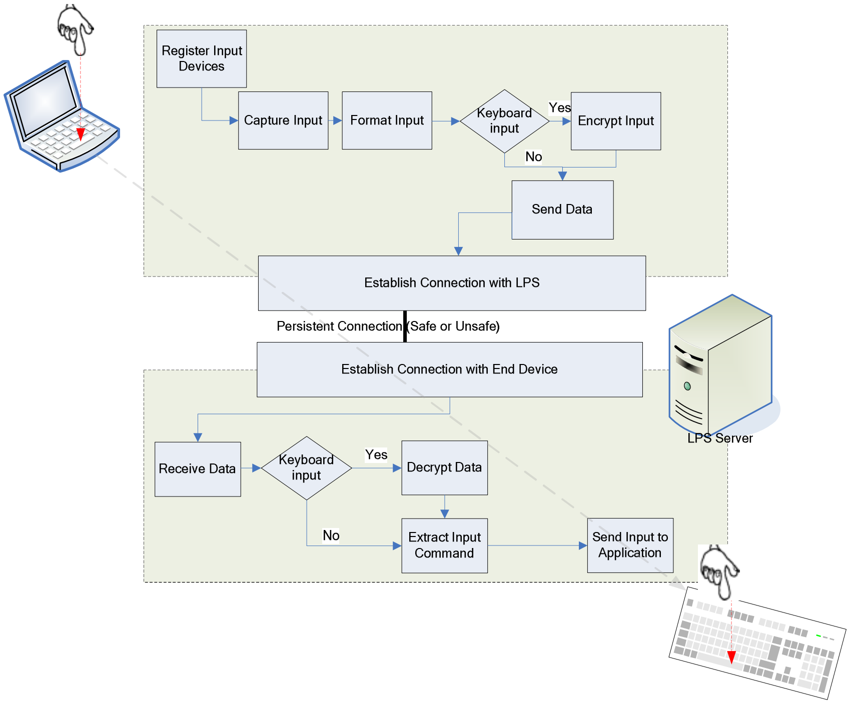


Figure 4 Encrypted Command Channel.

Once the encrypted message has arrived at the server side, the server decrypts it obtaining the initial keyboard commands that the client captured. If the received massage is not a keyboard one, the server bypasses the decryption stage, delivering the commands at the running game instance. The algorithm procedure of this step is presented in Algorithm 1.

```
//-- Client Encrypts and Send Keyboard Data
int encrypt(string message) {
//pk_size is the public key size
server.send_data(keyboard_type);
//notify the server for
//keyboard command that follows
unsigned char *encrypted;
int enc_size =
RSA_public_encrypt(strlen(message)+1,
(unsigned char *)message,
encrypted, PUBLIC_KEY, PADDING);
if (enc_size != pk_size)
{
Error("Ciphertext should match length of key");
return(-1);
}
//-- send encrypted data
return server.send_data((char
*)encrypted,enc_size);
}

//-- Server Receives and Dencrypts Keyboard
Data
int dencrypt(string encrypted) {
unsigned char *decrypted ;
char temp [MSG_SIZE];
//-- receive encrypted data
client.receive_data(temp,kp_size);
memcpy((char *)encrypted,
temp,kp_size*sizeof(char));
int decr_length = RSA_private_decrypt(kp_size,
encrypted, decrypted, PRIVATE_KEY,PADDING);
if(!decrypted){
Error("Encryption failed");
return -1;
}
Retrieve_vkey(decrypted);
}
```

Algorithm 1 Encryption and Decryption of keyboard messages

7. Server / Client Infrastructure

In this section, we describe the infrastructure that was implemented within the scope of the Games At Large project both at the server and the client side.

7.1 Server Side Infrastructure

As long as we are creating an environment with one server and multiple clients, it is essential to analyze how each end device will be able to capture all the commands from the input devices. This is because the unique server of the system should receive data that are sent over the network and execute the commands on the specific procedure that runs each game.

The "gateway" of the servers is the LPS (Local Processing Server). The main goal of Local Processing Server is to run multiple games simultaneously on the server, whereas each game runs in its own game environment and is streamed to an end-device. The game environment is an isolated and encapsulated "sandbox," providing the environment for game execution. The procedure, that makes the simultaneous running of multiple games possible, decouples the game execution from the game output, directed to display card/PC monitor, and all user-facing I/O, directed to the keyboard/mouse/HID. The LPS server also implements the encryption policy of the system by generating random RSA public/private key pairs for any newly connected clients and by decrypting the keyboard commands that come from the clients.

The "sandbox" environment for the server is created dynamically according to: a) the current occupancy of the resources and the hardware requirements that the game sets on the server, b) the software requirements on the client side and c) the current network condition. In order to be able to run a game on the server, the system monitors in a periodical manner the hardware resources of the server and the network conditions (jitter, latency and bandwidth). Additionally, according to the end device specifications (hardware and software), the server decides on the manner that the game will be executed on the client side.

7.2 Client Side Infrastructure

The possible different clients of the Games At Large environment are: (a) a Laptop with Windows XP / Vista environment, (b) a Set-Top Box with either Linux or Windows CE and (c) an enhanced handheld device with either Windows CE or a Linux version for small screen devices.

Each client implementation should consist of the following components: (a) a device discovery module, (b) the game browser and game launcher modules, (c) authentication modules, (d) input capturing and command transferring modules and finally, (e) decoders in order to run the streamed game that is sent from the server.

The device discovery module is used to seek for an appropriate LPS to connect to and introduce itself to the LPS with the End Device capabilities. The Games At Large Game Browser, which queries the Games Service on the LPS for listing the available games to the user, enables the user to browse the list of available games and select one to launch. Personalization of the UI should be available to the user/provider for enabling different views for users (i.e. Browser skins). When the user selects a game and requests to launch it, the Games At Large Game Browser issues a Start Game request to the Games At Large Client Game Launcher. The Game Browser will show to the user only the list of games that can run on the End Device by filtering the list according to the capabilities of the End Device compared with each game requirement.

Authentication communicates with the Games At Large Game Browser to authenticate the user against the LPS authentication module that authenticates the user against the Management Server. The Client Game Launcher controls all modules on the client side. The Game Launcher communicates with the LPS discovered by the device discovery module. The capture controller captures the Human Input Device (HID) controls and transfers them to the Controller Emulator on the LPS via the network layer, using the Controller protocol.

Each client should also implement the necessary RSA functions for the encryption module. For this cause, we are utilizing the OpenSSL RSA library which is available for the aforementioned platforms [14].

7.2.1 Capturing commands in Windows OS

As already mentioned, the end devices of the system can be multiple and thus they may use various operating systems, whereas the server is based on Windows operating system. When the client utilizes Windows operating system, the implementation of the modules, and more specifically, the command capturing module, is implemented as a generic driver. This driver is able to recognize any input device and transform the command from them to keyboard and mouse commands, according to mapping files that are utilized for this scope. The aforementioned is a windows application, included in the game launcher of the client, which is called reverse channel module. The main assignments of the reverse channel module is a) to ensure that the connection to the server is established successfully, b) to capture commands from any input device, and c) to send the commands over the channel that is present between the client and the server.

7.2.2 Capturing commands in Linux OS

When the end device utilizes Linux operating system, there is no need for a low level driver (as in the Windows case) to be implemented as a client-side program in order to capture the input devices' input. On the contrary, the command capturing is feasible through the evdev Xorg input driver and the evbug capturing implementation, both of which are available to any modern Linux kernel. Finally, the keycodes from any input device are translated to keyboard and mouse commands and are then transmitted to the server for execution in the game instance.

8. System Evaluation

Extensive testing that is done by teams of the project on the issue of delay over the network has proved that in order to support games that require instant action (like racers or shooters) it is essential to have at most 50ms latency.

In order to be able to have latency of less than 50ms and as long as a large amount of them must be used for the transfer of the game graphics from the server to the client it is important that the return channel has as lowest latency as possible. The evaluation proves that an encrypted channel with dynamic encryption on each command sent from the client to the server requires 10 to 20ms regardless the game that is played. What we need to do is to lessen this latency in order to have the minimum overhead from the return channel.

To achieve lower latency we apply a different type of encryption on the channel. At the initialization phase we apply all the encryption as it is described in section 6 but not in order to send commands but in order to create a secure channel. As long as the secure channel is initiated we are assured that each byte on the channel is encrypted by the channel itself. This means that we are transferring the encryption of the data to a lower level of the ISO/OSI network layering system. From the application level of encryption we are moving to a network level by creating a single secure channel.

This approach has led to less latency on the return channel which is almost as fast as a pinging command from the client to the server. Nevertheless, we already know that we are sending a very small amount of data per second which is usually hundreds of bytes and in average even less. The latest testing that was done on the return channel (specific results cannot be presented due to the confidence terms of the project) prove that the latency of the return channel is at most 5ms.

Considering the scaling of the system, although it is out of the scope of this specific document we can present some preliminary results. By using a server which was in the state of the art during the year 2007 (intel core2 duo 2GHz, 2GB RAM), we have managed to play simultaneously more than 15 casual games from different clients. When talking about demanding games the testing that was done proved that we can run 4 demanding at the same time.

As the project is currently under its third year of running we are not able currently to present extensive results on the evaluation of the system.

9. Conclusions

In this study we have described the command execution channel as well as the encryption module of the Games at Large project, an IP project with the vision to research, develop and implement a new architecture to provide users with a richer variety of entertainment experience in their entire houses, hotel rooms, cruise ships and Internet Cafés, incorporating unprecedented ubiquitous game-play. We are researching and utilizing new technological techniques to transfer graphical data while reducing latency and ensuring QoS in a cost-effective manner. Main focus is given on studying and supporting the use of video games within four different focus areas: User's home, Hotels, Internet Café, and Elderly Houses. A multi-layer approach cuts horizontally across the Games at Large focus areas, aiming to assess the conditions under which a Games at Large platform may frame within and improve the state of the art of each business domain, through performing the following, logically consecutive activities: collecting user requirements, researching and developing common Technologies, implementing and integrating those technologies within the required Servers and prototype CE Devices, running technology verification and Training and evaluating all results.

10. Future Work

As the system is implemented, more and more features are included on the release versions. These include modules that utilize network specific characteristics in order to adapt on the possible network environment (QoS support). These characteristics are expected to guarantee a minimum level of quality to any connected client, either utilizing a fast wired network or a noisy wireless one. Additionally, efforts are made towards the direction of creating software for every possible operating system in order to enable more end-devices to be connected to the Games at Large Environment. Furthermore, in our plans is also the incorporation of a media streaming server to the gaming infrastructure which will allow the connected clients to enjoy their preferred music or video clips with music/video on-demand characteristics.

References

[1] C. Bouras, V. Poulopoulos, I. Sengounis, V. Tsogkas. "Networking Aspects for Gaming Systems", In Proceedings of the Third International Conference on Internet and Web Applications and Services pp. 650-655, 2008

[2] P.S.L.M. Barreto, H.Y. Kim, B. Lynn, and M. Scott. "Efficient algorithms for pairing-based cryptosystems," In Advances in Cryptology – CRYPTO 2002, volume 2442 of LNCS, pages 354–368. Springer-Verlag, 2002.

[3] S. Bjork, J. Falk, R. Hansson, P Ljungstrand, Pirates! Using the Physical World as a Game Board, Interact 2001

[4] C. Bouras, V. Poulopoulos, I. Sengounis, V. Tsogkas "Input here - Execute there through networks: the case of gaming". The 15th Workshop on Local and Metropolitan Area Networks (LANMAN 2007), Princeton, NJ, USA, 10 - 13 June 2007

[5] P. Casas, D. Guerra, I. Irigaray, User Perceived Quality of Service in Multimedia Networks: a Software Implementation, Joint Research Group of the Electrical Engineering and Mathematics and Statistics Departments, 2006

[6] Games at Large project's official website, http://www.gamesatlarge.eu

[7] Half Life official website, http://orange.half-life2.com/

[8] IPTV, By 2010, One-Third of the Predicted 422m Broadband Households will be Able to Receive IPTV. http://www.findarticles.com/p/articles/mi_m0EIN/is_2006_Sept_26/ai_n16837715

[9] C Magerkurth, R. Stenzel and T. Prante, STARS - a ubiquitous computing platform for computer augmented tabletop games. In Extended Abstract of UbiComp '03, Springer, 267—268 2003

[10] S. Sattam, Al-Riyami and K. G. Paterson, "Certificateless Public Key Cryptography," Lecture Notes in Computer Science, pp. 452 - 473, 2003

[11] Second Life official website, http://www.secondlife.com

[12] G. J. Simmons, "Symmetric and Asymmetric Encryption," in ACM Computing Surveys (CSUR), vol. 11, no. 4, ACM Press New York, NY, USA 1979, pp. 305-330.

[13] Y. Tzruya, A. Shani, F. Bellotti, A. Jurgelionis, Games@Large - a new platform for ubiquitous gaming, BroadBand Europe 2006, Geneva, Switzerland, November 2006

[14] J. Viega, M. Messier, and P. Chandra, 2002. Network Security with OpenSSL, 1st Ed. O'Reilly, Cambridge, MA.

[15] World of Warcraft official website, www.worldofwarcraft.com

[16] Worldwide online access. http://www.emarketer.com/Report.aspx?bband_world_jun06&src=report_summary_reportsell

[17] Albert Lai , Jason Nieh, Limits of wide-area thin-client computing, Proceedings of the 2002 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, June 15-19, 2002, Marina Del Rey, California

[18] Citrix MetaFrame 1.8 Backgrounder. Citrix White Paper, Citrix Systems, June 1998.

[19] T. Richardson, Q. Stafford-Fraser, K. R. Wood, and A. Hopper. Virtual Network Computing. IEEE Internet Computing, 2(1), Jan./Feb. 1998.

[20] A. Shaw, K. R. Burgess, J. M. Pullan, and P. C. Cartwright. Method of Displaying an Application on a Variety of Client Devices in a Client/Server Network. US Patent US6104392, Aug. 2000.

Distributed GIS Approach for Flood Risk Assessment

Rifaat Abdalla
Disaster and Emergency Management
Program, York University, 4700 Keele
Street,Toronto, ON, Canada M3J 1P3
Email: abdalla@yorku.ca

Abstract – **Web-based Geographic Information Systems (WbGIS) provides key decision support capabilities for the disaster and emergency management community. Perspective visualization and simultaneous access to emergency management data are important capabilities that WebGIS can provide in support of informed decision-making process. By using a case study on a section of the Don Valley in Toronto, Canada, this paper will present a WebGIS based interdisciplinary approach for flood risk assessment and will demonstrate the utility of WebGIS in simulating different what-if scenarios under different water surface elevations. A visual model of the extent and the impact was published in the web using GeoServNet (GSN), a proprietary WebGIS package. The article highlights the capabilities of WebGIS and addresses some of the key issues that prevent a proper emergency response. Issues like time of geospatial data acquisition, maintenance, processing and update can be a challenge during emergencies. This article is of importance to decision makers in public safety and national security domains, as well as to military personnel working at the operational level.**

1. INTRODUCTION

Numerous fields including environmental planning and management, agriculture, hydraulics engineering, and earth science contribute to flood simulation research and prediction studies. This supports the process of reaching to quantitative understanding and accurate simulation of flooding scenarios based on GIS. The availability of accurate data and efficient modeling tools are bounding factors for effective flood risk assessment [1,2]. Modeling environmental and physical processes for the purpose of disaster and emergency management is a complex process because real-world phenomenon are typically 3D, time dependent and complex [3]. These factors make it difficult to obtain a complete qualitative and quantitative understanding of these processes. Is not a challenge to represent 3D features on a desktop system and publish the output on a webservice [2]. Spatial data collection is further developed by new technologies, such as Light Detection and Ranging (LIDAR), Synthetic Aperture Radar (SAR) and high-resolution satellite imagery by GeoEye and Quickbird has further enhanced 3D spatial information processing, and contributes to making geospatial information acquisition and processing much faster and cheaper [4, 5]. All these developments in computing power, software, internet bandwidth, and data acquisition have enhanced the power of 3D WebGIS and its ability to show change and communicate complex geospatial phenomena. The added dimensionality of 3D WebGIS allows geospatial modelers to move themselves from primitive representation of fence diagrams, isometric surfaces, multiple surfaces, stereo, and block diagrams [6]. The continuous improvements in hardware and software technology will ensure that Web-based 3D GIS become easier to implement, with a wide range of applications.

This study aims at presenting an interdisciplinary approach for publishing and visualizing flood information using GSN and examining the utility and efficiency of 3D WebGIS decision-support capabilities in support of disaster and emergency management operations. What-if simulated scenarios are used in delineating different water surface elevations for the assessment of possible impact on critical infrastructure and land use classes.

An integrated WebGIS approach for flood risk assessment is introduced in the first and the second section of this paper. The third section will introduces an overview of the fundamentals of disaster risk management. A case study will be presented in the fourth section to show the procedures to be used for integrating different data sources and modeling tools to provide hydraulics simulation driven visual models. The fifth section will present the results and discusses the advantages and challenges for using distributed GIS in flood risk assessment. The last section will present findings and conclusions of this work.

2. GEOSPTATIAL INFORMATION TECHNOLOGY

The distribution of geospatial information over a network of connected computing nodes has facilitated effective disaster management operations, through timely access to data and through the ability to effectively interpret and visualize disaster management scenarios. Estimating disaster damage and predicting its impact are among the contributions of the progressive development in Geospatial Information and Communications Technology (GeoICT). WebGIS as a GeoICT element is a centrally managed and distributed computing architecture. Distributed computing is a generic term that includes other terms like Internet, Intranet, the web, network-centric, and more. The growing trend is to distribute computing services across a physical infrastructure of networked data storage devices and computer processors. This environment includes both a two and three tier model where the physical locations of the data storage and application processing are not on the same machine.

3. DISASTER MANAGEMENT

Many researchers including [18, 23, 24] have struggled to define disaster. Although disasters have the common result of leaving behind devastation and loss, there is no precise definition for the term "disaster". The bottom line in disaster management is that loss of life and property should be eliminated or minimized, basic needs should be ensured, and business continuity should be secured. The basic requirements for disaster management can be achieved only through interdisciplinary efforts and there are no research methods that are unique to this field [14]. According to [21, 22] modern disasters are complex and diverse phenomena with a greater potential for adverse impact. For many years the effects, impacts, and issues pertaining to protection from disasters have been the focus of many researchers including [15- 20]

Flood disasters impact the economy, natural resources, lives, property and physical infrastructure. The magnitude of the impact depends on the emergency measures and the steps taken by the concerned authorities during the preparedness and the response phases. Providing accurate and quick information over the internet could help reduce the loss. A standardized community-based risk assessment protocol was developed by Emergency Management Australia [24]. This framework is based on six major elements as shown in Figure 1 and discussed below:

1- Risk Context: The first phase is related to the establishment of risk context. Issues related to the problem at hand and the approaches of solving it are discussed in this phase.
2- Hazard Vulnerability: The second phase involves identifying risks in terms of hazard and vulnerability. The scope and nature of a hazard must be identified, as well as the setting of the community at risk.
3-Risk Analysis: The third phase in this process is risk analysis. In this phase, tools of problem analysis, for instance modeling software, are used to analyze risks associated with the problem identified in the risk context phase.
4-Risk Evaluation: The fourth phase is risk evaluation, which involves prioritizing the risk and comparing it against risk evaluation criteria. Risk thresholds are also established in this phase.
5-Communication: The last phase in this process deals with treating risks according to the result of the evaluation. Results obtained from the risk evaluation phase will be communicated to the concerned stakeholders to allow them to implement disaster management measures.

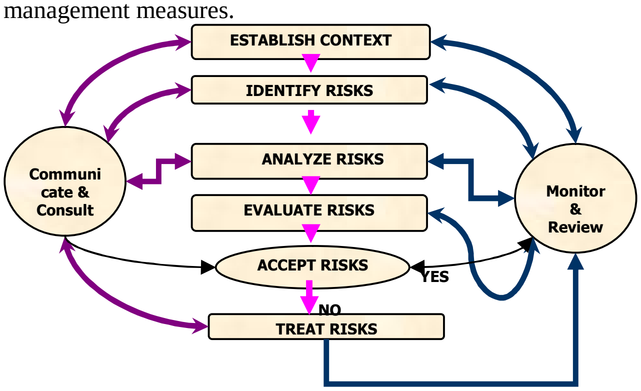


Figure 1 Flowchart showing the elements of risk assessment process (after EMA, 2002)

4. WEBGIS FOR DISASTER AND EMERGENCY MANAGEMENT

Disasters are dynamic processes [8] and are spatially oriented [9]. According to [10] most current tools that are used for disaster management focus on the temporal component of the four phases of disaster management, leaving an obvious gap in dealing with the spatial element. Emphasis on the spatial dimension makes GIS technologies ideal for simulating the complex spatial relationships during extreme situations, while still being able to integrate other modeling tools.

The importance of WebGIS stems from its accessibility to many users. There are many authorities involved in planning, decision-making, and communications during disaster management operations. Desktop GIS does not provide instant and effective multi-user platforms for the same project, which require distributed GIS capability. .WebGIS provide ease of use in terms of the technical background required from user perspective.. Many decision-makers with limited or no GIS background can access geospatial information simultaneously. Decision-makers are generally divided into two general groups: response teams working in the field and decision-makers working in emergency operation centers (EOC). The EOC group works in different subgroups; communications, planning, and prediction, various sources of information can be gathered and used for disaster response and WebGIS is mostly used by the planning group as well as by field personnel and relief workers who need to access information about the current situation.

5. CASE STUDY

This section demonstrates the utility of GIS interoperability for enhancing emergency management operations. The focus here is not in showing modeling results; rather it provides demonstration scenario that addresses various issues related to how GIS interoperability can be used for emergency management. In particular, issues related to data and systems heterogeneity.

The Don River is a unique river system, as it flows through the core of the Greater Toronto Area (GTA). The origin of the Don River is located at the Oak Ridges Moraine where the headwaters are fed by numerous aquifers. The river then flows for 38 kilometers to Lake Ontario and provides drainage for 360 square kilometer of land [11]. The section used for this study is part of the huge watershed in the province of Ontario. This section is located within the extent of North York municipal boundaries. The location of the study area is shown in figure 2.

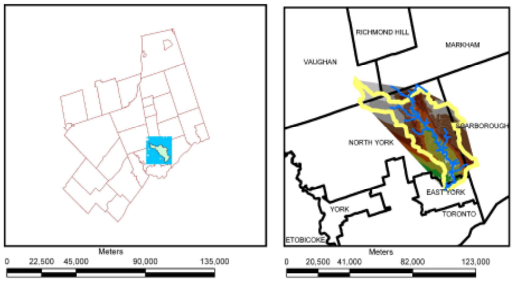


Figure 2. Map of Greater Toronto Area showing the study area.

Topographic digital maps in the form of shapefiles and a 10 meter spacing Digital Elevation Model (DEM) were collected. Topographic sheets of the area were used as reference for conducting the study. Hypothetical flow data were simulated using Canadian Hydrographic Service (CHS) data as a reference. The land use classes of the study area were used in flood analysis especially to know the extent of flood damage with respect to different classes in study area. The detailed land use map of the study area is show in figure 3.

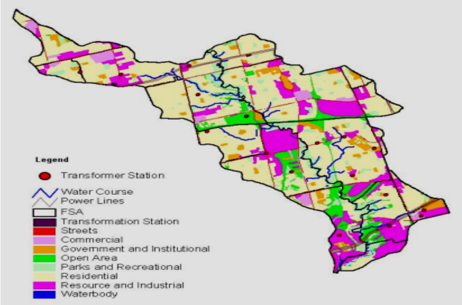


Figure 3. Detailed Land use of the Don Valley Watershed Area, Toronto, Canada

6. METHODOLOGY

ArcView GIS, Hydrologic Engineering Centre-River Analysis System (HEC-RAS) produced by the US Corps of Engineers used in the present study. HEC-RAS is a one dimensional, steady state modeling software for hydraulics intended for calculating water surface profiles at cross-sections along a stream, for both steady and unsteady flow. HEC-GeoRAS is an ArcView GIS extension specifically designed to process geospatial data for use with HEC-RAS[5]. GeoServNet (GSN) is a distributed web-based 3D GIS developed by GeoICT Lab from York University. GSN is of three main modules: GSN Builder, GSN Administrator, and GSN Publisher. The Builder is used to build and index the raw data prior to its visualization. The Administrator is used to register the layers built under builder and to complete the server and security setting before their publishing. The Publisher is used for setting visualization parameters, layer display and rendering functions in addition to the 3D perspective visualization capability. The system architecture of GSN makes it an ideal WebGIS Java technology, which provides reliable, fast, secure, and cutting-edge interoperable GIS capabilities. The system architecture of GSN is shown in figure 4[12].

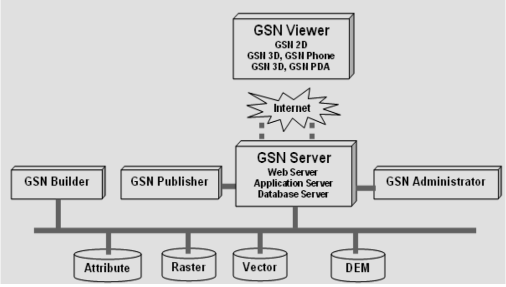


Figure 4. GSN Architecture

The methods followed in this study are: 1) Preparation The methods followed in this study are: 1) Preparation of different data layers, 2) Prediction of flooding Scenarios, 3) Preprocessing 4) Postprocessing 5) Calculating the flood damage 6) Publishing and visualization of data using GSN.

The preparation of an accurate database is imperative for successful WebGIS-based disaster and emergency management decision-making. This helps with providing accurate and timely flood information. In this phase, different data layers of the Don Valley in the form of shapefiles, grids or Digital Elevation Model (DEM) were inputted into the GIS database. The DEM was processed to produce a Triangulated Irregular Network (TIN) and a complete dataset was visualized using a standard web browser.

The analysis that was performed to simulate the Don Valley Watershed involved two major processing stages, preprocessing and postprocessing. Preprocessing is the first stage in the delineation of flooding scenarios. It involves a number of tools that are used for the creation of files used in developing the geometry of the Don Valley model. Five themes were created in this step; the stream centerline, banks, flow path, cross-sectional cutline, and land use themes. The stream centerline theme is used to establish the river reach network. It was necessary to create upstream endpoints before creating the reach in the downstream. River reach was labeled as an identifier and a reach name was assigned. Then it was possible to write the River Analysis System (RAS) GIS import file. The RAS GIS import file was created, and a complete file with a header, stream network, and cross section information was generated. After completing the process of developing stream network, five direct major steps were completed for conducting hydraulic simulation. The first step involved creating a new project for river simulation; the second step involved utilizing the GIS data. The GIS imported information is in the form of detailed geometry data that displays geometry derived. The third step was modeling the flow data; this was achieved in two different stages, the first was by entering flow profiles for the cross sections generated for the section of the study area, hypothetical flow rates were used in this research in order to support the concept of the approach, the second was entering water surface profiles.

Postprocessing is the final phase in conceptualizing the Don Valley watershed flood model. It is performed in three separate steps. The first step is the generation of a bounding polygon and cross section alignments. These are read directly from the GIS export file. Then, a water surface theme was created with the extent defined by the bounding polygon. The water surface elevations at each cross section were applied along the cross section alignments and the cross sections were treated as break-lines in the elevations, creating the TIN. The final step is to mesh the water surface TIN with the terrain TIN to produce the floodplain. Different flooding scenarios were predicted using different water surface profiles in the postprocessing step. Different water surface profiles were created from different water surface elevations using water flow rate parameters. Thus, it was possible to identify the affected settlements, affected population, and affected infrastructure, which is of special importance in emergency management. Immediately after a flood, the responsible authorities or organizations need the information about the affected area, affected people, affected infrastructure i.e. roads and rail networks, for evacuating or transportation of the required relief material.

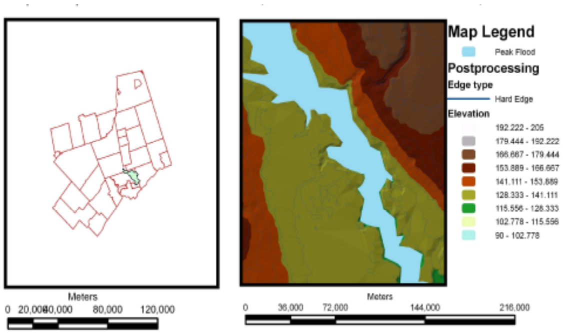


Figure 5. Triangular Irregular Network (TIN) Model for the study Area.

When the flood scenarios under different flood water surface elevation levels were predicted and damages assessed, the results were published and visualized in 3D web environment using the GSN.

7. RESULTS AND DISCUSSION

As discussed in the first section, WebGIS has become increasingly popular in disaster management. However, the integration of GIS data with remotely sensed information has posed a challenge in many occasions [13]. Determination of which type of data required for GIS modeling is a crucial issue. In this case, the first question considered was what is the data required for building this model? Is it a simple vector data for drawing maps or a complete set of data for this type of analysis? Is it specific data for specific software or data to be digitized directly?

At the preprocessing stage, RAS themes were generated. The attribution of these shapefiles according to "from - to" vector topology parameters were very important in keeping the flow direction and the DEM elevation details prior to the step of extracting the details of the model. The information gathered from the DEM integration with the digitized cross section allowed for the represention of the channel geometry. This was achieved for twelve different cross sections and provided the 3D stream channel parameters. HEC-RAS is used to represent the stream channel in 3D. Figure 6 is showing the geometric data of the Don Valley Watershed. These geometric data can be also edited in HEC-RAS.

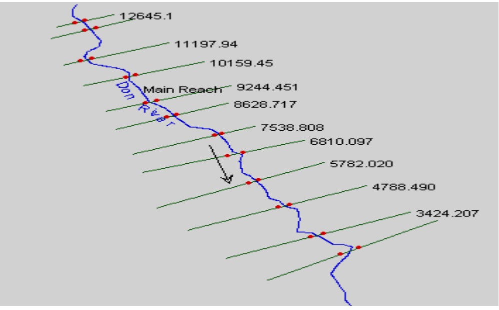
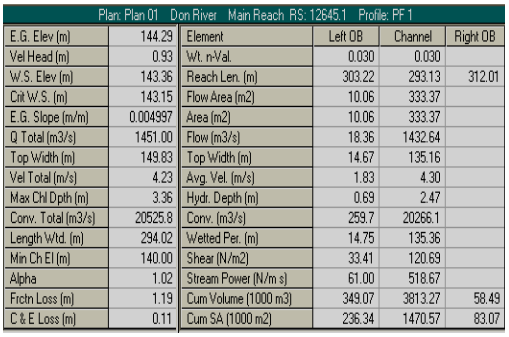


Figure 6. Geometric data of Don Valley Watershed.

Figure 7 summarizes the parameter used and a cross section output at a station of Don River, which contain i.e. elevation, slope, steady flow data, velocity, and channel depth. Different steady flow data at different stations of Don River are shown in respect of different profiles. These steady flow data can be edited and updated in HEC-RAS.



Plan: Plan 01 Don River Main Reach RS: 12645.1 Profile: PF 1

E.G. Elev (m)	144.29	Element	Left OB	Channel	Right OB
Vel Head (m)	0.93	Wt. n-Val.	0.030	0.030	
W.S. Elev (m)	143.36	Reach Len. (m)	303.22	293.13	312.01
Crit W.S. (m)	143.15	Flow Area (m2)	10.06	333.37	
E.G. Slope (m/m)	0.004997	Area (m2)	10.06	333.37	
Q Total (m3/s)	1451.00	Flow (m3/s)	18.36	1432.64	
Top Width (m)	149.83	Top Width (m)	14.67	135.16	
Vel Total (m/s)	4.23	Avg. Vel. (m/s)	1.83	4.30	
Max Chl Dpth (m)	3.36	Hydr. Depth (m)	0.69	2.47	
Conv. Total (m3/s)	20525.8	Conv. (m3/s)	259.7	20266.1	
Length Wtd. (m)	294.02	Wetted Per. (m)	14.75	135.36	
Min Ch El (m)	140.00	Shear (N/m2)	33.41	120.69	
Alpha	1.02	Stream Power (N/m s)	61.00	518.67	
Frctn Loss (m)	1.19	Cum Volume (1000 m3)	349.07	3813.27	58.49
C & E Loss (m)	0.11	Cum SA (1000 m2)	236.34	1470.57	83.07

Figure 7. Cross section output at a River Station of Don River.

At the Postprocessing stage, the floodplain was delineated using the RAS GIS export file generated by HEC-RAS. In this stage, Water Surface TIN was created from the cross sectional cutline themes and the bounding polygon theme for the respective water surface profile names. Then from the Water Surface TIN, the respective floodplain was delineated. As shown in Figure 8, three different flooding situations under different water surface profiles are shown. The flood Scenario 1, 2 and 3 were calculated using water surface profiles PF 1, PF 3 and PF 5 respectively.

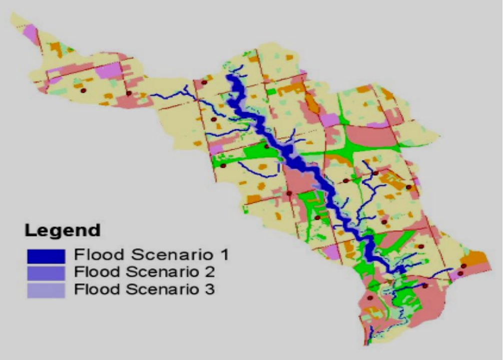


Figure 8. Different Flood Scenario at different flood situations.

Figure 8 describes the different flooding scenarios under different water surface profiles which contain different parameter e.g. water surface elevation, water discharge, and velocity. It shows the nearby infrastructure that is at risk.

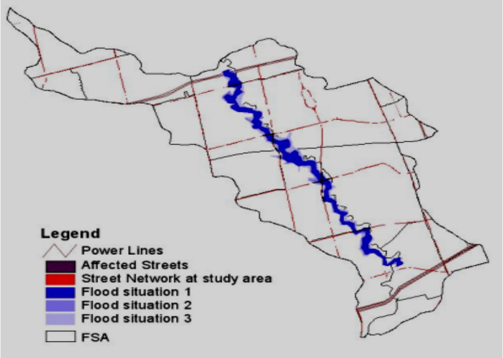


Figure9. Affected Roads under different flood situations

Once a flood hits an area, different organizations and authorities are engaged in emergency response and immediate measures are taken to manage the flood efficiently. Important information such as which areas are affected severely, affected population, affected infrastructure, and affected land use classes affected and detailed information about the affected area In this case study. The roads and rail networks are of special importance because it is used to evacuate people to shelters or safer places and to transport relief materials to relief centers in the shortest amount of time. Roads and different land use classes affected by different flood levels were calculated, which will provide the relevant organizations to know the locations of roads and other land use classes affected for emergency relief operation and efficient flood management. The 3D visualization capability of GSN is useful for better visualizing the flood-affected area in 3D web environment especially in the response phase emergency management. The published flood data was visualized over the Web, as shown in figure 9 and figure 10 respectively.

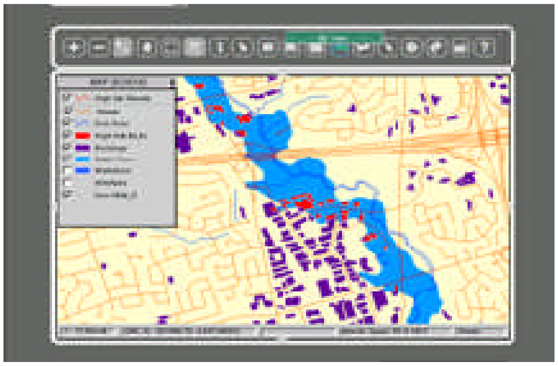


Figure 10. The flood extent of the study area. The purple polygons represent the building footprints. The red polygons are the buildings, which affected by the flood.

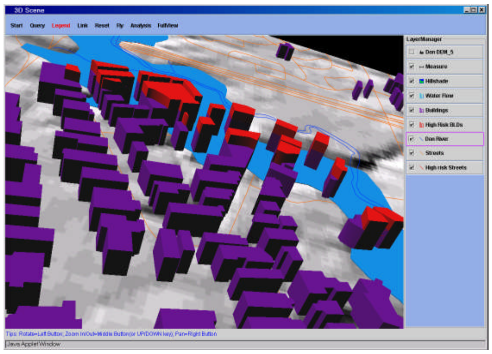


Figure 11. 3D view of the flood extend and impact shown in figure (10). The difference between 2D shown in Figure 8 and 3D is very obvious.

WebGIS provide important interoperability capabilities for emergency management departments, in form of data exchange between local, provincial and federal decision making authorities, in particular for data sharing specifications and standards. The City of Toronto demonstrated the need for using GIS Interoperability in handling many emergency management situations including the modeling of West Nile Virus, and Severe Acute Respiratory Syndrome (SARS) and in planning for two major mass festivals on World Youth Day in 2001 and the rolling stones concert in 2003. [17]

Ontario Emergency Management Doctrine [25] is the two major resources used when dealing with a disaster and or emergency. According to the Ontario Emergency Management Act 2005, there are many stakeholders responsible for deploying resources for emergency management. The Toronto and Region Conservation Authority (TRCA) is responsible for flood simulation, floodplain mapping, and water surface measurement. All hydraulics and hydrology data are under the custody of the TRCA. During emergency management situations, the TRCA will provide flood models and data through interoperable access to all decision-makers involved. Semi-real time situational awareness models can be generated and accessed on-demand based on stakeholder needs.

The City of Toronto emergency services is responsible for providing services in a variety of situations, ranging from simple road maintenance closure to extreme disastrous situations, e.g. civil infrastructure collapse.

In emergency situations the city police department utilizes GIS interoperability in a very efficient way. This includes accessing data provided by the emergency mapping department of the city and the data provided by the TRCA.

Toronto Emergency Medical Services (EMS) utilizes GIS data and information for predicting areas at high risk of experiencing an emergency and for planning how they can dispatch their services to these areas. Another important utility of GIS interoperability for EMS is that, in emergency situations, it is not easy to prioritize your response to calls from different parts of the city.

A key role for the provincial authorities represented by Emergency Management Ontario (EMO) is to monitor emergency situations and provide support on an as needed bases. GIS interoperability could provide EMO with improved situational awareness models by assembling data from all different departments and make the data available for basic analysis and visualization.

GIS interoperability provides key benefits in support of effective disaster and emergency management operations. It also allows different decision-making authorities to access information at the same time and provide transparent open access to different data sources. This benefit helps decision-makers to access multiple servers, thereby obtaining data and services that are not within organizational boundaries. Through GIS interoperability, it is possible for the sharing of a standardized data format that can be used for data transfer and information sharing in a simple manner

GIS interoperability helps to provide a simple and accessible means of integration. This is crucial since emergency management operations stand to benefit considerably from a process that allows for timely gathering, modeling and analysis of information. GIS interoperability, through its standardized protocols has allowed GIS users to utilize a simple and standard service. Through the transparency of GIS interoperability, emergency management stakeholders can readily access external GIS data and systems. Transparency here refers to the process of accessing and sharing data between systems without complicated protocols. Scalability is another advantage for GIS interoperability, as it allows for data expansion. This flexibility in scalability in systems and services is useful in emergency management operations, which, due to their dynamic, fast-paced nature require that previously unexpected situations be accommodated.

The weaknesses with GIS interoperability are related to policy and procedures. From an operational perspective, there are particular issues related to the degree to which data conversion hinders efficient data interoperability. This may arise if, for example, a particular department is using engineering data in a Computer Aided Design (CAD) form and another department is using data in GIS shapefiles formats. These two formats can be made to be compatible by

converting CAD data into shapefiles format. However, the time required for this process depends on data size and system capabilities. Data maintenance and data update is another issue that represents an obstacle for implementing GIS interoperability. Where there is no clear policy that identifies roles and responsibilities for each node in an interoperable system, data update, maintenance and management can be a challenge. On the management side, issues related to implementing GIS interoperability are related to corporate technology procurement policies, which can contribute to delayed implementation. Access rights to sensitive information, such as infrastructure and emergency management, represent another issue.

8. CONCLUSIONS

The discussed approach was dedicated to describing an integrated approach for using GIS as a tool for spatial analysis and visualization in flood simulation. This integrated approach utilizes GIS as a core technology for spatial analysis and visualization, and also integrates with other tools such as HEC-RAS for hydraulics modeling software. The dedicated approach is of special importance because it provides an interdisciplinary solution for solving real world problem. Linking WebGIS with environmental hydraulics and disaster management results in an integrated, collective, and interdisciplinary solution for addressing the reason for disaster and emergency management, which is the protection of life and property.

Despite the hypothetical nature of the scenario, it has shown the benefits that disaster management decision-makers can gain by adopting advanced and integrated WebGIS solutions in their day-to-day operations. An additional factor is the utility of 3D aspects to the simulation scenario. The contribution of the 3D visualization perspective demonstrates that WebGIS is efficient and useful for showing the impact of flooding, with particular emphasis on spatial extent of flood impact, and making it accessible for multiple users, simultaneously. This can aid decision-makers and planners to have efficient counter disaster measures and effective response plans.

REFERENCES

[1] R. Abdalla, and K. Niall, “Flood Emergency Management Scenario,” Proceedings of the Advanced Geographic Information Systems and WebServices (GEOWS) 2009. Cancun, Mexico, Feb. 1-7, 2009. IEEE Xplore.

[2] C. Tao, "Online GIServices," *Journal of Geospatial Engineering*, vol. 3, pp. 135-143, 2001.

[3 L. T. Styaert, "A Perspective on the State of Environmental Simulation Modelling," in *Environmental Modelling with GIS*: Oxford Press, 1993, p. 17

[4] G. A. Schlutz, "Use of Remote Sensing Data in a GIS Environment for Water Resources Management," in *Remote Sensing and Geographic Information Systems for Design and Operations of Water Resources Systems*, Rabat 1997.

[5] S. Doyle, M. Dodge, and A. Smith, "The potential of web-based mapping and virtual reality technologies for modeling urban environments. ," *Computers, Environment and Urban Systems*, vol. 22, pp. 137-155, 1998.

[6] F. Olivera and D. R. Maidment, "GIS Tools for HMS Modelling Support," in *The 19th ESRI Users Conference*, San Diego, CA., 1999.

[7] X. Yang, M. C. Damen, and R. A. Zuidam, "Satellite Remote Sensing and GIS for the analysis of channel migration changes in the active Yellow River Delta, China," *International Journal of Applied Earth Observation and Geo-information*, vol. 12, pp. 146-157., 1999

[8] D. Alexander, Natural Disasters. New York: Chapman & Hall, 1993.

[9] W. L. Waugh, "Geographic Information-Systems - the Case of Disaster Management," *Social Science Computer Review*, vol. 13, pp. 422-431, Win 1995.

[10] A. Montoya-Morales, "Urban Disaster Management: A case study of Earthquake Risk Assessment in Cartago, Costa Rica," Enschede, The Netherlands: Ineternational Institute For Geo-Information Science And Earth Observation (ITC), 2002.

[11] R. Abdalla, S. Liang, J. Sorrell, and V. Tao, "Visualization of Flood Mitigation Models using 3D Web-based GIS," *The Journal of the American Society of Professional Emergency Planners*, vol. 2003, pp. 65-76, 2003.

[12] GeoICT, "GeoServNet Manual," York University GeoICT Lab, 2003, p. 367.

[13] S. Maitra, "Environmental Impact Assessment for DAM Construction using GIS/Remote Sensing," in *The 21st, ESRI Users Conference*, San Diego, California., 2001.

[14] R. A. Stallings and International Research Committee on Disasters., Methods of disaster research. Philadelphia: Xlibris, 2002

[15] K. R. Murthy, Disaster management. Delhi: Dominant Publishers & Distributors, 2004.

[16] R. Palm, M. Hodgson, R. Blanchard, and D. Lyons, Earthquake Insurance in California, Environmental Policy and Individual Decision-Making. Boulder, CO: Westview Press, 1990.

[17] T. D. Schneid and L. Collins, Disaster management and preparedness. Boca Raton, Fla.: Lewis Publishers, 2001.

[18] C. Streeter, "Disaster and Development: Disaster Preparedness and and Mitigation as an Essential Component of Development Planning," Social Development Issues, vol. 13, 1991.

[19] K. J. Tierney, M. K. Lindell, and R. W. Perry, Facing the unexpected : disaster preparedness and response in the United States. Washington, D.C.: Joseph Henry Press, 2001.

[20] B. Wisner, "Bridging "Expert" and "Local" Knowledge for Counter-Disaster Planning in Urban South Africa," GeoJournal, vol. 37, pp. 335-348, 1995.

[21] E. L. Quarantelli, "The Environmental Disasters of the Future Will be More and Worse but the Prospect Is Not Hopeless," Disaster Prevention and Management, vol. 2, pp. 11-25, 1993.

[22] C. B. Rubin, Emergency Management in the 21st Century: Coping with Bill Gates, Osama bin-Laden and Hurricane Mitch. Working Paper no. 104, 2000.

[23] N. Britton, "Developing an Understanding of Disasters," Australia and New Zeland Journal of Science, vol. 22, pp. 254-271, 1986.

[24] R. Dynes, Organized Behaviour in Disaster. Lexington, MA: Health Lexington Books, 1970.

[25] Emergency Measures Ontario. (EMO). "Emergency Plans Act ": Government of Ontario, 2003.

A Legal Evaluation of Pseudonymization Approaches

Thomas Neubauer
Vienna University of Technology
Vienna, Austria
neubauer@ifs.tuwien.ac.at

Mathias Kolb
Secure Business Austria
Vienna, Austria
kolb@securityresearch.ac.at

***Abstract*—Privacy is one of the fundamental issues in health care today and a fundamental right of every individual. Several laws were enacted that demand the protection of patients' privacy. However, approaches for protecting privacy often do not comply with legal requirements or basic security requirements. This paper highlights research directions currently pursued for privacy protection in e-health and evaluates common pseudonymization approaches against legal criteria taken from Directive 95/46/EC and HIPAA. Thereby, it supports decision makers in deciding on privacy systems and researchers in identifying the gaps of current approaches for privacy protection as a basis for further research.**

***Keywords*-security, privacy, pseudonymization, e-health**

I. Introduction

Privacy is a trade-off between the patient's demands for privacy as well as the society's need for improving efficiency and reducing costs of the health care system. Electronic health records (EHR) improve communication between health care providers and access to data and documentation, leading to better clinical and service quality [2]. The EHR promises massive savings by digitizing diagnostic tests and images (cf. [3]). The pervasiveness of electronic devices has resulted in the almost constant surveillance of everyone and the permanent storage of personal data that is used and analyzed by corporations or intelligence services. With informative and interconnected systems comes highly sensitive and personal information that is often available over the Internet and – what is more concerning – hardly protected. It is a fundamental right of every individual to demand privacy because the disclosure of sensitive data may cause serious problems for the individual. Insurance companies or employers could use personally identifiable information to deny health coverage or employment. Although a variety of laws were enacted that demand the protection of privacy, only a few of the existing approaches comply with the current legal requirements. The individuals' rights are difficult and costly to pursue because they are limited in the absence of a dedicated authority to oversee and enforce compliance. The disclosure of personal data may be avoided through the use of privacy enhancing technologies (PET), such as anonymization, or more importantly, pseudonymization. Whereas anonymity allows unlinkability and maybe unobservability, it prevents any useful two-way communication. Pseudonymization ensures that a user may use a resource without disclosing his identity, but can still be accountable for that use [4].

This paper presents an evaluation of six current privacy enhancing technologies that specifically aim at protecting medical data by using pseudonymization and, thus, are used as a basis for EHR systems. The paper answers two major questions: (i) Which pseudonymization approaches adhere to the current privacy laws and (ii) what are the major drawbacks of current pseudonymization approaches. In the scope of this paper we regard evaluation as the "systematic assessment of the operation and/or the outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy" (cf. [5]). Based on the categorization of House [6] and Stufflebeam & Webster [7] we use a combination of objectivist approaches: The Testing programs approach and the Objectives-based approach. The objectives used for the evaluation are taken from the legal acts HIPAA and the EU Directive. This evaluation provides management decision makers such as chief privacy officers and chief security officers with a funded decision-making basis for the selection of privacy-enhancing technologies in the e-health area. As literature does not provide evaluations focusing on the comparison of PETs in e-health in literature so far, this paper provides a major contribution to the research area of privacy.

II. Legal Background

Nowadays, society is collecting all kinds of information. In daily life, several types of information are tracked, which are highly sensitive and can even be damaging to individuals and organizations [8][9][10]. For example, the supermarket tracks which items have been bought, mobile phone providers keep track of customer movements, airlines know what type of seat and meal is preferred and hotel chains keep records of room preferences. The exchange and storage of this information became very cheap and simple over the Internet. For this reason it is more important than ever to protect the privacy of individuals. In more than 30 countries, privacy laws protect the data of individuals [11]. The content of these privacy laws varies in each country, but they are mostly based on the Organization for Economic

Cooperation and Development (OECD) Guidelines on the Protection of Privacy and Transborder Flows of Personal Data [12].

Throughout history, collected information of individuals has been abused in several ways. Regarding the individual's privacy, historically the phrase "to be let alone", defined at the US Supreme Court in 1834, became famous. In the years during World War II, the German government abused census data to identify people of certain ethnic, religious or other targeted groups (cf. [13][14]. As various states gained in power and size, the first privacy laws were introduced in order to protect minorities. In 1948 the United Nations ratified a right to privacy in article 12 of the Universal Declaration of Human Rights. The UN declaration defines privacy as "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation". UN member countries are morally, if not legally, bound by such declarations. Everyone has the right to the protection of the law against such interference or attacks. A citizen's right of privacy is also recognized in the Article 8 of the European Convention for the Protection of Human Rights and Fundamental Freedoms from 1950. In 1966 a Computer Bill of Rights was suggested, followed by a Rights to Privacy Act in 1967 was proposed, which banned wiretapping and electronic eavesdropping.

The first national data-protection law was passed 1973 in Sweden, followed by the United States in 1974 and West Germany in 1977 [15]. In the United States, privacy has not gained much political attention. Discussions on privacy have been driven often by events in Europe. In the 1970s, concerns over privacy reached new heights, because of the abuse of wiretapping, tax, bank and telephone records during the Watergate scandal [13]. These concerns gave birth to the Privacy Act of 1974, which applies only to records of personal information held by federal agencies. These agencies are allowed to keep records only if relevant and necessary. They are not allowed to create secret files of an individual without giving the right to copy their own files. Furthermore, agencies are not permitted to disclose these records without the agreement of the individual - except within the agency for routine use or law enforcement [13].

By the end of the seventies more and more European States had passed privacy laws. To spread these laws across Europe, the Organization for Economic Cooperation and Development (OECD) published the Guidelines on the Protection of Privacy and Transborder Flows of Personal Data and the Convention of the Council of Europe in 1980/1981 that defined the provisions for the protection of individuals with regard to the automatic processing of personal data. To protect private electronic communications from unauthorized access by the government, the Electronic Communications Privacy Act of 1986 and the Computer Matching and Privacy Protection Act of 1988 have been introduced in the US. There are currently no privacy acts in the US that could be compared to the European acts. There are a handful of laws which cover the use of private data in health care [17][18][19], the electronic commerce industry [20], the cable-television industry [21] and a few other areas. A definition of personal data is given in Section 8(8) of the Online Privacy Protection Act (OPPA) [22]:

> '... information collected online from an individual that identifies that individual, including first and last name, home and other physical address, e-mail address, social security number, telephone number, any other identifier that the Commission determines identifies an individual, or information that is maintained with, or can be searched or retrieved by means of, data described above ...'

In 1995 the European Union (EU) passed the Data Protection Directive (95/46/EC) [23]. This directive applies to all personal data, which is collected or processed either electronically or in old-fashioned paper-filing systems. Article 2(a) of the Data Protection Directive (95/46/EC) defines personal data as:

> '... any information relating to an identified or identifiable natural person (data subject); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity ...'

Moreover, the Data Protection Directive (95/46/EC) is based on eight principles to which all data controllers are subject. These principles limit the usage of collected personal data [24][23][25]:

1) The data must be processed fairly and lawfully.
2) The data must be collected for explicit and legitimate purposes and used accordingly.
3) The data must be accurate and where necessary, kept up to date.
4) Organizations have to provide mechanisms to correct, delete or block data.
5) The data that identifies individuals must not be kept longer than necessary.
6) The data must be processed in accordance with the rights of the data subject.
7) Every organization must ensure the security and integrity of personal data, that they are processing.
8) It is not permitted to transfer personal data outside the European Union unless the country ensures an adequate level of protection

Furthermore, to ensure fair and lawful processing of the collected data, the data controller has to inform the data subjects which data will be collected and used. The individual also must be informed of the type of third parties the collected data will be disclosed to and the data subject must have the option to decline [23][25][26][24]. Especially sensitive data like in the health care sector need more privacy protection than non-sensitive data in other sectors. Sensitive medical data like the state of medical health, for example being HIV positive or having chronic illness, could harm a person if they are accessed by unauthorized persons. For example, an employer who accesses medical data of her employees, could use this information to dismiss an employee. Another example could be an insurance company denying a contract because of a chronic illness.

In the European Union, the Data Protection Directive (95/46/EC) [23] already implements protection for sensitive data, which are related to racial and ethnic background, political affiliation, religious or philosophical beliefs, trade-union membership, sexual preferences and health [23][25][13][24]. Besides this Data Protection Directive, an additional Working Document [27] has been released by the Article 29 Working Party of the European Union, which provides guidelines for the interpretation of the data protection legal framework for EHR systems and explains some of the general principles. The Working Document also gives indications on the data protection requirements for setting up EHR systems, as well as for the applicable safeguards. The processing of sensitive data is generally prohibited but is tolerated under specific circumstances [25]. Some of these circumstances are:

- if the data subject explicitly agrees on the processing of her sensitive data.
- if the processing of data is allowed by law.
- if the subject is unable to agree on the processing, e.g., due to unconsciousness.

Furthermore the Protection Directive (95/46/EC) defines the rights for the individual. Some of these rights are:

- to receive information about the processing of their own data,
- to receive a copy of all personal data held by the data controller,
- the prevention of direct marketing and automated decision-making,
- to seek damages for breach of the data protection principles.

In 2006 the United States Department of Health and Human Service Health issued the Health Insurance Portability and Accountability Act (HIPAA) which demands the protection of patients data that is shared from its original source of collection (cf. [16][17][18][19][28][29][30]). It is based on five principles:

1) Consumer control of medical information,
2) Boundaries that limit disclosure of medical treatment and
3) Payment accountability for violation of patient's rights with specific federal penalties,
4) Public responsibility for protecting public health, conducting medical research, improving quality of care and fighting health care fraud or abuse, and
5) Security of health information by organizations entrusted with that information.

The five principles only apply to individually identifiable health information, which is:

- created by or received from health care providers, employers or the clearinghouse.
- related to the provision of health care or the past, present or future medical condition.
- identifies or could reasonably be used to identify an individual.
- has been transmitted electronically or maintained in any other form or medium.

However, the act does not include other medical data, for example car insurance that has medical coverage or general sickness absence in the workplace that is not the subject of the health plan [24].

The disclosure of Protected Health Information (PHI) is permitted in certain cases. For example, the data is disclosed to the individual itself, the data is de-identified to carry out health plan's own treatment, payment or health care operations. Furthermore, the data owner could give consent to the processing of her medical data. To protect the privacy of individuals, many rights have been set up under the Health Insurance Portability and Accountability Act. Individuals have the right:

- to inspect or copy their own information,
- to request amendment or correction of erroneous or incomplete information,
- to request the restriction of use or disclosure,
- to give authorization for certain uses and disclosures.

III. Description of Pseudonymization Approaches

This chapter describes current pseudonymization approaches in detail. Thereby, we differentiate between three approaches. Firstly, there is the plain-text approach in which all data is readable for everyone. This approach could be compared with the traditional paper-record system. Secondly, there is the encrypted-text approach in which all data are encrypted and only accessible to persons with the key to decrypt this data. Thirdly, there is the pseudonymization approach, in which only the reference between the data and the data owner is encrypted. Table I gives an overview of the approaches:

A. Peterson Approach

Peterson [32] claims to provide a system for providing personal medical information records to an individual without jeopardizing privacy. The main ideas behind the approach are (i) the encryption of patient's data, (ii) the universal access to medical records by any (also unauthorized)

Description	*Name*	*References*
Plain-text approach	Approaches of Pommerening	[31]
Encrypted-text	Approach of Peterson	[32]
	Elektronische Gesundheits Karte	[33][35][36][34][37][38][39][40]
Pseudonymization approach	Pseudonymization of Information for	
Privacy in e-Health	[41][42][43][44][45]	
	Approach of Thielscher	[46]
	Approach of Slamanig and Stingl	[47][48][49]

Table I
OVERVIEW OF PSEUDONYMIZATION APPROACHES

person while (iii) the patient is responsible for granting privacy.

The user registers at the provider's website, receives a unique Global Key (GK) and server side key ($SSID$) generated by the provider and has to provide a unique Personal Encryption Key (PEK) as well as a password. The server returns a unique global key GK, which has to be different from the PEK. GK, PEK and password are stored in the Data Table. The user is demanded to enter a PEK until he provides a unique one. After registration the user may print the GK on an ID Card (on paper).

This approach consists of three database tables, the user table, the security table and the personal data table. The user table contains the GK, the PEK, a password and a foreign key to the security table. The security table contains a primary key, the method of encryption for the PEK, a server side encryption key and method and a foreign key to the personal medical data table. This table contains a primary key and the data, which is double encrypted with the PEK and the server side encryption key. Data is stored double encrypted in the database. If the user wants to retrieve data from the database, the user enters the GK or PEK, which are sent to the server through a firewall and checked if they match any entry in the database. The user enters an arbitrarily key and gets immediate access to the records without authentication.

In case of an emergency, the health care personnel can retrieve the medical data of the patient by entering the global key GK, or if the patient is responsive to verbal commands, she can tell them the private encryption key PEK. The system looks up the database for the entered GK or PEK and returns the decrypted medical data. The server looks up the $SSID$ and all corresponding data table row numbers needed for retrieving the (medical) data entries from the database. The records are decrypted using (i) the PEK and the personal encryption method and (ii) the server side encryption key $SSEK$ and the server side encryption method and delivered to the user. To modify or delete this medical data, the patient has to enter her password, which has been provided at registration time.

Table II shows the different access levels of this approach. If a person knows the global key GK or PEK or both, but does not have a password, she is able to view medical data sets. To be able to add, modify or delete medical datasets, the person has to provide the password. Peterson argues, that these access levels protect patient's privacy, because the data does not contain any identifying information. So, for an attacker, it would be of no interest to receive anonymous data.

Global Key	*Personal Key*	*Password*	*Resulting Action*
No	No	No	Access Denied
Yes	No	No	View Only
No	Yes	No	View Only
Yes	Yes	No	View Only
No	No	Yes	Access Denied
Yes	No	Yes	View and Edit
No	Yes	Yes	View and Edit
Yes	Yes	Yes	View and Edit

Table II
APPROACH OF PETERSON: ACCESS LEVELS [32]

B. Pseudonymization of Information for Privacy in e-Health (PIPE)

PIPE (cf. [50][42][51][52]) is a architecture that provides the following contributions compared to other methodologies: PIPE allows (i) the authorization of health care providers or relatives to access defined medical data on encryption level, (ii) provides a secure fall-back mechanism, in case the security token is lost or worn out, (iii) stores the data without the possibility of data profiling, and (iv) provides secondary use without establishing a link between the data and its owner.

The client is a service, which provides an interface to legacy applications, manages requests to local smart card readers and creates a secure connection to the server. The server, also called Logic (L), handles requests from clients to the storage. The data in the storage is divided into two parts, the personal data and the pseudonymized medical data. The link between personal data and pseudonymized medical data is protected through a hull-architecture. The hull-architecture (see Figure 1) contains a minimum of three

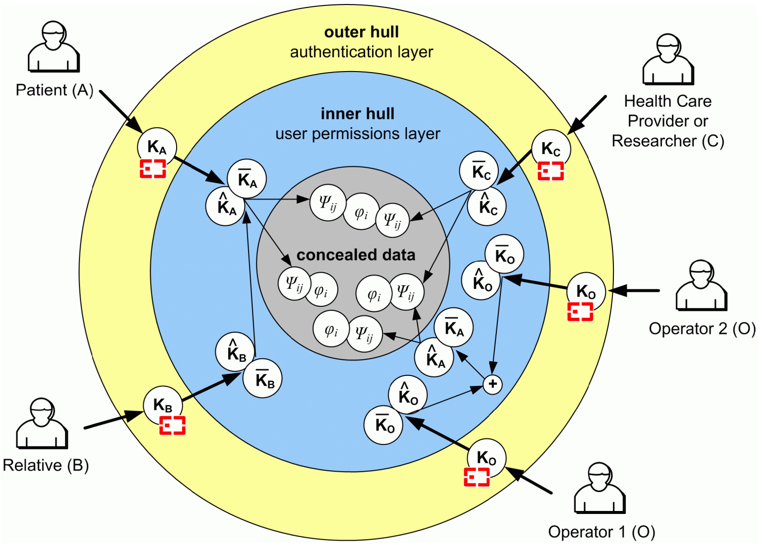


Figure 1. PIPE: Layered model representing the authorization mechanism

security-layers: the authentication layer (outer hull), the user permission layer (inner hull) and the concealed data layer. To reach the next hull, there are one or more secrets, for example, symmetric or asymmetric keys or hidden relations, in every hull-layer. A definition of all system attributes can be found in table III. PIPE defines users with different roles comprising patient A, relative B, health care provider C and operator O. The patient is the owner of her data and has full control of her datasets. She is able to view her medical data, add and revoke health care providers and she may define relatives, who have the same rights as herself. Health care providers can be authorized by the patient to see and create subsets of anamnesis data. The operators provide a backup in case the token needs to be replaced.

- The authentication layer contains an asymmetric key pair, e.g., the patient's outer public key K_A and outer private key K_A^{-1}. These keys are stored on a smart card and are protected with a pin code. The outer private key is used to decrypt the keys of the permission hull-layer.
- The permission layer contains an asymmetric key pair and a symmetric key, e.g., the patient's inner public key $\widehat{K}_A$, inner private key $\widehat{K}_A^{-1}$ and symmetric key $\overline{K}_A$. The symmetric key is encrypted with the inner private key and is used to en-/decrypt pseudonyms in the concealed data layer. If a patient associates a relative, her inner private key $\widehat{K}_A^{-1}$ is encrypted with the relative's inner public key $\widehat{K}_B$. So, the relative is able to decrypt the patient's symmetric key $\overline{K}_A$ with her inner private key $\widehat{K}_B^{-1}$, until the patient's inner private key $\widehat{K}_A^{-1}$ is changed.
- The concealed data layer contains hidden relations, which are called pseudonyms. Each medical data set is associated with one or more pseudonyms ψ_{i_j}. As the patient is the owner of her medical data and the person with security clearance, she owns the so called root-pseudonym ψ_{i_0}. These pseudonyms are calculated with an algorithm, which is based on a secret key. In our case, this secret key is the symmetric key of the user. Only instances, who are able to decrypt one of these pseudonyms ψ_{i_j}, can rebuild the link between the patient and her medical data.

To find the pseudonyms to rebuild the link to the medical data, the authors introduced keywords. Keywords are selected on creation time of the medical data or when another user is authorized. They are encrypted with the symmetric key of the root user and the user, who is being authorized. After the keywords are stored in the database, the user can select any of this keywords to find the pseudonym.

C. Electronic health card (eGK) architecture

The electronic health card architecture [33][34][35][36][37][38][39][40] is an approach of the Fraunhofer Institute supported by the Federal Ministry of Health Germany. The EGK is designed as a service-oriented architecture (SOA) with some restrictions: The health card can only be accessed locally on the client side. Services should use remote procedure calls for communication due to performance and availability issues. Therefore, the system architecture is divided into five layers:

- The *presentation* layer defines interfaces to communicate with the user,
- the *business logic* layer combines different services, which are processed automatically,
- the *service* layer provides special functional uncoupled services,
- the *application* layer realizes the user right and data management, and

	Patient	*Relative*	*HCP*	*Operator*	*Logic*
abbreviation	A	B	C	O	L
unique identifier	A_{id}	B_{id}	C_{id}	O_{id}	
(outer public key, private key)	(K_A, K_A^{-1})	(K_B, K_B^{-1})	(K_C, K_C^{-1})	(K_O, K_O^{-1})	(K_L, K_L^{-1})
(inner public key, private key)	$(\widehat{K}_A, \widehat{K}_A^{-1})$	$(\widehat{K}_B, \widehat{K}_B^{-1})$	$(\widehat{K}_C, \widehat{K}_C^{-1})$	$(\widehat{K}_O, \widehat{K}_O^{-1})$	
inner symmetric key	$\overline{K}_A$	$\overline{K}_B$	$\overline{K}_C$	$\overline{K}_O$	$\overline{K}_L$
key share				$\sigma_\iota(K)$	
medical data / anamnesis	φ_i				
pseudonym	ψ_{i_j}				

Table III
PIPE: Definition of System Attributes

- the *infrastructure* layer contains all physical hardware and software management, for example, data storage, system management, virtual private networks, etc.

With this layered architecture, the system provides several service applications such as emergency data, electronic prescription, electronic medical report or a electronic health record system. The system includes a ticketing concept to realize some uncoupled action in combination with security mechanisms, to comply with the privacy policy: All data, which will be stored in the virtual file system is encrypted with a one-time symmetric key, called session key. This session key is encrypted with the public key of the patient. To decrypt the data, the patient has to decrypt the session key with his private key and finally the data will be decrypted with this session key. A user is authenticated by using a Challenge-Response approach. Therefore the system generates a random number. This number will be encrypted with the public key of the user. Only the user is allowed to decrypt this random number with the private key, which is stored on her health card and can send it back to the eGK system. Furthermore, the ticketing concept manages the access rights to the system. A file or directory in this virtual file system has a default ticket-toolkit and any amount of private ticket-toolkits, called t-node (see Figure 2). The user defines a private ticket-toolkit for every other user in the system. This private ticket-toolkit could have stronger or looser access policies as the default ticket-toolkit. The ticket-toolkit contains a ticket-building tool, a ticket-verifier, the access policy list and a encrypted link to the directory or file. Every user holds a root directory in the virtual file system, which does not have a parent node. Furthermore, any directory contains unencrypted links to the ticket-toolkits of their child nodes. This technique enables the system to perform a fast selection of sub nodes (select * from t-nodes where parentID = directoryID).

To be able to find the root node of a specific user, the query service maps a unique identifier, for example the insurance number to the internal user and returns a ticket-toolkit containing a encrypted link to the root node. If there is no private ticket-toolkit available for the user, who performed the request, the system returns a default ticket-toolkit, which is based on a challenge. If the user is able to solve this challenge, she will get the access rights, which have been defined in the default access policy. Both, the hybrid encryption and the challenge response technique are based on the asymmetric key pair, which is stored on the patients' health card. Neither the operating company nor any public administration organization could recover the data, which has been stored in the system, if the patient lost the smart card or the card is worn out. To overcome this problem, the eGK architecture optionally provides the possibility to store a second private ticket-toolkit for every entry. This private ticket-toolkit uses an asymmetric key pair, which is stored on an emergency card. The architecture does not specify this emergency card, but recommends to use the card of a family member or a notary.

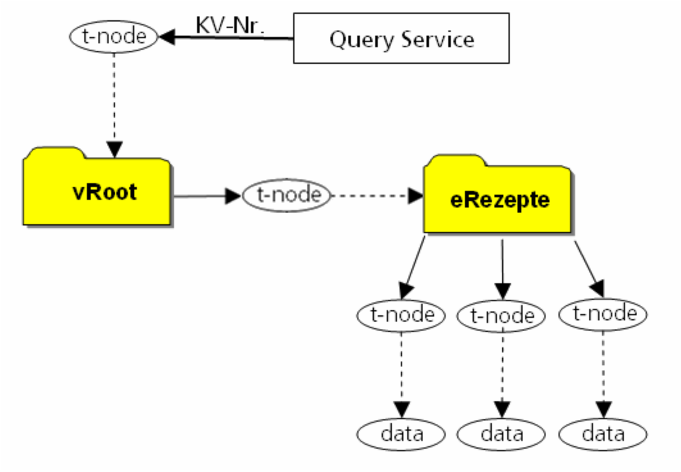


Figure 2. eGK: Virtual file system [33]

D. Thielscher Approach

Thielscher [46] proposes an electronic health record system, which uses decentralized keys stored on smart cards. The medical data are split into identification data and the anamnesis data and stored into two different databases. The key stored on the smart card of a patient is used to link the patient identity to her datasets. Therefore, this key generates a unique data identification code (DIC), which is also stored

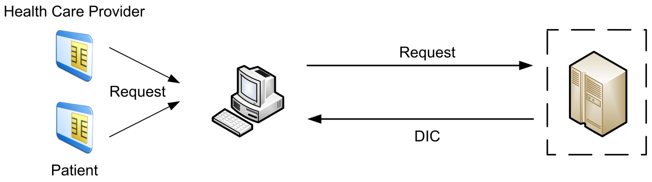


Figure 3. Thielscher: Architecture [46]

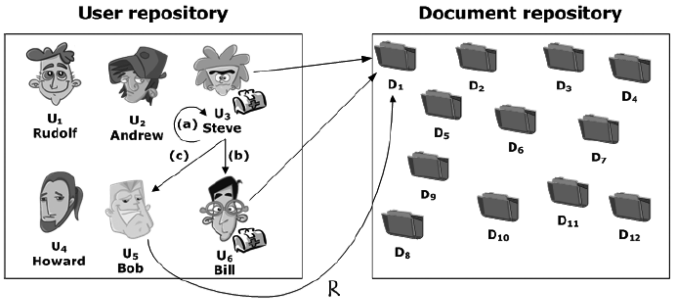
in the database. Such a DIC does not contain any information to identify an individual. Data identification codes are shared between the patient and health care providers to authorize them to access the medical data set. For more security the authorization is limited to a certain time period. After this period any access attempt is invalid. The system provides a mechanism in case of an emergency. Some parts of the patient's individual health data is stored directly on the smart card. A health professional has immediate access to this data in case of an emergency. Moreover, the system includes an emergency call center which is authorized to access the central database for requests and to read the data in case of an emergency. Therefore, the health professional has to confirm their identity to the call center.

E. Approach of Slamanig and Stingl

Stingl and Slamanig [48][49] propose a concept for an e-health portal with a public user repository and a document repository with encrypted medical documents. The link between these repositories is realized by a 5-tuple authorization concept $(U_S, U_R, U_C, U_P, D_i)$, which contains the identifiers of the sender U_S, the receiver U_R, the data creator U_C, the concerning user (i.e. patient) U_P, and the document reference with the decryption key D_i. All tuples except for the receiver are encrypted with the receiver's public key. Authorizations and the amount of disclosed information depend on the tuples used:

- $(U_1, U_1, U_1, U_2, D_1)$: User 1 creates this tuple concerning user 2 for accessing document 1.
- $(U_1, U_3, -, -, D_1)$: User 1 authorizes user 3 to access document 1 without disclosing information on the data creator and the concerning user.
- $(U_1, U_2, U_1, U_2, D_1)$: User 1 authorizes the concerning user 2 to access document 1 disclosing himself as the data creator.

In order to provide unlinkability, each user has a set of sub-identities, realized as independently chosen pseudonyms, with individual asymmetric keypairs. One of these sub-identities is defined as public identity used for authorizations, while the others are kept secret. Upon receipt of an authorization, the recipient first decrypts the relation with the private key of the public sub-identity, replaces the receiver tuple with one of his secret sub-identities, and then reencrypts the remaining tuples with the corresponding



...	
a)	(U_3,U_3,U_3,U_5,D_1)
b)	$(U_3,U_6,\_,\_,D_1)$
c)	(U_3,U_5,U_3,U_5,D_1)
...	

Figure 4. Slamanig and Stingl: Repositories and Shares [49]

public key such that the authorization tuple cannot be identified by any observer, except for the corresponding user. To prevent so called *disclosure attacks* (users forced to disclose their medical data, e.g., at job interviews) a special sub-identity can be chosen which includes only non-critical data. Highly sensitive data can be hidden in another sub-identity [54]. As fall-back mechanism, the authors mentioned that the distributed key backup to N users using a (t, N)-threshold secret sharing scheme could be implemented, because the users private keys are essential for the system.

The authors also propose the application of techniques such as anonymous authentication and obfuscation to further improve the patients' privacy. Obfuscation can be realized by intentionally producing collisions when selecting pseudonyms such that the pseudonyms are not unique, obfuscating the exact links between pseudonyms and documents. But obfuscation produces computational overhead because of invalid returned tuples (tuples actually not possessed by the user need to be identified as such by decrypting them and checking their semantic content). Anonymous authentication provides unlinkability between individual access operations but needs to be executed for each transaction individually.

In [55] and [56] they propose the application of their concept for personal health records (PHR). The medical documents are organized in virtual folders (where the content does not need not be disjunct) which in turn are controlled by sub-identities. In addition to identity pseudonymization, the folders and documents are pseudonymized as by foreign key encryption such that the documents, folders, and sub-identities cannot be linked by an observer [48].

Anonymous authentication and pseudonymization in the form of sub-identities provide a great deal of unobservability, both from the static and dynamic viewpoint. The usage of a special sub-identity managing only non-critical information also prevents exposure of sensitive data as a result of disclosure attacks. While reencryption of the authorization

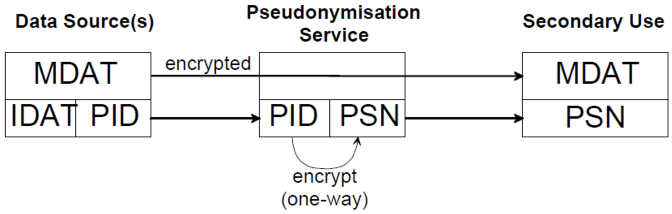


Figure 5. Pommerening: Data Flow for One-Time Secondary Use [31]

tuple after receipt ensures unobservability, it also prevents that the sender can revoke this authorization. In fact, the sender cannot control if the recipient authorizes a third person without the consent of the data owner. For finding a particular document, the user needs to select the correct sub-identity and folder and decrypt all document references (and document information) to determine the desired document; a query mechanism is not provided. Finally, because of the fully encrypted documents, secondary use is not possible without decryption by an authorized user.

F. Pommerening Approaches

Pommerening [31] proposes different approaches for secondary use of medical data. He differs between one-way and reversible pseudonyms. The first approach is based on data from overlapping sources for one-time secondary use. In this case, overlapping sources could be, e.g., data from different EHRs or biomaterial banks, which have been collected on another examination. To connect the data, a unique identifier (PID) is introduced. Figure 5 shows the pseudonymization workflow. A pseudonymization service encrypts the PID with a hash algorithm, and the medical data (MDAT) is encrypted with the public key of the secondary user. The secondary user can decrypt the medical data and merge the data of a person, but cannot identify it.

The second approach is also based on one-time secondary use, but with the possibility to re-identify the patient. Therefore, Pommerening extends the first approach with a PID service, which stores a reference list containing the identity of the patient (IDAT) and the associated PIDs. In case the patient should be notified, the pseudonymization service decrypts the pseudonym (PSN) and sends the request to the PID service, which notifies the data source owner.

The third approach fits the need of a research network with numerous secondary users. It supports long-term observation, e.g., of a patient with chronic diseases and allows to send research results to the patient or her responsible health care provider. The export and pseudonymization procedure is shown in figure 6. Therefore a physician exports her local database to the central researcher database. The identification data will be replaced with a PID in the PID service. For each secondary use the data will be exported through the pseudonymization service. The PID is encrypted by the pseudonymization service with a project specific key to ensure that different projects get different pseudonyms.

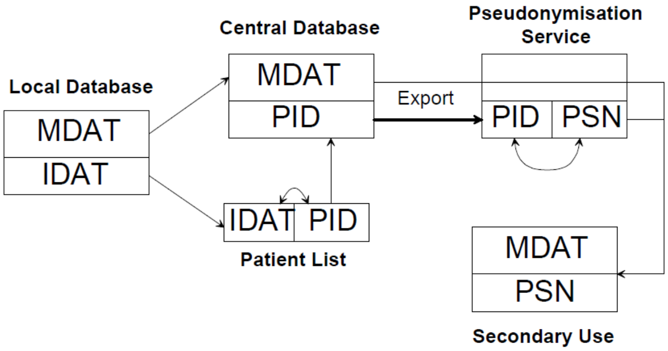


Figure 6. Pommerening: Data Flow for many Secondary Uses [31]

IV. Legal Evaluation

Pseudonymization approaches (e.g., used for securing electronic health record systems) have to adhere certain requirements to accord with privacy laws in the European Union or United States. The following set of requirements has been extracted from the Directive 95/46/EC of the European Parliament (DPA) and the Health Insurance Portability and Accountability Act (HIPAA) (cf. [23][24][53][17][18][19]).

- *User authentication*: The system has to provide adequate mechanisms for user authentication. This could be done, for example with smart cards or finger print.
- *Data ownership*: The owner of the medical data has to be the patient. The patient should be able to define who is authorized to access and create her medical records.
- *Limited access*: The system must ensure that medical data is only provided to authenticated and authorized persons.
- *Protection against unauthorized and authorized access*: The medical records of an individual have to be protected against unauthorized access. This includes system administrators who should not be able to access these medical records, for example, through compromising the database.
- *Notice about use of patients data*: The patient should be informed about any access to her medical records.
- *Access and copy own data*: The system has to provide mechanisms to access and copy the patients own data.
- *Unobservability*: Pseudonymized medical data should not be observable and linkable to a specific individual in the system.
- *Secondary use*: The system should provide a mechanism to export pseudonymized data for secondary use and a possibility to notify the owner of the exported data, if new medicaments or treatment methods are available.

Legal Requirements	*DPA*	*HIPAA*	*PIPE*	*eGK*	*Po*	*Pe*	*Th*	*St*
User authentication	x	x	x	x	-	o	x	x
Data ownership	x	x	x	x	-	-	x	o
Limited access	x	x	x	x	o	-	x	x
Protection against unauthorized and authorized access	x	x	x	x	o	-	o	x
Notice about use of patients data	x	x	x	x	-	-	-	-
Access and copy own data	x	x	x	x	o	x	x	x
Unobservability	x	x	x	x	x	-	x	x
Secondary use	-	x	x	o	x	-	-	-

Table IV
EVALUATION OF PSEUDONYMIZATION APPROACHES

Abbreviations		
Po	...	approach of Pommerening
Pe	...	approach of Peterson
Th	...	approach of Thielscher
Sl	...	approach of Slamanig and Stingl

Legend for DPA and HIPAA		
x	...	defined and accurate with the law
-	...	undefined in the law

Legend for pseudonymization approaches		
x	...	fully implemented
o	...	partially implemented
-	...	not implemented

Table V
LEGEND FOR TABLE IV

Table IV applies the legal criteria defined above to the selected pseudonymization approaches. Characteristics that are accurate with the law or fully implemented are denoted with x, whereas characteristics that are not accurate with the law or not implemented are denoted with $-$ and o indicates properties that are partially implemented.

Fulfilling legal requirements is an important precondition in order to guarantee security. However, since legal requirements are often defined in a generic way they leave room for interpretation. This results in a variety of approaches that are often vulnerable to typical attack scenarios. Table VI presents a list of typical attack scenarios and evaluates these criteria against the pseudonymization approaches described earlier.

- *Insider abuse*: Medical personnel may abuse their access rights for their own purposes. For example, they may want to know how family members or celebrities are being treated [57]. Insiders do not only abuse their privileges for their own purposes, they may release information to outsiders for spite, revenge or profit [57].
- *Social engineering*: is a common method to get information about a person. Therefore, an attacker could bribe or mislead an administrator of the pseudonymization system. For example, the attacker could fake her identity to get a new security token.
- *Data Disclosure*: Data mining attacks are a major threat for the disclosure of sensitive data as shown by Sweeney (cf. [58]). Sweeney was able to combine medical data with an electronic version of a city's voter list. The attacker can collect statistics and information about the data. In the worst case scenario the attacker could reconstruct the pseudonyms.
- *Attacker deletes data*: If an attacker breaks into the system, she may have the possibility to delete data. Therefore the system should be able to detect such changes and inform the system administrator about this attack and request a restoration of the datasets.
- *Attacker modifies data*: An attacker, who has broken into the system, may also change some datasets. Therefore, the system should digitally sign all records in order to detect modifications.
- *Attacker authorizes internal users*: An attacker could try to authorize an internal user or herself to be able to gain access to medical data of other users.
- *Attacker authorizes external users*: An attacker could try to authorize an external user or herself to be able to gain access to medical data of other users.
- *Administrator accesses data*: Administrators of the

Possible security issues	*PIPE*	*eGK*	*Po*	*Pe*	*Th*	*Sl*
Insider abuse	-	-	x	x	x	-
Social engineering	o	o	x	x	x	o
Data Disclosure	-	-	o	x	o	-
Attacker deletes data	o	x	x	x	x	o
Attacker modifies data	-	-	x	x	x	o
Attacker authorizes internal users	-	-	o	x	-	x
Attacker authorizes external users	-	-	o	x	-	-
Administrator accesses data	-	-	x	x	-	-
Administrator accesses cryptographic keys	-	-	o	x	o	-

Table VI
COMPARISON OF SECURITY ISSUES AND PSEUDONYMIZATION APPROACHES

Abbreviations		
Po	...	approach of Pommerening
Pe	...	approach of Peterson
Th	...	approach of Thielscher
Sl	...	approach of Slamanig and Stingl

Legend for pseudonymization approaches		
x	...	security issue
o	...	possible security issue
-	...	no security issue

Table VII
LEGEND FOR TABLE VI

pseudonymization system could access the database if the data is pseudonymized only by disclosure.

- *Administrator accesses cryptographic keys*: If system administrators have access to the private keys of individuals, she may have the possibility to decrypt all pseudonyms and link anamnesis to individuals. Every attacker who gets administration privileges could steal the database containing the keys.

Most of the approaches implement the requirements of *user authentication*, *data ownership*, *limited access* and serve control mechanisms *against unauthorized and authorized access*. The implementation of the requirement *protection against unauthorized and authorized access* is inadequate. Additional requirements, which enhance the security of the system and the containing datasets, are widely implemented. The approaches of Pommerening and Peterson only pseudonymize data on export. The approaches of Pommerening have the drawback that the generated pseudonyms from the PID service are stored in a reference patient list, to be able to re-build the link to the patient. To enhance the security, this list can be stored at a third party institution, but this measure does not prevent an abuse of the list through an insider of the third party institution. The system permits attackers to steal the database with all data linked to individuals. Moreover, system administrators could abuse their access privileges to release information to outsiders for revenge, profit or their own purposes [57]. An attacker could bribe an insider of the third party institution to get access to the patient list or the identifying data of some pseudonyms. The Peterson approach has some major security issues. Although the data is doubly encrypted an attacker getting access to the database gets access to all data stored on the server because the keys needed for decrypting the data are (i) also stored in the same database and (ii) what is even more important the relation between the tables (thus between the identification data and the medical data) are stored in clear text. An attacker getting access to the database can decrypt all information and, as the password is stored in the database as well as the keys, the attacker may change data stored in the database. The PEK is selected by the user but must be unique in the system. This behavior does not only open a security leak because the user trying to chose a key is informed about the keys that already existing in the system. An attacker could use the keys reported as existing for immediate access to the medical data associated with this key. Moreover, this behavior is impractical and inefficient in practice as the user might have to select dozens of keys before he enters a valid one. Peterson tried to prevent the following types of attacks: insider abuse, disclosure of weakly pseudonymized data and databases being stolen. He did so by defining that no identifiable data is allowed to be stored. However, the system is not able to check if identifiable words exist in the data. Thielscher's approach comes with the shortcoming, that the pseudonyms are stored centrally in the patient mapping list for recovery purposes. To prevent attacks to this list, Thielscher keeps this list off-line, but this mechanism cannot prevent insider abuse or social engineering attacks. The usage of a patients-pseudonyms list as fall-back mechanism could lead to security issues. The work-around of Thielscher to keep the patients-pseudonyms list off-line promises a higher level of security, but does not prevent the system against social-engineering or insider attacks. Furthermore, it does not provide protection if the attacker gets physical access to the computer. Another drawback of the system is the emergency call center. This call center can abuse their access privileges to get access to medical data of any patient. The drawback of the approach of Slamanig and Stingl is that an attacker (a person who gets access privileges on a document) may authorize other users, send faked medical documents or disclose medical data. For example, the requirements to send a faked medical document are, (i) access to the database, (ii) the public pseudonym U_P of the user, which the attacker wants to harm, (iii) any public pseudonym to fake the sender U_S and creator U_C, (iv) the public pseudonym and the public key K_R of the receiver U_R, for example the employer, and (v) a harmful document D_i. After the attacker has all the required information, she inserts a new tuple into the authorization table. After the next login of the receiver, the system replaces the public pseudonym of the user with a private pseudonym of the receiver. The authors suggest obfuscation to handle this problem. The approach does not prevent tuple reordering and, thus, allows the attacker to modify data.

PIPE, eGK and Slamanig/Stingl store the data pseudonymized in the database. Attackers who get access to the database or system administrators cannot link the data to individuals. All those approaches provide a high level of security. Even if the attacker breaks into the database, she would not be able to link and read the stored data. Maybe, the attacker could do a data profiling attack and get some informations from the unencrypted keywords, if these contain any identifiable words. The only way to link the data to an individual is by doing a social engineering attack and fake the identity of the person, the attacker wants to attack. Therefore, the attacker would have to fake a official photo identification in order to get a new smart card to access the system. Another method to link data to an individual is by doing a data mining or data profiling attack.

V. Conclusion

Health care require the sharing of patient related data in order to provide efficient patients' treatment. As highly sensitive and personal information is stored and shared within highly interconnected systems (e.g., electronic health records), there is increasing political, legal and social pressure to guarantee patients' privacy. Although, legislation demands the protection of patients' privacy, most approaches that lay claim to protect patients' privacy fail in fulfilling legal requirements.

This paper gave an overview of research directions that are currently pursued for privacy protection in e-health and evaluated common pseudonymization approaches against legal criteria taken from legal acts and literature. Thereby, this paper answered the questions (i) which pseudonymization approaches adhere to the current privacy laws and (ii) what are the major drawbacks of pseudonymization approaches. In order to answer the first research question, seven legal requirements have been extracted from relevant legal acts. These requirements could be used for the future development of pseudonymization approaches. At the moment, only two out of the six evaluated pseudonymization approaches fulfill the legal requirements. Therefore, only two out of the six approaches can actually be considered for use in the European Union and United States. Moreover, the results of the evaluation show that newer approaches already consider legal demands and fulfill more legal requirements of the European Union and the United States. An additional security evaluation, carried out to answer the second research question, shows that there are major drawbacks in most of the systems. Some approaches use a pseudonym-patient mapping list, which could very easily be abused by an insider of the system, for example a system administrator. A more secure way was presented by eGK, where all data is linked to backup security tokens. However, if both security tokens are accidentally destroyed, for example by fire, all data would be lost forever. Only two approaches suggest a solution to share the keys of the security token in the system using a threshold scheme. PIPE is the only approach which implements such a fall-back mechanism.

From the six candidates that were evaluated, only two can be seriously considered for use in practice. The result show that more contemporary approaches fulfill more of the legal requirements of the European Union and the United States. Whereas the eGK approach encrypts patients' data, PIPE leaves the decision of encrypting patients' data up to the user. Therefore, PIPE turns out to be the more appropriate option if secondary use is demanded. Apart from this difference both approaches - eGk and PIPE - provide a similar level of security and fulfill the majority of the applied criteria. The results of the evaluation can support decision makers (such as chief security officers) especially in health care in their decision process when it comes to the selection of a system for protecting patients' data according to legal requirements posed by HIPAA or the EU Directives. Furthermore, the results may assist researchers in identifying the gaps of current approaches for privacy protection as a basis for further research.

VI. Acknowledgments

This work was supported by grants of the Austrian Government's FIT-IT Research Initiative on Trust in IT Systems under the contract 816158 and was performed at Secure Business Austria, a competence center that is funded by the Federal Ministry of Economy, Family and Youth of the Republic of Austria as well as by the provincial government of Vienna.

References

[1] Thomas Neubauer and Mathias Kolb. Technologies for the pseudonymization of medical data: A legal evaluation. In *Proceedings of the IEEE International Conference on Systems (ICONS)*, 2009.

[2] S. Märkle, K. Köchy, R. Tschirley, and H. U. Lemke. The PREPaRe system – Patient Oriented Access to the Personal Electronic Medical Record. In *Proceedings of the 17th International Congress and Exhibition on Computer Assisted Radiology and Surgery*, number 1256 in International Congress Series, pages 849–854, 2001.

[3] Frank R. Ernst and Amy J. Grizzle. Drug-related morbidity and mortality: Updating the cost-of-illness model. Technical report, University of Arizona, 2001.

[4] Common criteria for information technology security evaluation, ISO/IEC 15408:1999.

[5] C. H. Weiss. *Evaluation: Methods for studying programs and policies*. Prentice Hall, 2nd edition, 1998.

[6] E. R. House. Assumptions underlying evaluation models. *Educational Researcher*, 7(3):4–12, 1978.

[7] D. L. Stufflebeam and W. J. Webster. An analysis of alternative approaches to evaluation. *Educational Evaluation and Policy Analysis*, 2(3):5–19, 1980.

[8] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, 1998.

[9] Bruce Schneier. Risks of data reuse. Schneier on Security - Blog, June 2007. Last access 28.09.2009.

[10] Bruce Schneier. Our data, ourselves. Schneier on Security - Blog, May 2008. Last access 28.09.2009.

[11] Alfred Kobsa. Personalized hypermedia and international privacy. *Commun. ACM*, 45(5):64–67, 2002.

[12] Organisation for Economic Cooperation and Development (OECD). Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data c(80)58/final, 1980.

[13] Solveig Singleton. Privacy and Human Rights: Comparing the United States to Europe. In *The Future of Financial Privacy*, pages 186–201, 1999.

[14] William Seltzer. Population Statistics, the Holocaust, and the Nuremberg Trials. *Population and Development Review*, 24(3):511–552, 1998.

[15] Colin John Bennett. *Regulating Privacy: Data Protection and Public Policy in Europe and the United States*. Cornell University Press, 1992. ISBN: 0801480108.

[16] United States Department of Health & Human Service. HIPAA Administrative Simplification: Enforcement; Final Rule. *Federal Register / Rules and Regulations*, 71(32), 2006.

[17] U.S. Department of Health & Human Services Office for Civil Rights. Summary of the HIPAA Privacy Rule, 2003.

[18] U.S. Department of Health & Human Services Office for Civil Rights. Your Health Information Privacy Rights.

[19] U.S. Congress. Health Insurance Portability and Accountability Act of 1996. *104th Congress*, 1996.

[20] Federal Trade Commission. Children's online privacy protection act. United States federal law, 15 U.S.C. §6501-6506, October 1998.

[21] U.S. House of Representatives. U.S. Code - Title 47 - Telegraphs, Telephones, and Radiotelegraphs - Chapter 5 - § 551.

[22] H. R. 84. Online privacy protection act of 2005. 109th Congress, 1st Session, Bill, October 2005. This bill never became law.

[23] European Union. Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities*, L 281:31–50, 1995.

[24] Stephen Hinde. Privacy legislation: A comparison of the US and European approaches. *Computers and Security*, 22(5):378–387, 2003.

[25] *Data protection in the European Union - Citizen Guide*. European Union, 2001.

[26] Gerhard Steinke. Data privacy approaches from US and EU perspectives. *Telematics and Informatics*, 19(2):193–200, 2002.

[27] European Union, Article 29 Working Party. Working document on the processing of personal data relating to health in electronic health records, February 2007.

[28] Tim Churches. A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BMC Medical Research Methodology*, 3(1), 2003.

[29] George J. Annas. Hipaa regulations - a new era of medical-record privacy? *The new england journal of medicine*, 348(15):1488–1490, 2003.

[30] David Baumer, Julia Brande Earp, and Fay Cobb Payton. Privacy of medical records: IT implications of HIPAA. *ACM SIGCAS Computers and Society*, 30(4):40–47, 2000.

[31] Klaus Pommerening and Michael Reng. *Medical And Care Compunetics 1*, chapter Secondary use of the Electronic Health Record via pseudonymisation, pages 441–446. IOS Press, 2004.

[32] Robert L. Peterson. Patent: Encryption system for allowing immediate universal access to medical records while maintaining complete patient control over privacy. *US Patent US 2003/0074564 A1*, 2003.

[33] Fraunhofer Institut. Spezifikation der Lösungsarchitektur zur Umsetzung der Anwendungen der elektronischen Gesundheitskarte, March 2005.

[34] Jörg Caumanns, Herbert Weber, Arne Fellien, Holger Kurrek, Oliver Böhm, Jan Neuhaus, Jörg Kunsmann, and Bruno Struif. Die eGK-Lösungsarchitektur Architektur zur Unterstützung der Anwendungen der elektronischen Gesundheitskarte. *Informatik-Spektrum*, 29(5):341–348, 2006.

[35] Jörg Caumanns. Der Patient bleibt Herr seiner Daten: Realisierung des eGK-Berechtigungskonzepts über ein ticketbasiertes, virtuelles Dateisystem. *Informatik-Spektrum*, 29(5):323–331, 2006.

[36] Andreas Rottmann. CAMS dirigiert die eGK. *Datenschutz und Datensicherheit - DuD*, 30:153–154, 2006.

[37] Jan Neuhaus, Wolfgang Deiters, and Markus Wiedel. Mehrwertdienste im Umfeld der elektronischen Gesundheitskarte. *Informatik-Spektrum*, 29(5):332–340, 2006.

[38] Bernd Blobel and Peter Pharow. Wege zur elektronischen Patientenakte. *Datenschutz und Datensicherheit - DuD*, 30(3):164–169, 2006.

[39] Gerd Bauer. Aktive Patiententerminals. *Datenschutz und Datensicherheit - DuD*, 30(3):138–141, 2006.

[40] D. Wilhelm, A. Schneider, and C. F. J. Götz. Die neue Gesundheitskarte. *Der Onkologe*, 11(11):1157–1165, 2005.

[41] Bernhard Riedl, Thomas Neubauer, and Oswald Boehm. Patent: Datenverarbeitungssystem zur Verarbeitung von Objektdaten. *Austrian-Provisional-Application, Application No. A 1928/2006*, 2006.

[42] Bernhard Riedl, Thomas Neubauer, Gernot Goluch, Oswald Boehm, Gert Reinauer, and Alexander Krumboeck. A secure architecture for the pseudonymization of medical data. In *Proceedings of the Second International Conference on Availability, Reliability and Security*, pages 318–324, 2007.

[43] Bernhard Riedl, Veronika Grascher, and Thomas Neubauer. Applying a threshold scheme to the pseudonymization of health data. In *Proceedings of the 13th IEEE Pacific Rim International Symposium on Dependable Computing (PRDC'07)*, pages 397–400, 2007.

[44] Bernhard Riedl, Veronika Grascher, Stefan Fenz, and Thomas Neubauer. Pseudonymization for improving the privacy in e-health applications. In *Proceedings of the Forty-First Hawai'i International Conference on System Sciences*, page 255, 2008.

[45] Bernhard Riedl, Veronika Grascher, Mathias Kolb, and Thomas Neubauer. Economic and Security Aspects of the Appliance of a Threshold Scheme in e-Health. In *Proceedings of the Third International Conference on Availability, Reliability and Security*, pages 39–46, 2008.

[46] Christian Thielscher, Martin Gottfried, Simon Umbreit, Frank Boegner, Jochen Haack, and Nikolai Schroeders. Patent: Data processing system for patient data. *Int. Patent, WO 03/034294 A2*, 2005.

[47] Christian Stingl, Daniel Slamanig, Dominik Rauner-Reithmayer, and Harald Fischer. Realisierung eines sicheren zentralen Datenrepositories. In *Tagungsband DACH Security*, 2006.

[48] Christian Stingl and Daniel Slamanig. Berechtigungskonzept für ein e-Health-Portal. In Günter Schreier, Dieter Hayn, and Elske Ammenwerth, editors, *eHealth 2007 - Medical Informatics meets eHealth*, number 227, pages 135–140. Oesterreichische Computer Gesellschaft, 2007.

[49] Daniel Slamanig and Christian Stingl. Privacy aspects of ehealth. In *Proceedings of the Third International Conference on Availability, Reliability and Security*, pages 1226–1233, 2008.

[50] Bernhard Riedl, Thomas Neubauer, and Oswald Boehm. Patent: Datenverarbeitungssystem zur Verarbeitung von Objektdaten. *Austrian Patent, Nr. 503291, September*, 2007.

[51] Bernhard Riedl, Veronika Grascher, and Thomas Neubauer. A secure e-health architecture based on the appliance of pseudonymization. *Journal of Software*, 3:23–32, 2008.

[52] Thomas Neubauer and Bernhard Riedl. Improving patients privacy with pseudonymization. In *Proceedings of the International Congress of the European Federation for Medical Informatics*, number 136 in Studies in Health Technology and Informatic, pages 691–696, 2008.

[53] Gerrit Hornung, Christoph F.-J. Götz, and Andreas J. W. Goldschmidt. Die künftige Telematik-Rahmenarchitektur im Gesundheitswesen. *Wirtschaftsinformatik*, 47:171–179, 2005.

[54] Daniel Slamanig and Christian Stingl. How to preserve patient's privacy and anonymity in web-based electronic health records. In *Proceedings of the 2nd Internatinal Conference on Health Informatics (HEALTHINF 2009)*, 2009.

[55] Daniel Slamanig and Christian Stingl. Sophisticated methods to prevent insider attacks against PHR systems. In *Proceedings of the IADIS International Conference on e-Health*, 2009.

[56] Daniel Slamanig and Christian Stingl. Ein sicheres patientenzentriertes konzept für personal health records. In *eHealth 2009 - Health Informatics meets eHealth*, 2009.

[57] Thomas C. Rindfleisch. Privacy, information technology, and health care. *Commun. ACM*, 40(8):92–100, 1997.

[58] Latanya Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

Business Intelligence Based Malware Log Data Analysis as an Instrument for Security Information and Event Management

Tobias Hoppe
Chair of Business Informatics
Ruhr-University of Bochum
Bochum, Germany
thoppe@winf.rub.de

Alexander Pastwa
Steria Mummert Consulting
Dusseldorf, Germany
alexander.pastwa@steria-mummert.de

Sebastian Sowa
Institute for E-Business Security
Ruhr-University of Bochum
Bochum, Germany
sebastian.sowa@rub.de

***Abstract*—Enterprises face various risks when trying to achieve their primary goals. In regard to the information infrastructure of an enterprise, this leads to the necessity to implement an integrated set of measures which should protect the information and information technological assets effectively and efficiently. Furthermore, tools are needed for assessing risks and the performances of measures in order to guarantee continuous effort to protect the enterprises' assets. These tools have to be able to support the handling of the vast amount of security relevant data generated within the enterprise information infrastructure and their analysis. Both tasks are typical for security information and event management. In this context, the current paper introduces an approach for malware log data analysis by using business intelligence methods. Thereby, examples are given which are derived from the results of a project being conducted with a world-wide operating enterprise.**

Business Intelligence; Data Mining; Malware; Online Analytical Processing; Security Information and Event Management

I. INTRODUCTION

In general business management research as well as in the field of business informatics, it is a well known fact that the effective as well as efficient processing of information constitutes one of the most important drivers for the success of an enterprise [DBKDA 2009, 1; 2]. For this purpose, adequate information systems are used. The organization's functions and processes highly depend on information and on those information systems, which semi- or fully automatically support information processing [3].

Considering that already a temporary unavailability of essential information systems may lead to existential dangers, special attention must be paid to measures which ensure that all devices and applications of the information infrastructure being necessary for the information processing activities are used. Furthermore, breaches in the confidentiality, integrity, and the non-repudiability in regard to information assets or information processing technologies may constitute perceptible impairments or even existential crises [4]. The protection of these security objectives therefore is one of the central goals of information management, which generally aims to support the executives with an optimally designed and run information infrastructure. Tasks and responsibilities focusing on the achievement of the aforementioned security objectives are attributed to the subdivision respectively -function of information security management.

An integrated bundle of measures (containing organizational, technical, logical as well as physical measures) is needed for the realization of the defined security objectives [5; 6]. Here, information security management includes the steering and controlling of measures as well as their initial planning. This process must be seen as a continuous operation to guarantee a sustainable realization of the desired level of protection [7; 8]. In this context, information again incorporates a very important role – it forms the basis for any possible modification of the measures aiming to hold or improve the level of protection which is defined by the executives on the basis of an analysis of threats and economic impacts.

As subdivision or sub-function of the information security management of an enterprise, the security information and event management (SIEM) discussed in this paper typically uses a wide range of information from various elements of the information security architecture. The information security architecture is defined as the part of the information infrastructure which contains all components to enforce the defined information security objectives. Further more, these components can be used for the management and re-engineering of the relevant security concepts. From this background, the architectural elements compromise all access controls, operating system cores, firewalls and further measures to guarantee safe communication, for instance [9].

As comprehensive as the amount of elements of the information security architecture is, as comprehensive is the amount of data generated from its elements. As consequence, the task of data evaluation is complex and time consuming. Therefore, a critical success factor for executives of SIEM has to be seen in the quality and not the quantity of data relevant for the decisions about the conceivable modifications of security measures.

Due to the amount and complexity of data that have to be analyzed, questions about adequate tools, methods and models to support the analysis process arise. Here, one of the most successful applied approaches in the business management context is business intelligence (BI). This paper shows how BI can be used to answer two questions which are relevant for SIEM: 1. How do malware causing attributes relate to each other? 2. How does malware spread in the IT landscape and how long does it reside in the system? For these purposes, known malware which occurred within a certain timeframe will be analyzed.

After dealing with the theoretical backgrounds concerning SIEM in Chapter II, Chapter III introduces the concept of business intelligence. Chapter IV shows how Online Analytical Processing (OLAP) can be applied for SIEM. Chapter V then focuses the research objectives of this paper from the perspective of data mining whereas Chapter VI refers to its results. Chapter VII gives a brief conclusion and finally, Chapter VIII exemplifies future work.

II. Theoretical Background – SIEM

Before presenting how BI, in particular OLAP and data mining, may support the goals of SIEM, the following paragraphs characterize specific problems of data analysis as well as the requirements for designing a BI system. In the first step, terms and definitions which are relevant for the overall conceptual coherences are introduced.

A. *Relevant Terms and Definitions*

Information as the first relevant term used in the discussion of information security management topics can linguistically be derived from the Latin informatio. In this turn, informatio stands for the explanation or interpretation of ideas as well as it can be used in the meaning of education, training or instruction. This gives a first consideration about an accurate and precise definition: Information in this paper is defined as an explanatory, significant assertion that is part of the overall knowledge as well as it is seen as specific, from human beings interpreted technical or non-technical processed data [10; 11].

The just given definition of information is precisely in line with the ISO/IEC standards which explain that information "can exist in many forms. It can be printed or written on paper, stored electronically, transmitted by post or by using electronic means, shown on films, or spoken in conversation" [7; 8]. This – mostly trivial – way to use the term information unfortunately does not reflect the common sense in the information security community. There, it is quite often assumed to only affect electronic data, and thereby information security management has mostly to deal with IT. Although this paper focuses on data gathered from technological elements, it is stressed that this only covers one aspect of the entire tasks of information security management executives.

As consequence of the appreciation of information, also information security has to cover technical as well as non-technical challenges. In this context, the ISO explains that whatever "form the information takes, or means by which it is shared or stored, it should always be appropriately protected. Information security is the protection of information from a wide range of threats in order to ensure business continuity, minimize business risk, and maximize return on investments and business opportunities" [7; 8].

The term SIEM combines security information management and security event management. In both areas, the focus lies on the collection and analysis of security relevant data in information infrastructures respectively the security infrastructures. Thereby, the security event management emphasizes the aggregation of data into a manageable amount of information in order to deal with events and incidents immediately (for example, in a timely fashion).

In contrast to security event management, security information management primarily focuses on the analysis of historical data aiming to improve the long term effectiveness/efficiency of the information security infrastructure [12].

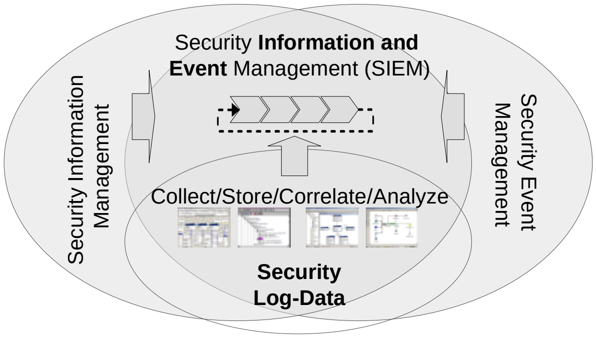

Security **Information and Event** Management (SIEM)
Security Information Management
Security Event Management
Collect/Store/Correlate/Analyze
Security Log-Data

Figure 1. Conceptual Architecture of SIEM

As shown in Figure 1, SIEM then stands for the amalgamation of security information management and security event management into an integrated process of planning, steering, and controlling security relevant information on the basis of the data collected from the information security architecture. Carr states: "Security information and event management (SIEM) systems help to gather, store, correlate and analyze security log data from many different information systems" [13].

B. *Selected Challenges SIEM is facing*

The analysis of security relevant data collected from the information security architecture is a challenging task because of the following reasons:

- Amount of data
- Heterogeneity of data formats
- Heterogeneity of the data contents
- Limited personnel and budget

As consequence of the various information security architecture elements and the number of protocols, the amount of data gathered is massive. Thus, considerable manual effort is needed to gather relevant information about security threats. Furthermore, the data collected exist in various formats, making evaluation difficult and time-consuming. The heterogeneity of the data contents also impedes a simple and flexible analysis. Depending on the system and the action performed, the data may contain information about incidents or threats due to email or internet use, for example. In addition, data may be recorded, since specific ports are used by gateways and firewalls, for instance. Therefore, the possibility of manually analyzing data which are derived from the information security architecture elements is severely limited due to the sheer volume of data as well as the heterogeneity of data formats and contents.

Two further aspects must be considered. Typically, information security management divisions have only a small fraction of personnel, and the budget is also limited. As well as in other entities of an enterprise, the resources also spent for SIEM have to be managed economically. Thus, SIEM faces the same requirements as the other organizational units of the entire enterprise. The executives have to allocate resources in such a way that the specific entity contributes to the enterprise's goals as much as possible [14]. To sum up, the following aspects are identified as the primary requirements for SIEM:

- Extraction of information and knowledge
- Establishment of an integrated and continuous management process
- Effective and efficient data evaluation
- Support for network management
- Support for compliance management

By identifying relevant information and deducting knowledge from the existing volume of data, SIEM strives to guarantee the protection of information and information system values. To achieve this goal, it is necessary to conduct SIEM as an integrated, continuous management process. In turn, this process is dependent on the information relevant to the decision makers. This information again is extracted from the data pool. From the background of the limitations of data evaluation as described above, it is crucial to establish appropriate (what means highly effective and efficient) practices and mechanisms to support the data processing for the needs of the SIEM executives.

As consequence of the numerous elements installed in the enterprise information security architectures, the number of protocols as well as the amount of data generated is enormous. Depending on the system and the action performed, log data may contain information about incidences or threats (due to email or internet use, for example). In addition, the data relevant for security information and event management (SIEM) may be recorded because specific ports were used by gateways and firewalls, for instance [15].

III. The Concept of Business Intelligence for Supporting SIEM

After describing the challenges of SIEM, the current chapter focuses on the introduction of the concept of business intelligence (BI).

Business intelligence stands for a conceptual framework which bundles numerous approaches, tools and applications used for the analysis of business relevant data [16]. The general aim of BI is to support effective and efficient business decision making for what purpose a data warehouse is built up. Usually a data warehouse serves as the central storage system of a BI system. For implementing a BI application serving the goals of SIEM, a reference architecture has to be defined initially. Here, Figure 2 shows the layers and elements of an architecture that serves as a basic guiding topology in this context.

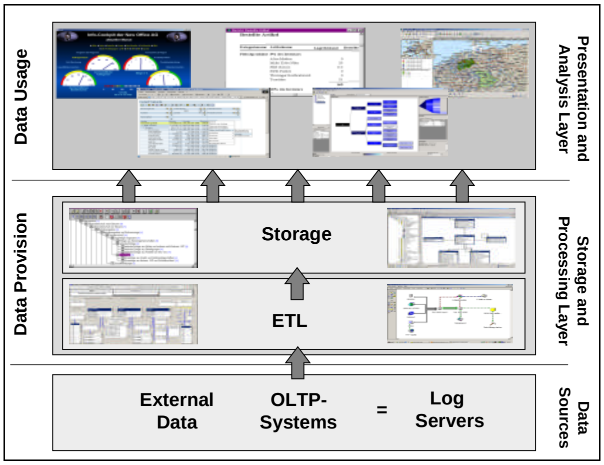


Figure 2. BI Reference Architecture [17]

A. Data Sources

At the lowest level of the BI reference architecture, various enterprises' operational systems as well as useful external data sources are located. They serve as data suppliers for the data warehouse as the integral part of the middle layer. The data primarily relevant for SIEM is gathered from the information security architecture elements which log information security relevant processes and incidents. This includes data about installed operating systems, versions of patches, installed anti-malware programs or information about the frequency of user password changes, for instance.

Potential threats can be identified by logging policy violations, malware reports, login-/logout-events and account-lockouts of users. This data is transported to log servers providing the input data for the data warehouse.

B. *Storage and Processing Layer*

One goal of a BI application is the consolidation of different data contents and formats towards a uniform perspective. For this task, an ETL (extraction, transformation, and load) component is combined with solutions for storing and preparing the data for later presentation / analysis [17]. This component constitutes a further module of the BI architecture and serves as the interface between the operational systems and the data warehouse [18]. It transfers the heterogeneous data into a consistent and multidimensional data perspective and loads the data into the data warehouse; in detail:

Extraction: Extraction deals with the selection and deployment of source data. Since relevant data typically exist in a very heterogeneous form, the ETL tool needs to access all data from the operational systems containing the security relevant log data.

Transformation: Transforming the source data into the target formats of the data warehouse is the central task of the ETL process. It can be further divided into the steps of filtering, harmonization, aggregation and enrichment. Filtering ensures that only the data necessary for the multidimensional analysis is loaded into the data warehouse. Log files usually contain lots of information not needed for analysis. For example, Windows event logs record a multitude of application and system information. But for the purposes of SIEM, only information security events are needed. Following, harmonization corrects the data of syntactical and semantic defects. Also an adjustment of codes, synonyms and homonyms as well as the unification of different definitions of terms will be conducted. For example, for the same person, a different user name could have been assigned in a Windows environment and in a UNIX or a Linux environment. In the multidimensional database, this user must be clearly identifiable, however. In a further step of transformation, the consistent, but in the lowest level of granularity existing data will be aggregated to improve analysis performance. Here, the aggregation of hosts to organizational units or geographical locations is a possibility. Enhancing the data by adding contextual information represents the last and very important step of the transformation process because the knowledge generated in the consequence enables to systematically substantiate decision making processes on a broader base.

Load: Finally, the extracted and transformed data is loaded into the data warehouse where it is permanently stored. For this purpose, batches are used. In order to ensure the adequate supply of information in regard to timeliness and quality, the question has to be answered how long the interval between the single batches should be. Thus, depending on the amount of data as well as on the information and communication technologies in use and the information needed by the decision makers, the data is transferred flexibly from the source systems into the data warehouse.

C. *Presentation and Analysis Layer*

The top layer of the BI reference architecture comprises all methods and tools which are capable to analyze the multidimensional data as well as to present analytical reports. Among the different possibilities in this context, OLAP and data mining methods play an especially prominent role:

Online Analytical Processing: OLAP is a software technology. It allows decision makers to accomplish fast, interactive and flexible requests to the relevant data stored in a multidimensional structure [19].

Data Mining: While OLAP focuses mainly on historical analysis, data mining is concerned with a prospective analysis. By applying various statistical and mathematical methods, data miners aim to identify so far unknown data patterns [20].

OLAP and data mining increase the prospect of analyzing security relevant data efficiently for the short term treatment (e.g., of malware threats) as well as for the long term improvement of the overall information security architecture. Especially in regard to the SIEM challenges, BI offers the chance to handle the accrued amount of data and to transfer the heterogeneous data into a consistent format that can be used for analyses and reports of SIEM relevant topics.

IV. Applying Online Analytical Processing for SIEM

Up to now, challenges of SIEM and characteristics of BI have been described. The following chapters focus on the combination of these fields, presenting an OLAP application for SIEM.

A. *Multidimensional Data Model*

Modeling an adequate multidimensional data structure is one of the crucial factors of success when designing a BI application. It forms the basis for the execution of the ETL process with which relevant data is loaded from the operational systems into the data warehouse. The resulting data construct can then be analyzed by typical OLAP operations: Slice, dice, drill down, and roll up. By using these operations, diverse occurrences of different perspectives can be determined and evaluated, like the frequency of malware infections within a certain period on a certain operating system, for instance. Figure 3 visualizes the arrangement of the dimensions mentioned above in the structure of a so called data cube [21].

Multidimensional data models consist of a fact table and further tables which serve to depict the so called dimensions. Dimensions stand for the relevant entities with which the metrics of the fact table can be analyzed [22]. Hence, dimensions are used to provide additional perspectives to a given fact [23]. In order to ensure the data quality, it is of vital importance to follow a systematic and holistic approach when defining the dimensions and selecting the facts.

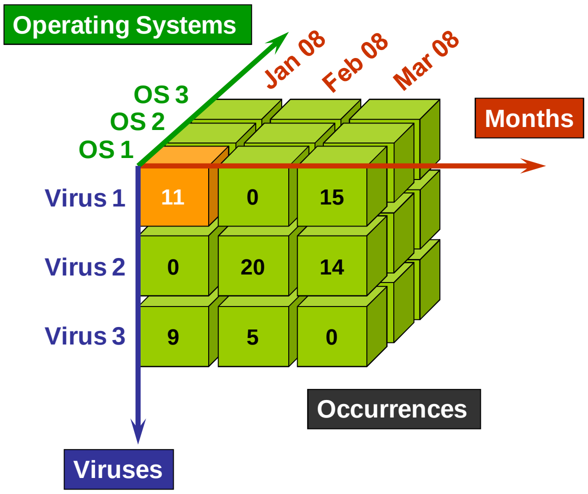


Figure 3. Example of a data cube for OLAP analyses

The content proposed in this paper refers to the key findings resulting from a cooperative project between a university and an industrial institution of leading presence. The goal was to develop a solution for a more sophisticated analysis of information security relevant data. The industrial institution uses a combination of several security systems. The generated log data is stored in a centralized relational database. Amongst others, main sources of the log data of interest are those from anti-malware solutions.

Figure 4 illustrates the business objectives of the business intelligence project and the way log data contributes to them.

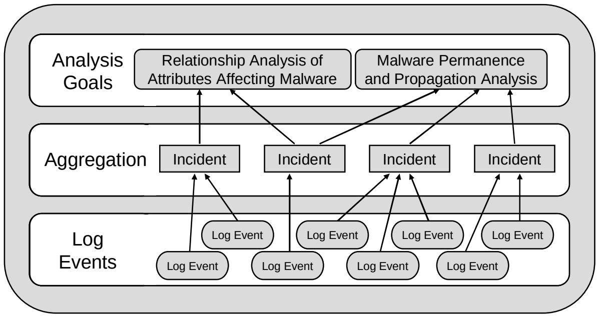


Figure 4. Aggregation of Log Events

A log event thereby is a specific, single event created by some log source and stored in the database. An example for a log event is the finding of a malware program. Log events are very numerous and hard to analyze, so they are aggregated to incidents. An incident thus covers one or more log events which belong together.

In order to aggregate malware logs, two cases must be considered:

(1) A malware which is detected at t_1 reappears on the same computer at t_2 and thus generates a new log file. For this case, the reappearance of the malware at t_2 is treated then as a new incident, if the malware has been deleted successfully in t_1 and the subsequent scan has not revealed a persistence of the malware. In addition, the malware events must have occurred on the same computer and must be caused by the same user.

(2) Further on, each log event indicating that a new malware has been detected on a computer becomes part of a new malware incident.

Figure 5 gives an overview of the input log data made available for the case study. Only known malware was in the focus of the upcoming analysis.

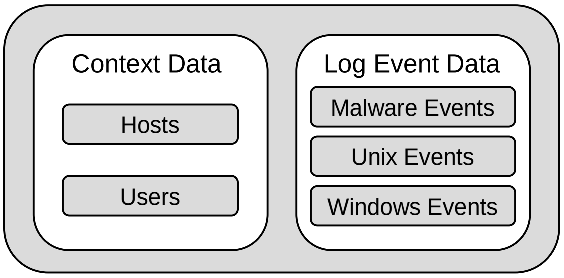


Figure 5. Overview of Input Data

The data set can be separated into actual log data and context data. The actual log data is divided into three types. On the one hand, logs contain log data originating from the Windows operating systems. On the other hand, for UNIX hosts, similar data was made available. The most interesting log data in respect to the paper is the malware log data. The malware event records contain information about the time, location, and type of malware found on a system.

The context data consists of records representing the computers (hosts) and the users of the enterprise's IT systems. These records offer data in several dimensions such as geographic and demographic information. The user records include fields containing information like the user's age and gender as well as his or her organizational status within the company. The host records include fields containing the computer's current status and the operating system running on it as well as information about the patch status of the operating system.

The resulting multidimensional data model, presented in Figure 6, illustrates the relations between the relevant dimensions containing different levels of hierarchy and the measures (facts).

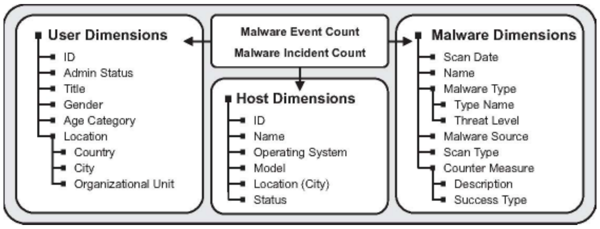


Figure 6. Multidimensional Data Model for Malware Analysis

The metrics Malware Event Count and Malware Incident Count can be analyzed according to these dimensions in any combination.

The User Dimensions include demographic information about users (e.g., gender, age category) who caused malware events as well as their admin status on the host where the malware was found. Additionally, their geographic location is tracked by the Location dimension, which consists of the hierarchy levels Country, City and Organizational Unit.

Information about the host computers on which malware events were found is provided by the Host Dimensions. The host Model is a description of the hardware. The host Status provides information about the host's current status in regard to anti-malware logging as well as information about the patch status of the operating systems.

The Malware Dimensions provide information about the Type of malware found and its Threat Level, which is either "high" or "low" by previous definition. E.g., cookies, adware, and joke programs are classified as low risks while malware such as viruses, trojans, and key loggers represent high risks. The Malware Source indicates the location of the malware; e.g., "local hard drive" or "internet browser files". Since anti-malware programs scan on a regular basis as well as on file access, the corresponding scan types are the elements of the dimension Scan Type. The countermeasures which are taken by the anti-malware software constitute the definition of another dimension (Counter Measure).

This multidimensional processed data also serve as data basis for the upcoming data mining process.

B. Prototyping an OLAP System for SIEM

Dashboards are usually used to visualize different, distributed information in a concentrated and integrated form. Relevant information is qualified in order to represent large quantities of information to the decision makers more clearly. Dashboards enable organizations to measure, monitor, and manage business objectives more effectively in the consequence [24]. In the case of SIEM, security dashboards are deployed in order to visualize security relevant data.

The dashboard illustrated in Figure 7 is currently set up to enable analysis of malware permanence and propagation. Here, the four reports merely provide descriptions of the data, indicating irregularities. Thus, they provide the starting point for a more accurate analysis, which is only possible within the individual organizational context. Since the original results of the data analysis are not allowed to be published due to confidentiality requirements, it has to be stressed that the following findings base on generated random data. Nevertheless, the results convey an impression about the possible outcomes of such an analysis.

Report no. 1 depicts the top five malware programs measured by the number of affected hosts, the number of affected users, and the duration of the malware in the institutions' IT systems. The malware "JS/Downloader-AUD" stands out, infecting 664 hosts and 431 users. It was present on at least one host on 322 days which is virtually every day in the given time frame of one year. This result implies that this particular malware either remains on the system or returns frequently.

A specific top five list of malware affections is helpful to identify particular pertinent malware and thus is a valuable tool for risk management. The types of malware visualized in the diagram can be filtered while the time period of the collected data can be adapted to one's need. The variability of such dimensions is a main feature of multidimensional OLAP analysis.

Report no. 2 illustrates the long-term development of the number of hosts and users infected with malware. Once countermeasures have been applied, this diagram can be used to control the measure effects. Scaling from quarters to months or even days, the diagram can also serve for medium to short-term controlling tasks and is thus another useful tool for risk management.

The reports no. 3 and 4 give details about the most frequent malware, in this case of the "JS/Downloader-AUD". The left diagram represents the success of malware elimination over time, the right one shows the presence of the malware in the IT systems over time. In this chart, strong excursions are to be recognized. Even after deleting the malware successfully, it seems that the malware re-emerges quickly. Further investigations concerning this malware should be accomplished.

During the project, several more dashboards were developed to enable users to analyze malware findings in regard to geographical aspects, for instance. The associated reports are represented as color coded maps in which significant occurrences of malware affection can be recognized rapidly. Further more, occurrences can be examined in detail by drilling down. With this opportunity, enterprises are able to identify locations which particularly cause the malware spreading. Thus, it can be derived in which organizational units security measures have to be improved immediately. Another dashboard visualizes user groups which cause various malware, by demographic characteristics. In this way, various age groups and/or gender-specific classes can be identified that correlate with increased malware affection. This information could be utilized to design specifically targeted awareness measures aiming to significantly reduce malware infections amongst the users and for other purposes.

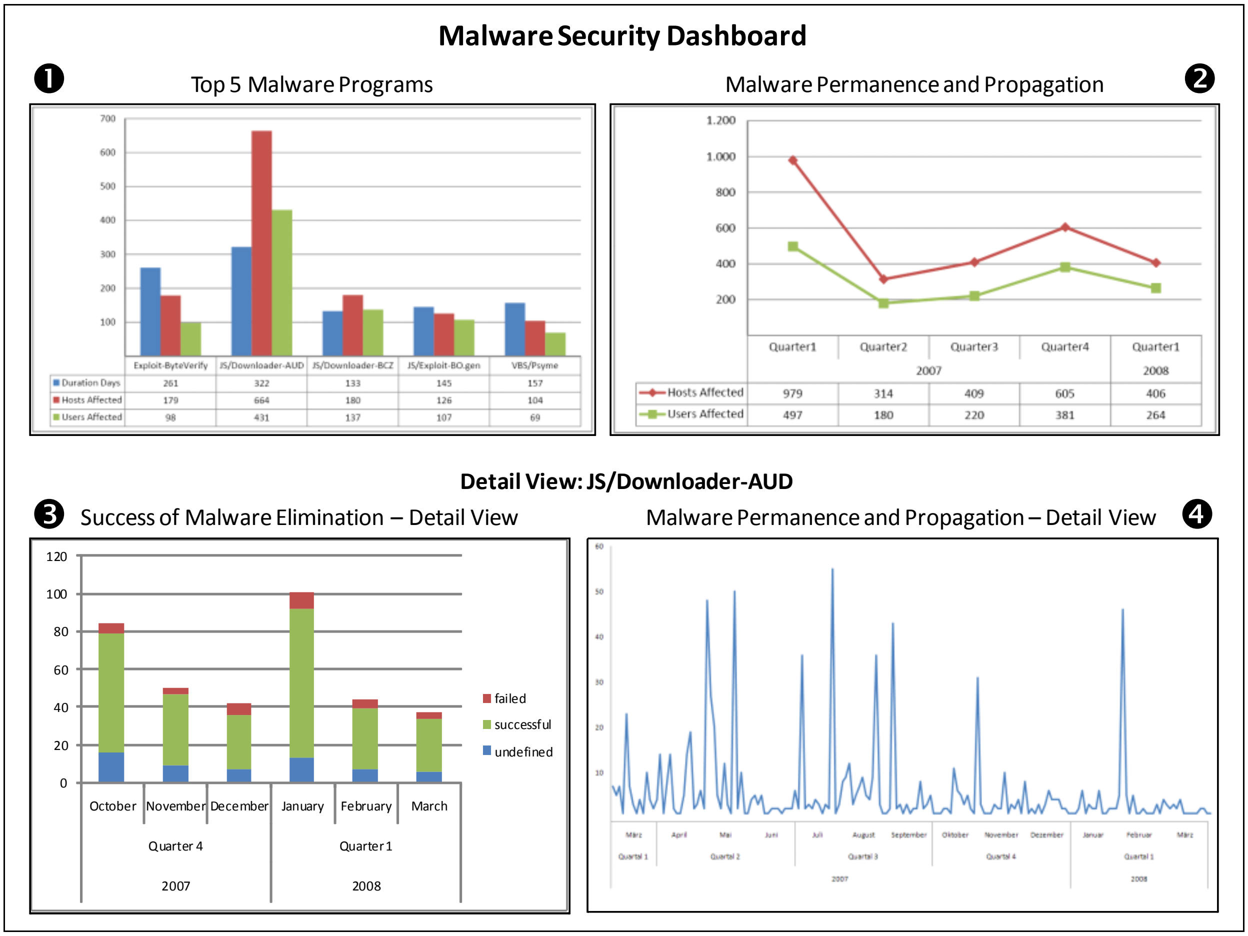


Figure 7. Excample of a Malware Security Dashboard

To sum up, all OLAP functions specified above can be used for detailed analysis in dashboards. First of all, dashboards give a general overview of the relevant measures, but also can be designed for presenting important details. Additionally, using a reporting tool, many other OLAP reports can easily be generated by accessing the data warehouse. Here, dimensions can be combined flexibly in order to analyze measures in regard to the perspectives of individual interest. In the consequence, OLAP enables a powerful descriptive analysis and effectively supports SIEM.

V. Data Mining Research Objectives

Undoubtedly, the simple storage of security relevant data alone does not enable to draw sensible conclusions from the data in order to support SIEM. Data by itself is of little direct value since potential insights are buried within and are often very hard to uncover. As described above, OLAP and dashboards are one way to analyze and visualize data which is modeled multidimensionally and stored in a data warehouse. Data mining is another option. The concept of data mining provides specific algorithms for data analysis, like association analysis, clustering, or classification [25]. These algorithms originate from diverse research fields, like statistics, pattern recognition, database engineering, and data visualization, for instance.

It has to be stressed, that the application of data mining algorithms must be accompanied by preparatory as well as post processing steps [25]. As Fayyad et al. point out, “blind application of data mining methods can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns” [20]. In order to conduct the necessary steps, and to analyze the data efficiently / effectively, the Cross Industry Process for Data Mining (CRISP-DM) was used [26]. CRISP-DM is an industry- and tool-neutral process model for data mining analysis which was and still is applied in several industry sectors successfully.

Actually, every single log event is potentially interesting for further investigative analysis. Since most organizational IT networks are in some way connected to the Internet and are thus subject to attacks from outside, the most popular application of data mining on log data is concerned with intrusion detection [27; 28]. In addition, questions to be answered by analyzing the log data could be why, where, when, and how long a malware incident happened and who was involved and responsible. In order to attain new and useful insights from the log data of interest, the following research objectives were identified.

A. *Objective 1: Relationship Analysis of Attributes Affecting Malware Infection*

One goal of applying data mining techniques is to identify interesting, unknown and relevant patterns in the data. Rules help to verbalize and quantify the patterns. The resulting set of rules can then be further analyzed by a human expert who decides how these rules will further be used in the process of SIEM. Among the different methodologies which are used to extract rules from a given data set, the authors of this paper focused on the association analysis. This method aims to discover interesting relationships between the attributes of a data set [29]. For this purpose, the two measures support and confidence are used. They indicate the interestingness of a relationship. Support quantifies how frequently a rule is applicable to a given data set, while confidence indicates how often items in B appear in transactions that contain A [29]. As depicted in Figure 8, the support of 2% means that in 2% of the whole set of hosts, Windows XP and a malware incident went along with each other. The confidence of 10% conveys that malware incidents occurred on 10% of all Windows XP hosts.

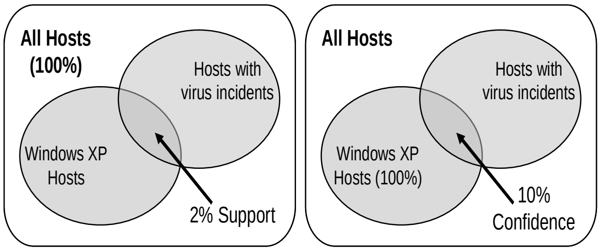


Figure 8. Support and Confidence

Mathematically, support and confidence can be represented as in the following equations, where A is the antecedent and B the consequent of the rule:

– Support $(A \rightarrow B) = P(A \cup B)$;

– Confidence $(A \rightarrow B) = P(B \mid A)$.

Since many relationships may exist between the attributes causing malware incidents, the following research objective has been stated: “Given malware incidents with certain attributes, find associations between those attributes, and state them as rules satisfying a minimum confidence and support.”

B. *Objective 2: Malware Permanence and Propagation Analysis*

Another interesting question is how malware spreads in the IT landscape and how long it resides in the system. Such a profile may contain data about the number of computers and users affected by malware incidents and the duration the malware resides within the IT infrastructure. Thus, the second objective of mining the security relevant data aims to analyze malware permanence and propagation.

Here, the k-means algorithm was applicable in order to cluster malware incident records in dependence of their similarity. Describing similarity is the main task of clustering algorithms. Similar records are put into the same cluster, whereas dissimilar records are allocated to different clusters. Thus, the second research objective was stated as “given a set of *n* malware incidents, group them by similarity into *k* clusters”.

VI. Data Mining Results

Since the results of the data mining analysis are not allowed to be published due to confidentiality reasons, the following findings also base on randomized data. Nevertheless, the results convey an impression on the possible outcomes of using data mining techniques for supporting SIEM.

A. *Findings of the Relationship Analysis*

Since the Apriori algorithm is appropriate for analyzing small or mid-size data sets, the authors have decided to apply this algorithm to provide an answer for research objective 1 [29]. Table I depicts an extract of random data which served as input in this context.

TABLE I. Overview of Database Extract

No.	User Age	User is Admin	Malware Risk
1.	IV	true	low
2.	V	false	low
3.	III	false	low
4.	II	true	high
n.	...	...	...

Each row represents a virus incident with three attributes. Thereby, the user ages are grouped into one of five classes with “I” for the youngest employees to “V” for the eldest ones. In order to find out which attributes are associated with high malware risks (or low malware risks, respectively), the different types of malware had to be assessed prior to the analysis. This was done by adding a new attribute to the data table for malware risks. Thus, it was possible to assign each user a “low” or “high” malware risk. Like done for OLAP, cookies, adware, and joke programs were classified as low risk while malware such as viruses, trojans and key loggers, was classified as high risk.

Since the data mining analysis focused malware affecting indicators, only those item sets were regarded which contain the risk attribute. In order to gain significant rules, support and confidence factors, as shown in Table II, were calculated.

TABLE II. ASSOCIATION RULES

hoher Malware-Befall, wenn		Support %	Confidence %
1.	user age category = E and user gender = male	9.5	82.7
2.	user age category = D and user gender = male and user is admin = false	5.3	75.6
...	...	...	...

niedriger Malware-Befall, wenn		Support %	Confidence %
...	user is admin = true and user gender = female	1.5	50.7
n.	user age category = E and user gender = female	8.7	60.9

The Apriori algorithm made it possible to separate the rule set. Rules with a confidence of less than 70% and a support of below 5% were not taken into account. The upper part of the table displays the rules which lead to high malware affection. The lower part displays those rules with low malware affection, respectively. The support of rule 1, as shown in the table, allows to conclude that in 9.5% of malware incidents the user's age category is IV, the user's gender is male, and the malware affection was high. The confidence of rule 1 indicates that in 82.7% of those malware incidents where the age category is IV and the user's gender is male, the malware affection is high.

It was tempting to interpret the rules indicating low malware affection similarly. However, the analysis only included records which already represented at least one incident. The "low malware affection" incidents merely occurred on hosts with less malware incidents. Thus, the last two rules have to be interpreted with specific attention, since they merely indicated lower affections than rules 1 and 2, for instance, but not a complete absence of it.

B. *Findings of the Malware Permanence and Propagation Analysis*

Data mining aiming to describe the permanence and propagation of malware incidents throughout the hosts of the enterprise was not performed in a straightforward fashion such as for the association analysis. The efforts put into this task are described now.

In order to narrow the analysis focus, measures for malware permanence and propagation were defined. The propagation of malware is described by the number of hosts and number of users a specific malware has affected. The duration of a malware infection can serve as measure for malware permanence. With background of these measures, concrete data sources were defined. Here, the malware event data served as basis for what reason no further data preparation was necessary.

The most difficult measure to extract from the data was the duration of a malware infection. A malware infection in this context is defined as the duration in which the same malware was present on different hosts within the entire enterprise. So, if a specific malware was identified on at least one host at the beginning of April and again in the middle of April, one is dealing with two separate infections. The malware incident data thus was aggregated once more to provide information about such infections. This time, the aggregation had to be performed along the date attribute of the malware incidents. Incidents with the same malware and similar dates were aggregated to the same malware infection group.

In order to identify similar dates, a grouping algorithm was applied. The algorithm devised for the present use case groups data objects by date and malware ID. The results were a number of classes, each containing a number of data objects with the same malware ID and a similar date. The algorithm performs the following steps for each identified malware ID:

(1) Sort all data objects by date.

(2) Create an initial empty group.

(3) Go through the data objects systematically and compare each date to the date of the previous one. If dates are similar, put the current data object into the just opened group. Otherwise, close the open group and create a new one containing the current object. Similarities between dates may be parameterized. In the case above, dates were considered dissimilar if they were more than 7 days apart.

Finally, the attribute "group" was added to each record. This attribute will have the value "0" if the record belongs to no group and a different number if it is part of a malware infection group. The result was a number of groups, each containing data objects with the same malware ID and a similar date. The grouping algorithm was parameterized during test runs in such a way that most groups contain either mostly malware incidents with high malware affection or mostly those with low malware affection.

After pre-processing the data, a cluster analysis was performed. Some findings are depicted in Figure 9. Due to the already mentioned confidentiality reason, real values must not be shown; hence, the results of the analysis cannot be discussed in detail. Since the k-means algorithm has been proven to be effective in producing good clustering results for many practical applications, this method was applied for clustering the malware incidents [30]. The attribute distributions indicate if the administrative privileges, the age, and the gender result in uncommon malware affection.

In total, eight clusters were identified. Figure 9 shows cluster 1 and 3 which were the most extensive ones. Cluster 1 includes 22%, whereas cluster 3 contains 19% of all malware incidents. Cluster 1 reveals that male users (cell 2) are likely to be affected by low-risk malware (cell 1) while in cluster 3 female users (cell 6) are in danger of being affected by high-risk malware (cell 5). Further on, cluster 1 indicates that middle-aged employees tend to be infected by malware (cell 3). The admin status does not seem to have influence on malware infection in this cluster (cell 4) what is surprising. This is also the case in cluster 3 (cell 8). In contrast to cluster 1, cluster 3 reveals that younger and elder employees tend to have malware on their computers (cell 7).

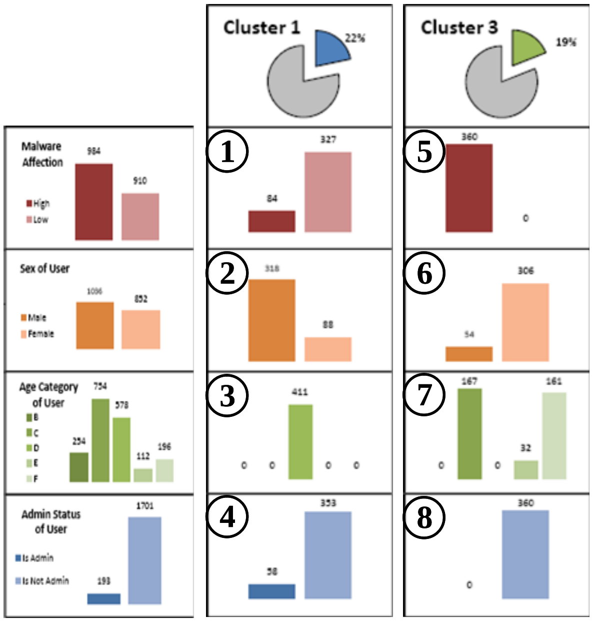


Figure 9. Results of the Cluster Analysis

VII. Conclusion

While BI systems are used in many enterprises to support classical business entities like the controlling or production one, they usually have little to no experience with BI systems in the context of SIEM. Taking the benefits of a classic BI system into account, this paper focused on the option of using OLAP and data mining techniques for the purposes of SIEM. Based on results of a project with an international enterprise, it can be derived that OLAP and data mining strongly support information security management teams. The gathered data can be analyzed more efficiently and patterns can be found which were previously hidden. Although the methods do not increase the detection ratio of malware directly, they support in finding internal (and external) factors which influence malware infestation. As result, measures (like awareness campaigns) could be set up to increase performances of running traditional measures like anti-virus and intrusion detection systems.

It has to be stressed that the quality of the data is crucial for success and that interpretation questions in regard to false positives and false negatives were not in the focus of this paper. Thus, the implementation of an adequate ETL process to transfer data from the source into the data warehouse correctly and consistently is as important as the validation of the accuracy of the data. In order to narrow the entire set of data to a manageable subset and to ensure that this subset matches the needs of the decision makers, the data relevance must be judged. In addition, an appropriate multidimensional data model which serves as the basis for flexible data analyses has to be designed.

While many research papers focused the analysis of log data e.g., for web marketing purposes, the analysis of security relevant log data has barely been explored. As result of the named project, it was exemplified that the so called native data mining methods are applicable for the analysis of security relevant log data.

Although the results presented in this paper are based on random data, rules were identified throughout the data mining project indicating that the age of a user has impact on malware affection on the one hand and that the user’s gender influences malware occurrences on the other hand. At the same time, it had to be stated that the admin status of a user does not seem to have influence on malware affection. However, the findings should not be generalized as they may relate to specific circumstances of the project conducted.

Due to the amount of data processed during the timeframe of the project, major efforts had to be made to ensure the quality of the log data in regard to its readiness for analysis. Though not being in the focus of this paper, it has to be stated that the application of a data mining process, like CRISP-DM for instance, is a crucial success factor in this context.

VIII. Future Work

Naturally, the results of the association analysis should provide information about relationships between the different attributes which influence the number of malware occurrences on the enterprise’s hosts. Easily understandable representations of such information are rules. A rule might say that “if a user has administrative privileges on a host, this host does not have an abnormal high number of malware incidents”.

As for research objective 1 discussed in this paper, it seems sensible to create another model based upon a different technique in order to support or disprove the rules generated by Apriori. This can be achieved by training a clustering model with the k-means algorithm. An association rule might be supported by the cluster analysis, if at least one cluster can be associated to it. A cluster representing the rule stated above might contain only those records in which the user possessed administrative privileges and the host was subject to a relatively low number of malware occurrences.

In order to serve the goals of SIEM, future research has to focus on further fields of log data analysis. For example, policy violations could be monitored by the use of data mining methods. Since enterprises usually have a bulk of policies (like password and access rules or the enforcement of regular updates of anti-malware and operating system software) to which the users and hosts have to comply to, the corresponding security data cannot be handled manually. By applying the described data mining techniques here, factors for violations of policy compliance could be identified efficiently as well as countermeasures could be set up in a timely fashion in the consequence. Thereby, identified policy violation issues should be categorized, rated, and visualized automatically in a clearly arranged manner. Thus, the information security management executives can be provided with high-quality information. Thereby, data mining is a promising option to identify patterns inside the data sets which were previously hidden.

Another way to perform data analyses and visualize the results is OLAP. This technology leads to efficient identifications of policy compliance violations for which corresponding countermeasures could be set up rapidly. The presented OLAP approach should not only be limited to the own enterprise. Also, the standard reporting modules of anti-malware software can be substantially improved by integrating a function which enables to use dashboards as presented in the paper.

To sum up, the possibilities of BI in the context of SIEM are manifold. Thereby, data mining techniques offer the promising chance to extract new knowledge out of the seemingly unstructured set of continuously logged data on the one hand. On the other hand, OLAP enables various powerful descriptive analyses of measures according to different perspectives of interest. This knowledge again enables to design new or adjust current measures resulting in an enhancement of the quality of the entire information security infrastructure of the enterprise using BI for SIEM.

REFERENCES

[1] R. Gabriel, T. Hoppe, A. Pastwa, and S. Sowa, “Analyzing Malware Log Data to Support Security Information and Event Management: Some Research Results”, Proc. First International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2009), IEEE Press, Mar. 2009, pp. 108-113, doi: 10.1109/DBKDA.2009.26.

[2] K.C. Laudon and J.P. Laudon, Management Information Systems, Managing the Digital Firm, Prentice Hall International, Upper Saddle River, 2005.

[3] J.-C. Laprie, “Dependability of Computer Systems: from Concepts to Limits“, Proceedings of the 6th International Symposium on Software Reliability Engineering, 1995, pp. 2-11.

[4] S.C. Shih and H.J. Wen, “Building E-Enterprise Security: A Business View”, Information Systems Security, Vol. 12, No. 4, 2003, pp. 41-49.

[5] R. Anderson, Security Engineering, A Guide to Building Dependable Distributed Systems, Wiley & Sons, New York et al., 2008.

[6] B. Schneier, Secrets and Lies, Wiley & Sons, New York et al., 2004.

[7] ISO/IEC 17799:2005, Information technology – Code of practice for information security management, 2005.

[8] ISO/IEC 27001:2005, Information technology – Security techniques – Information security management systems – Requirements, 2005.

[9] M. Nyanchama and P. Sop, “Enterprise Security Management: Managing Complexity”, Information Systems Security, Vol. 9, No. 6, 2001, pp. 37-44.

[10] J. Biethahn, H. Mucksch, and W. Ruf, Ganzheitliches Informationsmanagement, Band I, 5th Edition, Oldenbourg, München et al., 2000.

[11] R. Gabriel and D. Beier, Informationsmanagment in Organisationen, Kohlhammer, Stuttgart, 2003.

[12] A. Williams, “Security Information and Event Management Technologies”, Siliconindia, Vol. 10, No. 1, 2006, pp. 34-35.

[13] D.F. Carr, “Security Information and Event Management”. Baseline, No. 47, 2005, p. 83.

[14] D. Hellriegel, S.E. Jackson, and J.W. Slocum, Management, South-Western College Publishing, Ohio, 1999.

[15] B. Gilmer, “Firewalls and security”, Broadcast Engineering, Vol. 43, No. 8, 2001, pp. 36-37.

[16] M. Anandrarajan, A. Anandrarajan, and C.R. Srinivasan, Business Intelligence Techniques, Springer, Berlin et al., 2004.

[17] P. Gluchowski and H.G. Kemper, “Quo Vadis Business Intelligence? Aktuelle Konzepte und Entwicklungstrends”, BI Spektrum, Vol. 1, No. 1, 2006, pp. 12-19.

[18] W.H. Inmon, Building the Data Warehouse, Wiley, New York et al., 1996.

[19] E.F. Codd, S.B. Codd, and C.T. Salley, Providing OLAP to User Analysts, An IT Mandate, White Paper, s.l., 1993.

[20] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases”, AI Magazine, Vol. 17, No. 3, 1996, pp. 37-54.

[21] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, Fundamentals of Data Warehouses, Springer, Berlin et al., 2000.

[22] W.H. Inmon, J.A. Zachman, and J.G. Geiger, Data Stores, Data Warehousing and the Zachman Framework, McGraw-Hill, New York, 1997.

[23] P. Rob and C. Coronel, Database Systems: Design, Implementation, and Management, Boston, 2007.

[24] W.W. Eckerson, Performance Daschboards: Measuring, Monitoring, and Managing Your Business, Wiley & Sons, New York et al., 2006.

[25] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2006.

[26] P. Chapman, J. Clinton, R. Kerber, T. Khazaba, T. Reinartz, C. Shearer, and R. Wirth, “CRISP-DM 1.0 Step-by-Step Data Mining Guide”, 2000, URL: http://www.crisp-dm.org/CRISPWP-0800.pdf, 22.09.2009.

[27] D.G. Conorich, “Monitoring Intrusion Detection Systems: From Data to Knowledge”, Information Systems Security, Vol. 13, No. 2, 2004, pp. 19-30.

[28] K. Yamanshi, J.-I. Takechu, and Y. Maruyama, “Data Mining for Security”, NEC journal of advanced technology, Vol. 2, No. 1, 2004, pp. 13-18.

[29] V. Kumar, M. Steinbach, and P.-N. Tan, Introduction to Data Mining, Addison Wesley, Upper Saddle River, 2005.

[30] K. Alsabti, S. Ranka, and V. Singh, “An Efficient K-Means Clustering Algorithm”, 1998, URL: http://www.cs.utexas.edu/~kuipers/readings/Alsabti-hpdm-98.pdf, 22.09.2009.

Enhancing Information Reliability through Backwards Propagation Of Distrust

Panagiotis Metaxas
Computer Science Department
Wellesley College
106 Central Street, Wellesley, MA 02481, USA
Email: pmetaxas@wellesley.edu

***Abstract*—Search Engines have greatly influenced the way we experience the web. Since the early days of the web people have been relying on search engines to find useful information. However, their ability to provide useful and unbiased information can be manipulated by Web spammers. Web spamming, the practice of introducing artificial text and links into web pages to affect the results of searches, has been recognized as a major problem for search engines. But it is mainly a serious problem for web users because they tend to confuse trusting the search engine with trusting the results of a search.**

In this paper, first we discuss the relationship between Web spam in cyber world and social propaganda in the real world. Then, we propose "backwards propagation of distrust," as an approach to finding spamming untrustworthy sites. Our approach is inspired by the social behavior associated with distrust. In society, recognition of an untrustworthy entity (person, institution, idea, etc) is a reason for questioning the trustworthiness of those that recommended this entity. People that are found to strongly support untrustworthy entities become untrustworthy themselves. In other words, in the society, distrust is propagated backwards.

Our algorithm simulates this social behavior on the web graph with considerable success. Moreover, by respecting the user's perception of trust through the web graph, our algorithm makes it possible to resolve the moral question of who should be making the decision of weeding out untrustworthy spammers in favor of the user, not the search engine or some higher authority. Our approach can lead to browser-level, or personalized server-side, web filters that work in synergy with the powerful search engines to deliver personalized, trusted web results.

An earlier version of this paper was presented at [35].

***Keywords*-Web search; Information Reliability; Web graph; Link structure; Web Trust; Web Spam**

I. Introduction

Search Engines have greatly influenced the way we experience the web. Since the early days of the web people have been relying on search engines to find useful information. When the web was relatively small, Web directories were built and maintained that were using human experts to screen and categorize pages according to their characteristics. By the mid 1990's, however, it was apparent that the human expert model of categorizing web pages would not scale. The first search engines appeared and they have been evolving ever since.

But what influences their evolution? The way a user interacts with a search engine is through the search results to a query that he or she has issued. Search engines know that the quality of their ranking will determine how successful they are. If users perceive the results as valuable and reliable, they will come again. Otherwise, it is easy for them to switch to another search engine.

Research in Information Retrieval has produced a large body of work that, theoretically, produces high quality search results. Yet, search engines admit that IR theory is but one of their considerations. One of the major issues that influences the quality of ranking is the effect that web spam has on their results. *Web spamming* is defined as the practice of manipulating web pages in order to influence search engines rankings in ways beneficial to the spammers. Spammers aim at search engines, but target the end users. Their motive is usually commercial, but can also be political or religious.

We should mention here that, to people unfamiliar with web spam, the term is confused with email spam. Even though both term describe manipulation of information to confuse people in cyberspace, which is why we call them both "spam", they are very different in the way we experience them. In particular, email spam is pushed onto the users through email and we can learn to recognize it easily. Web spam, on the other hand, is misinformation that we pull through search engines, and thus it is very difficult to learn to recognize it. Sometimes, the term "adversarial information retrieval" is used to describe web spam. A more descriptive name for it would be "search engine ranking manipulation."

One of the reasons behind the users' difficulty to distinguish trustworthy from untrustworthy information comes from the success that both search engines and spammers have enjoyed in the last decade. Users have come to trust search engines as a means of finding information, and spammers have successfully managed to exploit this trust.

From their side, the search engines have put considerable effort in delivering spam-free query results and have developed sophisticated ranking strategies. Two such ranking strategies that have received major attention are PageRank

[5] and HITS [27]. Achieving high PageRank has become a sort of obsession for many companies' IT departments, and the *raison d'être* of spamming companies. Some estimates indicate that at least 8% of all pages indexed is spam [12] while experts consider web spamming the single most difficult challenge web searching is facing today[21]. Search engines typically see web spam as an interference to their operations and would like to restrict it, but there can be no algorithm that can recognize spamming sites based solely on graph isomorphism [4].

In this paper, we first explain why we need to understand web spamming beyond the technical details. Web spamming is a social problem first, then a technical one, and there is strong relationship between it and social propaganda. In fact, [34] presents evidence of its influence on the evolution of search engines. Then, we describe and evaluate an algorithmic way of discovering spamming networks automatically. Finally, we discuss a general framework for the long-term approach to web spam

A. Background

Web spamming has received a lot of attention lately [3], [4], [12], [13], [17], [20], [21], [23], [28], [31], [32], [37], [34]. The first papers to raise the issue were [32], [21]. The spammers' success was noted in [3], [10], [12], [13], [22]. Web search was explained in [1]. The related topic of cognitive hacking was introduced in [11].

Characteristics of spamming sites based on diversion from power laws are presented in [12]. Current tricks employed by spammers are detailed in [16]. An analysis of the popular PageRank method employed by many search engines today and ways to maximize it in a spamming network is described in [4]. TrustRank, a modification to the PageRank to take into account the evaluations of a few seed pages by human editors, employees of a search engine, is presented in [17]. Techniques for identifying automatically link farms of spam pages were presented in [45], [2].

A comprehensive treatment on social networks is presented in [43]. The connection between the Web and social networks was explicitly noted in [29], [38] and implicitly used in [5], [27]. In fact, Kleinberg's work explores many of these connections (e.g., [26]). Identification of web communities was explored in [28], [14]. Work on topic-sensitive and personalized web search is presented in [19], [24]. The effect that search engines have on page popularity was discussed in [8].

Research in the past has focused on the identification of web communities through the use of bipartite cores [28] or maximum flow in dense subgraphs [14]. Some of the background information on Web Spam and its connection to social propaganda was presented in [34].

The rest of this paper is organized as follows. The next section gives an overview of the problem of information reliability and web spamming. Section II-B has a short introduction to the theory of propaganda detection and the next section II-C discusses the relationship between the Web Graph and the trust social network. The following section II-D analyzes the evolution of search engines as their response to spam. Section III describes the backward propagation of distrust algorithm and the following section IV presents some of our experimental results running this algorithm. Finally, the last section V has our conclusions and a framework for the long-term approach to web spam.

II. On Information Reliability and Web Spam

A. Web Spam

The web has changed the way we inform and get informed. Every organization has a web site and people are increasingly comfortable accessing it for information on any question they may have. The exploding size of the web necessitated the development of search engines and web directories. Most people with online access use a search engine to get informed and make decisions that may have medical, financial, cultural, political, security or other important implications in their lives [10], [42], [22], [31]. Moreover, 85% of the time, people do not look past the first ten results returned by the search engine [40]. Given this, it is not surprising that anyone with a web presence struggles for a place in the top ten positions of relevant web search results. The importance of the top-10 placement has given birth to a new "Search Engine Optimization" industry, which claims to sell know-how for prominent placement in search results and includes companies, publications, and even conferences. Some of them are willing to bend the truth in order to fool the search engines and their customers, by creating web pages containing web spam [12].

The creators of web spam are often specialized companies selling their expertise as a service, but can also be the web masters of the companies and organizations that would be their customers. Spammers attack search engines through text and link manipulations:

Text manipulations: This includes repeating text excessively and/or adding irrelevant text on the page that will cause incorrect calculation of page relevance; adding misleading meta-keywords or irrelevant "anchor text" that will cause incorrect application of rank heuristics.

Link manipulations: This technique aims to change the perceived structure of the Web Graph in order to cause incorrect calculation of page reputation. Such examples are the so-called "link-farms," domain flooding (plethora of domains that re-direct to a target site), page "awards," (the spammer pretends to run an organization that distributes awards for web site design or information; the awarded site gets to display the "award", an image linking back to awarding organization, effectively increasing the visibility of the spammer' site), etc.

Both kinds of spam aim to boost the ranking of spammed web pages. So as not to get caught, spammers conceal their actions through cloaking, content hiding and redirection. Cloaking, for example, aims to serve different pages to search engine robots and to web browsers (users). The spamming pages could be created statically or dynamically. Static pages, for example, may employ hidden links and/or hidden text with colors or small font sizes noticeable by a crawler but not by a human. Dynamic pages might change content on the fly depending on the visitor, submit millions of pages to "add-URL" forms of search engines, etc. We consider the false links and text themselves to be the spam, while, strictly speaking, cloaking is not spam, but a tool that helps spammers hide their attacks. For a comprehensive treatment of the spamming techniques, see [16].

Since anyone can be an author on the web, these practices have brought into prominence a question of *information reliability*. An audience used to trusting the written word of newspapers and books is unable, unprepared or unwilling to think critically about the information obtained from the web. A recent study [15] found that while college students regard the web as a primary source of information, many do not check more than a single source, and have trouble recognizing trustworthy sources online. In particular, two out of three students are consistently unable to differentiate between facts and advertising claims, even "infomercials." Very few of them would double-check for validity. At the same time, they have considerable confidence in their abilities to distinguish trustworthy sites from non-trustworthy ones, especially when they feel technically competent. We have no reason to believe that the general public will perform any better than well-educated students. In fact, a recent analysis of internet related fraud by a major Wall Street law firm [10] puts the blame squarely on the questionable critical thinking skills of the investors for the success of stock fraud cases.

B. Social Theory of Propaganda

On the outset, it may seem surprising that a technical article discusses social propaganda. This is a subject that has been studied extensively by social scientists and might seem out of the realm of computing. However, the web is a social network, influenced daily by the actions (intentional or otherwise) of millions of people. In that respect, web researchers should be aware of social theories and practices since they may have applicability in their work. We believe that a basic understanding of social propaganda can be valuable to technical people designing and using systems that affect our social interactions. In particular, it can be useful to researchers that study Web Spam. We offer here a brief introduction to the theory of propaganda detection.

There are many definitions of propaganda, reflecting its multiple uses over time. One working definition we will use here is

Propaganda is the attempt to modify human behavior, and thus influence people's actions in ways beneficial to propagandists.

Propaganda has a long history in modern society and is often associated with negative connotation. This was not always the case, however. The term was first used in 1622, in the establishment by the Catholic Church of a permanent Sacred Congregation *de Propaganda Fide* (for the propagation of faith), a department which was trying to spread Catholicism in non-Catholic Countries [44]. Its current meaning comes from the successful Enemy Propaganda Department in the British Ministry of Information during WWI. However, it was not until 1938, in the beginning of WWII, that a theory was developed to detect propagandistic techniques. For the purposes of this paper we are interested in ways of detecting propaganda, especially by automatic means.

First developed by the Institute for Propaganda Analysis [30], classic Propaganda Theory identifies several techniques that propagandists often employ in order to manipulate perception.

- **Name Calling** is the practice of giving an idea a bad label. It is used to make people reject and condemn the idea without examining the evidence. For example, using the term "miserable failure" to refer to political leaders such as US President George Bush can be thought of as an application of name calling.
- **Glittering Generalities** is the mirror image[1] of name calling: Associating an idea with a "virtue word", in an effort to make us accept and approve the idea without examining the evidence. For example, using the term "patriotic" to refer to illegal actions is a common application of this technique.
- **Transfer** is the technique by which the propagandist carries over the authority, sanction, and prestige of something respected and revered to something he would have us accept. For example, delivering a political speech in a mosque or a church, or ending a political gathering with a prayer have the effect of transfer.
- **Testimonial** is the technique of having some respected person comment on the quality of an issue on which they have no qualifications to comment. For example, a famous actor who plays a medical doctor on a popular TV show tells the viewers that she only uses a particular pain relief medicine. The implicit message is that if a famous personality trusts the medicine, we should too.
- **Plain Folks** is a technique by which speakers attempt to convince their audience that they, and their ideas, are "of the people," the "plain folks". For example, politicians sometimes are seen flipping burgers at a neighborhood diner.
- **Card Stacking** involves the selection of facts (or falsehoods), illustrations (or distractions), and logical

[1] Name calling and glittering generalities are sometimes referred to as "word games."

(or illogical) statements in order to give an incorrect impression. For example, some activists refer to the Evolution Theory as a theory teaching that humans came from apes (and not that both apes and humans have evolved from a common ancestor who was neither human nor ape).

- **Bandwagon** is the technique with which the propagandist attempts to convince us that all members of a group we belong to accept his ideas and so we should "jump on the band wagon". Often, fear is used to reinforce the message. For example, commercials might show shoppers running to line up in front of a store before it is open.

The reader should not have much trouble identifying additional examples of such techniques used in politics or advertising. The next section discusses the relationship of propaganda to web spam, by first describing the similarity of social networks to the web graph.

C. The Web Graph as a Trust Network

The web is typically represented by a directed graph [7]. The nodes in the Web Graph are the pages (or sites) that reside on servers on the internet. Arcs correspond to hyperlinks that appear on web pages (or sites). In this context, web spammers' actions can be seen as altering the contents of the web nodes (mainly through text spam), and the hyperlinks between nodes (mainly through link spam).

The theory of social networks [43] also uses directed graphs to represent relationships between social entities. The nodes correspond to social entities (people, institutions, ideas). Arcs correspond to recommendations between the entities they connect. In this context, propagandistic techniques can be seen as altering the trust social network by altering one or more of its components (i.e., nodes, arcs, weights, topology).

To see the correspondence more clearly, we will examine some of the propagandistic techniques that have been used successfully by spammers: The technique of testimonials effectively adds a link between previously unrelated nodes. Glittering generalities change the contents of a node, effectively changing its perceived relevance. Mislabeled anchor text is an example of card stacking. And the technique of bandwagon creates many links between a group of nodes, a "link farm". So, we define web spam based on the spammers actions:

Web Spam is the attempt to modify the web (its structure and contents), and thus influence search engine results in ways beneficial to web spammers.

Table I has the correspondence, in graph theoretic terms, between the web graph according to a search engine and the trust social network of a particular person. Web pages or sites correspond to social entities and hyperlinks correspond to trust opinions. The rank that a search engine assigns to a page or a site corresponds to the reputation a social entity has for the person. This rank is based on some ranking formula that a search engine is computing, while the reputation is based on idiosyncratic components associated with the person's past experiences and selective application of critical thinking skills; both are secret and changing.

This correspondence is more than a coincidence. The web itself is a social creation, and both PageRank and HITS are socially inspired ranking formulas. [5], [27], [38], [1]. Socially inspired systems are subject to socially inspired attacks. Not surprisingly then, the theory of propaganda detection can provide intuition into the dynamics of the web graph.

PageRank is based on the assumption that the reputation of an entity (a web page in this case) can be measured as a function of both the number and reputation of other entities linking to it. A link to a web page is counted as a "vote of confidence" to this web site, and in turn, the reputation of a page is divided among those it is recommending[2]. The implicit assumption is that hyperlink "voting" is taking place independently, without prior agreement or central control. Spammers, like social propagandists, form structures that are able to gather a large number of such "votes of confidence" by design, thus breaking the crucial assumption of independence in a hyperlink. But while the weights in the web graph are assigned by each search engine, the weights in the trust social network are assigned by each person. Since there are many more persons than search engines, the task of a web spammer is far easier than the task of a propagandist.

D. Search Engine Evolution

In the early 90's, when the web numbered just a few million servers, the **first generation** search engines were ranking search results using the vector model ([39], [20]) of classic information retrieval techniques: the more rare words two documents share, the more similar they are considered to be.

According to the *vector model* in Information Retrieval [39], documents contained in a document collection D are viewed as vectors in term space T. Under this formulation, rare words have greater weight than common words, because they are viewed as better representing the document contents. In the vector model, document similarity $sim(D_1, D_2)$ between document vectors D_1 and D_2 is represented by the angle between them. A search query Q is considered simply a short document and the results of a search for Q are ranked according to their (normalized) similarity to the query. While the exact details of the computation of term weights were kept secret, we can say that the ranking formula R^{G_1} in the first generation search engines was based in the following

[2] Since HTML does not provide for "positive" and "negative" links, all links are taken as positive. This is not always true, but is considered a reasonable assumption. Recently, Google introduced the "nofollow" attribute for hyperlinks, as a tool for blog site owners to mark visitor opinions. It is very unlikely that spamming blog owners will use it, however.

Graph Theory	*Web Graph*	*Trust Social Network*
Node	web page or site	social entity
weight	rank (accord. to a search engine)	reputation (accord. to a person)
weight computation	ranking formula (e.g., pagerank) computed continuously	idiosyncratic (e.g., 2 recommenders) computed on demand
Arc	hyperlink	trust opinion
semantics	"vote of confidence"	"recommendation"
weight	degree of confidence	degree of entrustment
weight range	$[0 \dots 1]$	$[distrust \dots trust]$

Table I

GRAPH THEORETIC CORRESPONDENCE BETWEEN THE WEB GRAPH AND THE TRUST SOCIAL NETWORK. THERE IS A ONE-TO-ONE CORRESPONDENCE BETWEEN EACH COMPONENT OF THE TWO GRAPHS. A MAJOR DIFFERENCE, HOWEVER, IS THAT, EVEN THOUGH A PERSON MAY FEEL NEGATIVE TRUST (DISTRUST) FOR SOME ENTITY, THERE IS NO NEGATIVE WEIGHT FOR HYPERLINKS.

principle: the more rare keywords a document shares with a query, the higher similarity it has with it, resulting in a higher ranking score for this document:

$$R^{G_1} = f(sim(p, Q)) \tag{1}$$

The first attack to this ranking came from within the search engines. In 1996, search engines started openly selling search keywords to advertisers [9] as a way of generating revenue: If a search query contained a "sold" keyword, the results would include targeted advertisement and a higher ranking for the link to the sponsor's web site.

Mixing search results with paid advertisement raised serious ethical questions, but also showed the way to financial profits to spammers who started their own attacks using **keyword stuffing**, i.e., by creating pages containing many rare keywords to obtain a higher ranking score. In terms of propaganda theory, the spammers employed a variation of the technique of *glittering generalities* to confuse the first generation search engines [30, pg. 47]:

The propagandist associates one or more suggestive words without evidence to alter the conceived value of a person or idea.

In an effort to nullify the effects of glittering generalities, **second generation** search engines started employing additionally more sophisticated ranking techniques. One of the more successful techniques was based on the "link voting principle": Each web site s has value equal to its "popularity" $|B_s|$ which is influenced by the set B_s of sites pointing to s.

Therefore, the more sites were linking to a site s, the higher the popularity of s's pages. Lycos became the champion of this ranking technique [33] and had its own popularity skyrocket in late 1996. Doing so, it was also distancing itself from the ethical questions introduced by blurring advertising with ranking [9].

The ranking formula R^{G_2} in the second generation search engines was a combination of a page's similarity, $sim(p, Q)$, and its site's popularity $|B_s|$:

$$R^{G_2} = f(sim(p, Q), |B_s|) \tag{2}$$

To avoid spammers (and public embarrassment from the keyword selling practice), search engines would keep secret their exact ranking algorithm. Secrecy is no defense, however, since secret rules were figured out by experimentation and reverse engineering. (e.g., [37], [32]).

Unfortunately, this ranking formula did not succeed in stopping spammers either. Spammers started creating clusters of interconnected web sites that had identical or similar contents with the site they were promoting, a technique that subsequently became known as **link farms**. The link voting principle was socially inspired, so spammers used the well known propagandistic method of *bandwagon* to circumvent it [30, pg. 105]:

With it, the propagandist attempts to convince us that all members of a group to which we belong are accepting his program and that we must therefore follow our crowd and "jump on the band wagon".

Similarly, the spammer is promoting the impression of a high degree of popularity by inter-linking many internally controlled sites that will eventually all share high ranking.

PageRank and HITS marked the development of the **third generation** search engines. The introduction of PageRank in 1998 [5] was a major event for search engines, because it seemed to provide a more sophisticated anti-spamming solution. Under PageRank, not every link contributes equally to the "reputation" of a page $PR(p)$. Instead, links from highly reputable pages contribute much higher value than links from other sites. That way, the link farms developed by spammers would not influence much their PageRank, and Google became the search engine of choice. HITS is another socially-inspired ranking which has also received a lot of attention [27] and is reportedly used by the AskJeeves search engine. The HITS algorithm divides the sites related to a query between "hubs" and "authorities". Hubs are sites that contain many links to authorities, while authorities are sites pointed to by the hubs and they both gain reputation.

Unfortunately, spammers again found ways of circumventing these rankings. In PageRank, a page enjoys absolute reputation: its reputation is not restricted on some particular issue. Spammers deploy sites with expertise on irrelevant subjects, and they acquire (justifiably) high ranking on their expert sites. Then they bandwagon the irrelevant expert sites, creating what we call a **mutual admiration society**. In propagandistic terms, this is the technique of *testimonials* [30, pg. 74] often used by advertisers:

Well known people (entertainers, public figures, etc.) offer their opinion on issues about which they are not experts.

Spammers were so aggressive in pursuing this technique that they openly promoted "reciprocal links": Web masters controlling sites that had some minimum PageRank, were invited to join a mutual admiration society by exchanging links, so that at the end everyone's PageRank would increase. HITS has also shown to be highly spammable by this technique due to the fact that its effectiveness depends on the accuracy of the initial neighborhood calculation.

Another heuristic that third generation search engines used was that of exploiting "anchor text". It had been observed that users creating links to web pages would come to use, in general, meaningful descriptions of the contents of a page. (Initially, the anchor text was non-descriptive, such as "click here", but this changed in the late 1990's.) Google was the first engine to exploit this fact noting that, even though IBM's web page made no mention that IBM is a computer company, many users linked to it with anchor text such as "computer manufacturer".

Spammers were quick to exploit this feature too. In early 2001, a group of activists started using the anchor text "miserable failure" to link to the official Whitehouse page of American President George W. Bush. Using what became known as "Googlebomb" or, more accurately, **link-bomb** since it does not pertain to Google only, other activists linked the same anchor text to President Carter, filmmaker Michael Moore and Senator Hilary Clinton.

Using the anchor text is socially inspired, so spammers used the propagandistic method of *card stacking* to circumvent it [30, pg. 95]:

Card stacking involves the selection and use of facts or falsehoods, illustrations or distractions, and logical or illogical statements in order to give the best or the worst possible case for an idea, program, person or product.

The ranking formula R^{G_3} in the third generation search engines is, therefore, some secret combination of a number of features, primarily the page's similarity, $sim(p, Q)$, its site's popularity $|B_s|$ and its the page's reputation $PR(p)$:

$$R^{G_3} = f(sim(p, Q), |B_s|, PR(p)) \qquad (3)$$

Search engines these days claim to have developed hundreds of little heuristics for improving their web search results [18] but no big idea that would move their rankings beyond the grasp of spammers. As Table II summarizes, for every idea that search engines have used to improve their ranking, spammers have managed quickly to balance it with techniques that resemble propagandistic techniques from society. Web search corporations are reportedly busy developing the engines of the next generation [6]. The new techniques aim to be able to recognize "the need behind the query" of the user. Given the success the spammers have enjoyed so far, one wonders how will they spam the fourth generation engines. Is it possible to create a ranking that is not spammable? Put another way, can the web as a social space be free of propaganda?

This may not be possible. Our analysis shows that we are trying to create in cyberspace what societies have not succeeded in creating in their real space. However, we can learn to live in a web with spam as we live in society with propaganda, given appropriate education and technology.

III. An Anti-propagandistic Algorithm

Since spammers employ propagandistic techniques [34], it makes sense to design anti-propagandistic methods for defending against them. These methods need to be user-initiated, that is, the user decides which web site not to trust and then seeks to distrust those supporting the untrustworthy web site. We are considering trustworthiness to be a personal decision, not an absolute quality of a site. One person's gospel is another's political propaganda, and our goal is to design methods that help individuals make more informed decisions about the quality of the information they find on the web.

Here is one way that people defend against propaganda in every day life:

In society, distrust is propagated backwards: When an untrustworthy recommendation is detected, it gives us a reason to reconsider the trustworthiness of the recommender. Recommenders who strongly support an untrustworthy recommendation become untrustworthy themselves.

This process is selectively repeated a few times, propagating the distrust backwards to those who strongly support the recommendation. The results of this process become part of our belief system and are used to filter future information. (Note that distrust is not propagated forward: An untrustworthy person's recommendations could be towards *any* entity, either trustworthy or untrustworthy.)

We set out to test whether a similar process might work on the web. Our algorithm takes as input s, a web site, which is represented by the URL of the server containing a page that the user determined to be untrustworthy. This page could have come to the user through web search results, an email spam, or via the suggestion of some trusted associate (e.g., a society that the user belongs to).

The obvious challenge in testing this hypothesis would be to retrieve a neighborhood of web sites linking to the starting site s in order to analyze it. Since we are interested in back links to sites, we can not just follow a few forward links (hyperlinks on web sites) to get this information. Otherwise we would need to possibly explore the whole web graph. Today, only search engines have this ability. Thankfully, search engines have provided APIs to help with our task.

S.E.'s	Ranking	Spamming	Propaganda
1st Gen	Doc Similarity	keyword stuffing	glittering generalities
2nd Gen	+ Site popularity	+ link farms	+ bandwagon
3rd Gen	+ Page reputation + anchor text	+ mutual admiration societies + link bombs	+ testimonials + card stacking

Table II
CHANGES IN RANKING BY GENERATIONS OF SEARCH ENGINES, THE RESPONSE OF THE WEB SPAMMERS AND THE CORRESPONDING PROPAGANDISTIC TECHNIQUES.

Starting from s we build a breadth-first search (BFS) tree of the sites that link to s within a few "clicks" (Figure 1). We call the directed graph that is revealed by the back-links, the "trust neighborhood" of s. We do not explore the web neighborhood directly in this step. Instead, we can use the Google API for retrieving the back-links.

Referring to Figure 1, if one deems that starting site 1 is untrustworthy, and sites 2, 3, 4, 5 and 6 link directly to it, one has reasons to be suspicious of those sites too. We can take the argument further and examine the trustworthiness of those sites pointing to 2, ... 6. The question arises on whether we should distrust all of the sites in the trust neighborhood of starting site s or not. Is it reasonable to become suspicious of every site linking to s in a few steps? They are "voting in confidence" after all [5], [27]. Should they be penalized for that? Such a radical approach is not what we do in everyday life. Rather, we selectively propagate distrust backwards only to those that most strongly support an untrustworthy recommendation. Thus, we decided to take a conservative approach and examine only those sites that use link spamming techniques in supporting s. In particular, we focused on the biconnected component (BCC) that includes s (Figure 2).

A BCC is a graph that cannot be broken into disconnected pieces by deleting any single vertex. An important characteristic of the BCC is there are at least two independent paths from any of its vertices to s. Strictly speaking, the BCC is computed on the undirected graph of the trust neighborhood. But since the trust neighborhood is generated through the BFS, the cross edges (in BFS terminology) create cycles in the undirected graph (Figure 1). Each cycle found in the BCC must have at least one "ring leader", from which there are two directed paths to s, one leaving through the discovery edge and the other through the cross edge. We view the existence of multiple paths from ring leaders to s as evidence of strong support of s. The BCC reveals the members of this support group. The graph induced by the nodes not in the BCC is called "BFS periphery".

More formally, the algorithm is as follows:

```
Input:
  s = Untrustworthy starting site's URL
  D = Depth of search
  B = Number of back-links to record
```

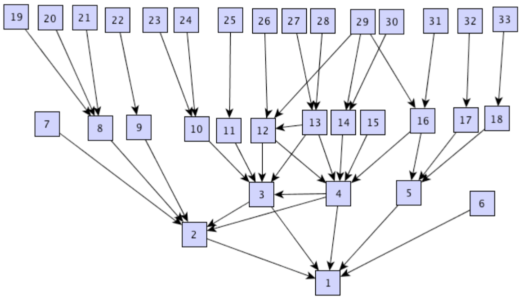


Figure 1. An example of a breadth-first search tree in the trust neighborhood of site 1. Note that some nodes (12, 13, 16 and 29) have multiple paths to site 1. We call these nodes "ring leaders" that show a concerted effort to support site 1.

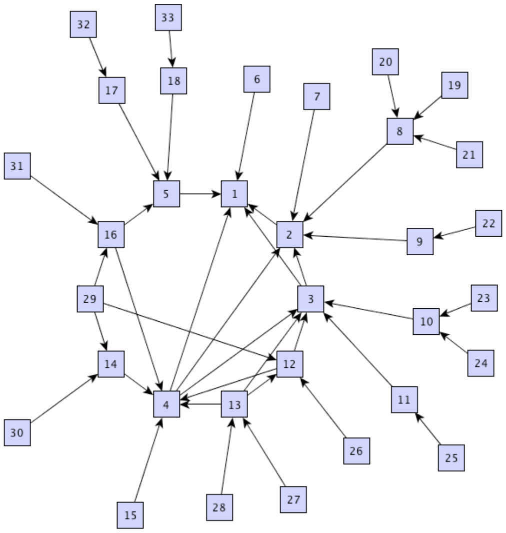


Figure 2. The BCC of the trust neighborhood of site 1 is drawn in a circular fashion for clarity. Note that the BCC contains the "ring leaders," that is, those nodes with multiple paths leading to s. The graph induced by the nodes not in the BCC is called "BFS periphery".

```
S = {s}
Using BFS for depth D do:
  Compute U={sites linking to sites in S}
    using the Google API
    (up to B back-links / site)
  Ignore blogs, directories, edu's
  S = S + U
Compute the BCC of S that includes s

Output: The BCC
```

A. Implementation Details

To be able to implement the above algorithm at the browser side, we restrict the following parameters: First, the BFS's depth D is set to 3. We are not interested in exploring a large chunk of the web, just a small neighborhood around s. Second, we limit the number B of back-link requests from the Google API to 30 per site. This helps reduce the running time of our algorithm since the most time-consuming step is the query to Google's back-link database. Finally, we introduced in advance a set of "stop sites" that are not to be explored further.

A *stop site* is one that should not be included in the trust neighborhood either because the trustworthiness of such a site is irrelevant, or because it cannot be defined. In the first category we placed URLs of educational institutions (domains ending in .edu). Academicians are not in the business of linking to commercial sites [36]. When they do, they do not often convey trust in the site. College libraries and academicians, for example, sometimes point to untrustworthy sites as examples to help students critically think about information on the web. In the latter category we placed a few well known Directories (URLs ending in yahoo.com, dmoz.org, etc.) and Blog sites (URLs containing the string ’blog’ or ’forum’). While blogs may be set up by well meaning people who are trying to increase the discourse on the web, blog pages are populated with opinions of many people and are not meant to represent the opinion of the owner. Anyone can put an entry into an unsupervised blog or directory, and following a hyperlink from a blog page should not convey the trustworthiness of the whole blog site. If the search engines were able to distinguish and ignore links inside the comments, blogs could be removed from the stop sites. No effort to create an exhaustive list of blogs or directories was made.

With these restrictions, our algorithm can be implemented on an average workstation and produce graphs with up to a few hundred nodes within minutes. As we mentioned, the most time demanding step is requesting and receiving the back-link lists from Google, since it requires initiating an online connection. No connections to the particular web sites was done during the creation of the trust neighborhood. Performing the BFS and computing the BCC of the graph assembled is done in time linear on the number of sites retrieved, so it is fast. We used the JUNG software library [25] to maintain the web subgraph and compute its BCC. The whole neighborhood can fit into the main memory of the workstation, so this does not require additional time.

IV. Finding Untrustworthy Neighborhoods that use Link Spam

There are several ways one can run into an initial untrustworthy site to use it as a starting site s. For example, search results for queries that happen to be controversial (e.g., "Armenian genocide", "morality of abortion" and "ADHD real disease") or happen to be the source of unreliable advertisement (e.g., "human growth hormone increase muscle mass"), contain plethora of responses that can be considered untrustworthy. In our experiments, we examined the trust neighborhoods of eight untrustworthy and two trustworthy sites. In Table III below these sites are labeled as U-1 to U-8 and T-1 to T-2, respectively. See Figure 3 for an example of U-1.

We run the experiments between September 17 and November 5, 2004. At the time of the experiment, all sites happen to have comparable PageRank, as reported by the Google Toolbar. In fact, U-1 and T-1 both had PageRank 6 while the remaining sites had PageRank 5. We recorded the PageRank numbers as reported by the Google Toolbar because this is always one of the first questions people ask and because the spamming industry seems to use it as a measure of their success. In fact, one can find spam networks inviting the creation of "reciprocal links" for sites that have at lease a minimum of PageRank 5, in order to increase their overall PageRank.

To determine the trustworthiness of each site we had a human evaluator look at a sample of the sites of the BCC. The results of our experiments appear on Table III. Due to the significant manual labor involved, only 20% of the total 1,396 BCC sites were sampled and evaluated. To select the sample sites, we employed stratified sampling with skip interval 5. The stratum used was similarity of the site to the starting site.

Each site in the sample was classified as either Trustworthy, Untrustworthy, or Non-determined. The last category includes a variety of sites for which the evaluator could not clearly classify.

We have two main results:

1. The trustworthiness of the starting site is a very good predictor for the trustworthiness of the BCC sites.

In fact (see Table 1), there were very few trustworthy sites in the trust neighborhoods of sites U-1 to U-8. The reason is, we believe, that a trustworthy site is unlikely (though not impossible) to deliberately link to an untrustworthy site, or even to a site that associates itself with an untrustworthy one. In other words, *the "vote of confidence" link analogy holds true only for sites that are choosing their links responsibly.* The analogy is not as strong when starting from a trustworthy site, since untrustworthy sites are free to link to whomever they choose. After all, there is some value in portraying a site in good company: Non-critically

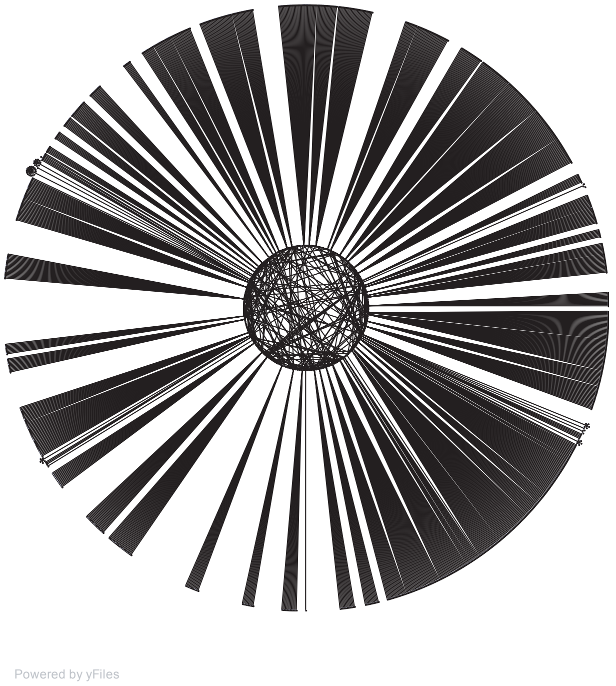


Figure 3. The trust graph of starting site U-1. The circularly drawn nodes in the middle form its largest biconnected component. This experiment found a trust graph of 1307 sites, 228 of which were connected with 465 edges into a **bi-connected component (BCC)**. The central, circularly drawn component is the BCC, while the sites drawn on the **BCC Periphery** were the remaining 1079 sites discovered by the BFS algorithm. Only 2% trustworthy sites were found in the BCC, while 74% of them were untrustworthy. In contrast, 31% trustworthy and 33% untrustworthy sites were found in the BFS periphery. The remaining sites were mostly directories or other non-determined sites.

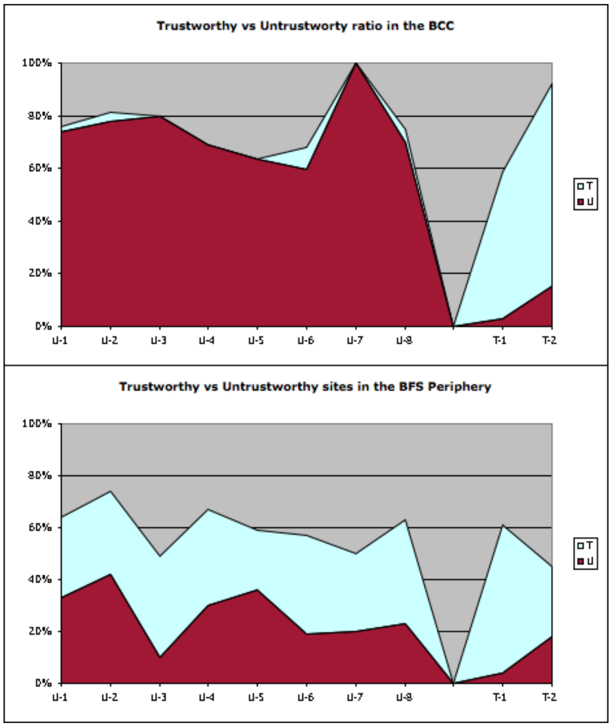


Figure 4. The trustworthy and untrustworthy percentages for trust neighborhoods of the BCC (top) and BFS peripheral (bottom) sites for the data shown in Table III. On the horizontal coordinates are shown 8 untrustworthy (on the left) and 2 trustworthy sites (on the right side of each graph). The vertical coordinates are the percentages of untrustworthy (U) and trustworthy (T) sites found in the neighborhood of each starting site. Comparing the left and right sides of the top graph, one can see that the trustworthiness of the starting site is a very good predictor for the trustworthiness of the BCC sites. Comparing the top and bottom graphs, one can see that the BCC is significantly more predictive of untrustworthy sites than the BFS periphery

thinking users may be tempted to conclude that, if a site points to "good" sites, it must be "good" itself.

2. THE BCC IS SIGNIFICANTLY MORE PREDICTIVE OF UNTRUSTWORTHY SITES THAN THE BFS PERIPHERY.

In particular (see Figure 4, top), in the BCC of an untrustworthy starting site, we found that, on average, 74% of the sites were also untrustworthy, while only 9% were trustworthy. In the BFS periphery (see Figure 4, bottom), these average percentages change to 27% untrustworthy and 11% trustworthy, with the rest non-determined. This suggests that the trustworthiness of sites in the BFS periphery is essentially unrelated to the trustworthiness of the starting site.

A. Future Directions: Incorporating Content Analysis

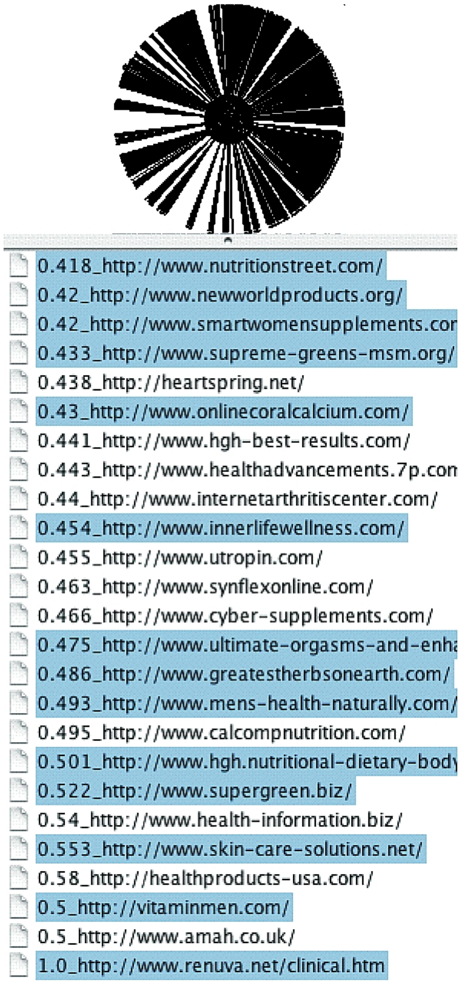
In our experiments we also devised a simple method to evaluate the similarity of the contents of each site to the starting site s. After the trust neighborhood was explored, we fetched and concatenated a few pages from each site (randomly choosing from the links that appeared in the domain URL) into a document. Then, we tried to determine the similarity of each such document to the document of the starting site. Similarity was determined using the $tf.idf$ ranking on the universe of the sites explored. We are aware that having a limited universe of documents does not give the best similarity results, but we wanted to get a feeling of whether our method could further be used to distinguish between "link farms" (spamming sites controlled by a single entity) and "mutual admiration societies" (groups of independent spammers choosing to exchange links). The initial results are encouraging, (see Fig. 5) showing a higher percentage of untrustworthy sites among those most similar to the starting site s.

Several possible extensions can be considered in this work. Generating graphs with more back-links per site, studying the evolution of trust neighborhoods over time, examining the density of the BCCs, and finding a more reliable way to compute similarity are some of them. We

S	$\|V_G\|$	$\|E_G\|$	$\|V_{BCC}\|$	$\|E_{BCC}\|$	**Trust**$_{BCC}$	**Untr**$_{BCC}$	**Trust**$_{BFS}$	**Untr**$_{BFS}$
U-1	1307	1544	228	465	2%	74%	31%	33%
U-2	1380	1716	266	593	4%	78%	32%	42%
U-3	875	985	97	189	0%	80%	39%	10%
U-4	457	509	63	115	0%	69%	37%	30%
U-5	716	807	105	189	0%	64%	23%	36%
U-6	312	850	228	763	9%	60%	38%	19%
U-7	81	191	32	143	0%	100%	30%	20%
U-8	1547	1849	200	430	5%	70%	40%	23%
T-1	1429	1566	164	273	56%	3%	57%	4%
T-2	241	247	13	17	77%	15%	27%	18%

Table III

SIZES OF THE EXPLORED TRUST NEIGHBORHOODS G AND THEIR BCC'S FOR EIGHT UNTRUSTWORTHY (U-1 TO U-8) AND TWO TRUSTWORTHY (T-1 AND T-2) STARTING SITES. $|V_G|$ CONTAINS THE NUMBER OF VERTICES AND $|E_G|$ THE NUMBER OF EDGES THAT OUR ALGORITHM FOUND IN THE TRUST NEIGHBORHOOD OF STARTING SITE s (STARTING FROM SITE s AND EXPLORING IN BFS MODE THEIR BACK-LINKS.) COLUMNS $|V_{BCC}|$ AND $|E_{BCC}|$ CONTAINS THE NUMBERS OF VERTICES AND EDGES OF THE LARGEST BICONNECTED COMPONENT WITHIN G. THE NEXT FOUR COLUMNS CONTAINS THE ESTIMATED PERCENTAGES OF TRUSTWORTHY AND UNTRUSTWORTHY SITES FOUND IN THE BCCS AND THE BFS PERIPHERIES (RESPECTIVELY). 20% OF EACH BCC AND 10% OF EACH BFS PERIPHERY WERE EVALUATED USING STRATIFIED SAMPLING.



0.418_http://www.nutritionstreet.com/
0.42_http://www.newworldproducts.org/
0.42_http://www.smartwomensupplements.cor
0.433_http://www.supreme-greens-msm.org/
0.438_http://heartspring.net/
0.43_http://www.onlinecoralcalcium.com/
0.441_http://www.hgh-best-results.com/
0.443_http://www.healthadvancements.7p.con
0.44_http://www.internetarthritiscenter.com/
0.454_http://www.innerlifewellness.com/
0.455_http://www.utropin.com/
0.463_http://www.synflexonline.com/
0.466_http://www.cyber-supplements.com/
0.475_http://www.ultimate-orgasms-and-enha
0.486_http://www.greatestherbsonearth.com/
0.493_http://www.mens-health-naturally.com/
0.495_http://www.calcompnutrition.com/
0.501_http://www.hgh.nutritional-dietary-bod
0.522_http://www.supergreen.biz/
0.54_http://www.health-information.biz/
0.553_http://www.skin-care-solutions.net/
0.58_http://healthproducts-usa.com/
0.5_http://vitaminmen.com/
0.5_http://www.amah.co.uk/
1.0_http://www.renuva.net/clinical.htm

Figure 5. The list of sites similar to the starting site U-1 (at the end of the list). The highlighted sites are those that participate in the BCC. The decimal number in front of the URL corresponds to its calculated content similarity to the starting site (which has similarity of 1.0 to itself).

also expect that the results would be strengthened if one considers tri- (or higher) connected components of the trust neighborhood. The Google API has been known to be filtering and restricting the number of the back-links it is reporting but it was the only tool available at the time of this research. Using the Yahoo Search API will likely improve the results we are getting.

V. Conclusions

In this paper we present a technique to identify spamming untrustworthy neighborhoods, developed by mimicking anti-propagandistic methods. In particular, we presented automatic ways of recognizing trust neighborhoods on the web based on the biconnected component around some starting site. Experimental results from a number of such instances show our algorithm's ability of recognizing parts of a spamming network. Even though it may not be possible to identify spamming sites solely through our algorithm, our work is complementary to the recent developments that recognize web spam based on link analysis [45], [2].

One of the benefits of our method is that we do not need to explore the web graph explicitly in order to find these neighborhoods, which would be impossible for a client computer. Of course, it would be possible to support a user's trusted and untrusted sites through some personalization service provided by search engines. To be usable and efficient, this service would require the appropriate user interface. For example, a search engine's Toolbar could have a "Web Spam" button similar to the "Spam" or "Junk" buttons that many email applications fashion these days. When a user encounters an untrustworthy site coming high up in the results of some search query, she would select the item and click on a "Distrust" button. The browser would add this site in the user's untrustworthy site collection and would run the algorithm that propagates distrust backwards. Next time the user runs a similar search query, the untrusted sites would be blocked or demoted.

Recently, Google has introduced SearchWiki, a method of supporting personalized opinions about search results [41], which could be adjusted to support this operation. We view this development as justified by our findings and, even though we do not know whether Google's decision to employ this tool was partially influenced by our results, we do think it is a step in the right direction.

The algorithm we described is a first step in supporting the trust network of a user. Ultimately, it would be used along with a set of trust certificates that contains the portable trust preferences of the user, a set of preferences that the user can accumulate over time. Organizations that the user joins and trusts may also add to this set. A combination of search engines capable of providing indexed content and structure [19], including identified neighborhoods, with personalized filtering those neighborhoods through the user's trust preferences, would provide a new level of reliability to the user's information gathering. Sharing ranking decisions with the end user will make it much harder for spammers to tune to a single metric – at least as hard as it is for propagandists to reach a large audience with a single trick.

ACKNOWLEDGMENT

The author's research was partially funded by a Brachman-Hoffman Fellowship.

The author would like to thank Joe DeStefano, Mirena Chausheva, Meredith Beaton-Lacoste, Scott Anderson and Scott Dynes for their valuable contributions. We would also like to thank David "Pablo" Cohn for his many useful suggestions that improved the presentation of the paper. The graphs shown in this paper were drawn using the yEd package [46].

REFERENCES

[1] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technology*, 1(1):2–43, June 2001.

[2] A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. Spam Rank – Fully automatic link spam detection. In *Proceedings of the AIRWeb Workshop*, May 2005.

[3] K. Bharat, B.-W. Chang, M. R. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 51–58. IEEE Computer Society, 2001.

[4] M. Bianchini, M. Gori, and F. Scarselli. PageRank and web communities. In *Web Intelligence Conference 2003*, Oct. 2003.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[6] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Comput. Networks*, 33(1-6):309–320, 2000.

[8] J. Cho and S. Roy. Impact of search engines on page popularity. In *Proceedings of the thirteenth international conference on World Wide Web*, May 2004.

[9] CNETNews. Engine sells results, draws fire. http://news.cnet.com/2100-1023-215491.html, June 21 1996.

[10] T. S. Corey. Catching on-line traders in a web of lies: The perils of internet stock fraud. Ford Marrin Esposito, Witmeyer & Glesser, LLP, May 2001. http://www.fmew.com/archive/lies/.

[11] G. Cybenko, A. Giani, and P. Thompson. Cognitive hacking: A battle for the mind. *Computer*, 35(8):50–56, 2002.

[12] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics. In *WebDB2004*, June 2004.

[13] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the twelfth international conference on World Wide Web*, pages 669–678. ACM Press, 2003.

[14] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization of the web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002.

[15] L. Graham and P. T. Metaxas. "Of course it's true; i saw it on the internet!": Critical thinking in the internet era. *Commun. ACM*, 46(5):70–75, 2003.

[16] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proceedings of the AIRWeb Workshop*, May 2005.

[17] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *VLDB 2004*, Aug. 2004.

[18] S. Hansell. Google keeps tweaking its search engine. New York Times, Jun. 3 2007.

[19] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the eleventh international conference on World Wide Web*, pages 517–526. ACM Press, 2002.

[20] M. R. Henzinger. Hyperlink analysis for the web. *IEEE Internet Computing*, 5(1):45–50, 2001.

[21] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[22] M. Hindman, K. Tsioutsiouliklis, and J. Johnson. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *Annual Meeting of the Midwest Political Science Association*, April 3-6 2003.

[23] L. Introna and H. Nissenbaum. Defining the web: The politics of search engines. *Computer*, 33(1):54–62, 2000.

[24] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the twelfth international conference on World Wide Web*, pages 271–279. ACM Press, 2003.

[25] JUNG. The JUNG framework developer team – release 1.5. http://jung.sourceforge.net/.

[26] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM Press, 2000.

[27] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[28] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1481–1493, 1999.

[29] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web and social networks. *IEEE Computer*, 35(11):32–36, 2002.

[30] A. M. Lee and E. B. Lee(eds.). *The Fine Art of Propaganda*. The Institute for Propaganda Analysis. Harcourt, Brace and Co., 1939.

[31] C. A. Lynch. When documents deceive: trust and provenance as new factors for information retrieval in a tangled web. *J. Am. Soc. Inf. Sci. Technol.*, 52(1):12–17, 2001.

[32] M. Marchiori. The quest for correct information on the web: hyper search engines. *Comput. Netw. ISDN Syst.*, 29(8-13):1225–1235, 1997.

[33] M. L. Maulding. Lycos: Design choices in an internet search service. *IEEE Expert*, January-February(12):8–11, 1997.

[34] P. Metaxas. On the evolution of search engine rankings. In *Proceedings of the 5th WEBIST Conference*, Lisbon, Portugal, March 2009.

[35] P. Metaxas. Using propagation of distrust to find untrustworthy web neighborhoods. In *Proceedings of the 4th International Conference on Internet and Web Applications and Services* (ICIW 2009), Venice, Italy, May 2009.

[36] A. Ntoulas, D. Fetterly, M. Manasse, and M. Najork. Detecting spam web pages through content analysis. In *World-Wide Web 2006*, May 2006.

[37] G. Pringle, L. Allison, and D. L. Dowe. What is a tall poppy among web pages? In *Proceedings of the seventh international conference on World Wide Web 7*, pages 369–377. Elsevier Science Publishers B. V., 1998.

[38] P. Raghavan. Social networks: From the web to the enterprise. *IEEE Internet Computing*, 6(1):91–94, 2002.

[39] G. Salton. Dynamic document processing. *Commun. ACM*, 15(7):658–668, 1972.

[40] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[41] The official Google blog. SearchWiki: Make search your own. http://googleblog.blogspot.com/2008/11/searchwiki-make-search-your-own.html, Nov. 20 2008.

[42] A. Vedder. Medical data, new information technologies and the need for normative principles other than privacy rules. In *Law and Medicine. M. Freeman and A. Lewis (Eds.), (Series Current Legal Issues)*, pages 441–459. Oxford University Press, 2000.

[43] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[44] D. Welch. Power of persuasion - propaganda. *History Today*, 49(8):24–26, 1999.

[45] B. Wu and B. Davison. Identifying link farm spam pages. In *Proceedings of the fourteenth international conference on World Wide Web*, May 2005.

[46] yWorks. yEd – java graph editor, v. 2.2.1. http://www.yworks.com/en/products_yed_about.htm.

Assuring Quality in Vulnerability Reports for Security Risk Analysis

Deepak Subramanian, Ha Thanh, Le and Peter, Kok Keong, Loh

School of Computer Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
Email: DEEP0018@ntu.edu.sg

***Abstract*-Web application scanners detect and provide some diagnoses for specific vulnerabilities. However, scanner performance as well as the damage potential of different vulnerabilities varies. This undermines the development of effective remediation solutions and the reliable sharing of vulnerability information. This paper describes the development of fuzzy classification metrics that are used to grade web application scanners and vulnerabilities so that scanner performance can be evaluated and confidence levels can be computed for vulnerability reports. These metrics help derive a level of assurance that will support security management decisions as well as enhance effective remediation efforts.**

Keywords
Fuzzy classifiers, confidence level, calibration, scanner, vulnerability, web application

I. Background and Motivation

Contemporary detection of software vulnerabilities in web-based systems is accomplished via web application scanners [21, 25]. However, depending on the capabilities of these scanners, different vulnerability reports generated will have a widely varying level of trustworthiness. This raises critical concerns especially when these reports are used to estimate system risks for management decisions and the development of remediation processes.

Risk analysis is inherently a complex process fraught with ambiguity and uncertainty. Traditional risk approaches are usually based on assumptions of known vulnerabilities or threats and are thus not suitable for contemporary web services and applications that exhibit a degree of platform inter-operability and dynamic content. In different web applications, some vulnerabilities are also more dangerous than others in terms of potential damage/risks [1, ICIMP 2009]. These issues create a challenge to develop a quality assurance mechanism for scanner generated reports. Qualified reports can then support a trusted level of analysis of system risk as well as being a more dependable resource of shared system security information.

Our assurance mechanism described in this paper focuses on supporting more reliable risk analysis of web-based systems and is based on fuzzy metrics that are used to calibrate scanner performance as well as vulnerabilities. Our approach also forms part of a system framework that achieves standardization of scanner reports across different web technologies [2-4]. The standardization is necessary since there has been a rise in the number and type of scanners and vulnerabilities. Scanner algorithms evolve as the vulnerabilities evolve. For any organization that had a greater requirement for security, it would be more advisable to rely on several algorithms instead of just one since these algorithms perform differently in different scenarios. The scanner results can then be collated to give a more significant output in a standardized format. Our approach would benefit the Security Administration and Audit groups of an organization or enterprise in ensuring scalable security enforcement and compliance against an unpredictable vulnerability backdrop.

In this paper, our approach utilizes the standardized scanner results generated, together with scanner performance and vulnerability calibrations, to compute associated confidence levels with these results.

The rest of the paper is organized as follows. Section II presents a review of existing research. Section III sums up the research issues based on the review and describes the requirements. Section IV details the design of the quality assurance metrics while section V presents the design of the scanner and vulnerability grading systems. Section VI exemplifies the calculation of the 1st and 2nd degree confidence levels for vulnerability reports while Section VII provides an illustrative example to describe the working of the framework. Section VIII concludes the paper followed by the references.

II. Review of Existing Research

It is difficult for decision makers to identify entire network threats and collect precise and adequate data to estimate all probable risks due to vulnerabilities or threats. Furthermore, risk analysis for web service security and applications is not only limited to determining recognized web threats, but should also estimate potential risks. A review of the more recent works that have had some influence in this research is described in this section.

In [14] an extended form of the *Pseudo-Order Preference Model (POPM)* was used to estimate the imprecise risk of web services based on richness of information and to determine their ranking using a weighted additive rule. A fuzzy logic based approach was used to calculate information characteristics provided by the web service. There are 3 models used in this process a) Pseudo-order preference model, b) Semi-order preference model and c) Complete-preorder preference model. Each model is executed if and only if a condition is reached. Each model was given an Outranking relation which the research states would affect the decision making capabilities for the risk analysis. The decision makers have been stated as useful parameters in helping the experts making their decisions. The future work in this model includes the selection of an appropriate threshold for the preference relation, defining an appropriate threshold for the Indifference preference. The use of fuzzy logic by this research, however, does not extend to measure differing security tool performance and vulnerability severity.

In the research [15], the software code has been taken and analyzed for security patterns. The paper shows the results of experimenting with J2EE code with a MySQL back-end and JBoss Application Server. The various software security patterns of Intercepting Validator, Guard of Secure Proxy with Secure Pipe, Container Managed Security and Secure Logger have been analyzed with how implementation of each affects the vulnerability being used on the system. A fuzzy approach has been used by having a linguistic value for every generated fuzzy equivalent range such as low, medium, high, very high etc., This has been used to analyze the effectiveness of the various security patterns. The effectiveness of the patterns against primary attack evens i.e. events that lead to execution of an attack has also been analyzed. While each pattern has a varying effectiveness in varying scenarios, the code with security patterns implemented has been proven to always be more secure to the ones that are not following the patterns. Future work of the project involves the creation of newer patterns that have not been mentioned above. This approach uses a whitebox methodology to test for code patterns that have a less likelihood of getting affected by certain vulnerabilities. Our approach compliments this research by ensuring if these code patterns have been designed and implemented securely.

The research [16] describes the need to prevent or manage the damage caused by security threats. The research describes that the web-server based applications must be made in such a way that they incorporate the ability to self-heal after an attack. The authors describe how this can be made possible by the basis of data obtained from anomaly detection. The anomaly detection data is then processed using a Discrete Finite Automate (DFA) to detect malicious web requests. An anomaly based detection combined with DFA needs to be trained in the beginning to find which anomaly detection data matches a true positive attack and the patterns of such requests are observed by the training algorithm. The patterns are then detected after the training and such requests are suggested to be blocked or sent to a more secure, but less functional server to protect the found and restored. This approach is suitable for systems that are holding highly critical data that cannot be changed but would need extensive training and validation and could be expensive to implement. It needs to be implemented at every server. It is not a preventive technique but a criterion for recovery when an attack is observed. Our approach, on the other hand, is a part of a framework that would be able to provide remediation based on the observed attack but not an automated recovery.

The research in [12] proposed a method for identifying and charting software exposure to un-patched vulnerabilities. Disclosed vulnerabilities are divided into 2 types. The first comprises of vulnerabilities that are publicly known with no patch available from the vendor. The second comprises of vulnerabilities that are publicly known with a patch available from the vendor. By calculating the Daily Vulnerability Exposure (DVE) for all un-patched vulnerabilities for a continuous period of time, an exposure chart is obtained. Using the chart's help, it is possible to ascertain how long a vendor takes to patch and if the patch is effective, by calculating the DVE after the patch date. The exposure chart could also be used to calculate the severity metrics used by the National Vulnerability Database (NVD). The DVE is a severity metric that is based on how much the vulnerability is graded in terms of time elapsed, from the date it is discovered till the date a patch is available. The vulnerabilities handled here are generally in the new and Relatively new categories of our approach. These have been described in the section V.

The research in [13] is oriented towards the quantitative characterization of the vulnerabilities in operating systems. A time-based model for the total vulnerabilities discovered is proposed and fitted to the data for Windows 98 and Windows NT 4.0. Being a time-based model, it is able to obtain an indication of the expectancy of the vulnerability being targeted based on the phase the system is in. An alternative effort-based vulnerability model analogous to software reliability growth models was also proposed. Both models fit well and the fit is significant, however, further development is necessary before confidence levels associated with the detection of vulnerabilities can be assessed.

III. Research Issues and Requirements

While data classification can act as an enabler for a more effective diagnosis and calculation of DVE helps in determining patch effectiveness [12], the widely varying detection capabilities encountered during scanning as well as the differing threat / risk levels posed by individual vulnerabilities have not been addressed.

The approaches [10][5][11] have been influential in validating the use of fuzzy logic in classification. The use of a neural network is an effective method in developing an inference engine. However, it needs to a lot of training data and this data can influence the working in a very significant way. By using suitable heuristics instead gives more stability to the system against any misclassification errors and also reduces any complexity that could be faced while training a neural network.

The scanner output could be right or wrong depending upon the algorithms used by the scanner and the database supported by the scanner. From the intrusion detection research stated above [5][11] it can be observed that the intrusion detection models also use several tools to first identify the various suspicious events and then have some decision making processes to deal with such observed events based on fuzzy diagnosis. This research also adopts such an approach to deal with the web vulnerabilities discovered by the scanners. However, this is the only similarity between the approaches in this research and the above stated works [10][5][11].

In our research, we address the variable detection capability of scanners and different threat / risk levels posed by individual vulnerabilities. Our approach grades web vulnerabilities and scanners quantitatively via expressions based on fuzzy truth values. The requirements of our approach are stated as follows:

1. Metrics for vulnerabilities and scanner performance may be calibrated empirically prior to analysis making this a more practical and flexible methodology.

2. Web vulnerabilities and scanner performance may be classified and ranked in a reliable and informative way.

3. Scales easily to cover new vulnerabilities, vulnerability variants and scanners.

4. Supports more effective management and remediation decisions and facilitates occurrence estimation of classified vulnerability.

IV. Preliminaries

In this section, we define and explain the terminologies that will be used in the rest of the paper.

Scanners are applications that use suitable algorithms to detect web vulnerabilities.
Fuzzy sets are sets whose elements have degrees of membership. An element mapping to the value 0 means that the member is not included in the fuzzy set, while a mapping to the universe of disclose, where the universe of disclose represents the entire set of members possible and the fuzzy sets they belong to. A diagrammatic representation is given in Figure 1. A value of 1 describes a fully included member. Mapped values strictly between 0 and 1 characterize the fuzzy members.

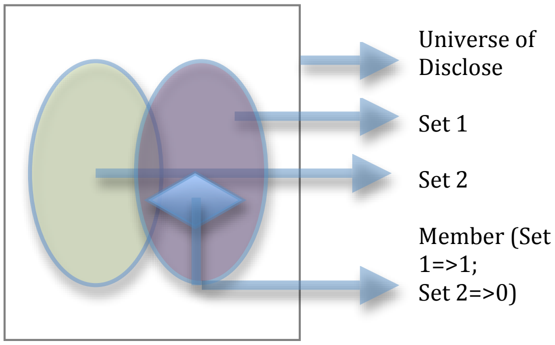


Figure 1. Venn Diagram of a fuzzy set

A *positive vulnerability* refers to a vulnerability that is present in the website at a specified instant and there is evidence to support it.

A *negative vulnerability* refers to a vulnerability that is not present in the website and can be proved to a satisfactory level.

The *calibration phase* is the time period during which fuzzy metrics are calibrated before the scanner is ready to generate vulnerability reports.

An *instance* of a vulnerability detected present (absent) by a scanner for a given website can be defined as that manifestation (or non-manifestation) of the vulnerability that occurs during a specified period of time where there has been no change in scanner algorithm, scanned website or vulnerability definition.

Test websites are those websites that have been custom designed to contain or not contain specified vulnerabilities for the purpose of testing the scanners. The test websites are used mainly in the calibration phase.

Ground truth is the true value of whether the vulnerability is present or absent. It is an absolute value i.e. the vulnerability is present or it is absent.

The *Likelihood Ratio* is the ratio of the probability that a particular vulnerability would be predicted when it matches the ground truth to the probability that it would be predicted erroneously.

Sensitivity is the proportion of correct detections of vulnerability presence out of all true instances of a particular scanner's detection. It can be computed in both a vulnerability specific way as well as in a scanner specific way. When calculated in a scanner specific way, it is averaged over all the vulnerabilities. It corresponds to the correct detection rate relative to ground truth.

Specificity is the proportion of false detections of vulnerability presence out of all false instances of a particular scanner's detection. It can be computed in both a vulnerability specific way as well as in a scanner specific way. When computed in a scanner specific way, it is averaged over all specified vulnerabilities.

Cross-site request forgery (CSRF) is an attack which forces an end user to execute unwanted actions on a web application in which the end user is currently authenticated.

Cross-site scripting (XSS) attacks occur when an attacker uses a web application to send malicious code, generally in the form of a browser side script, to a different end user.

V. Design of Classification Metrics

In our proposed design (Figure 2), *calibration* forms an important and integral part of the framework. The calibration process makes use of two grading systems: *scanner grading system* and *vulnerability grading system*. Scanner and vulnerability metrics are first calibrated by the respective grading systems before any confidence levels and diagnostics are computed.

Grading can increase the reliability of reports obtained by allowing for their evaluation based on the grades of the various scanners that detected the specified vulnerability and threat/risk posed by the vulnerability. Grading of scanners and vulnerabilities are computed based on *scanner specific truth-values* and *vulnerability specific truth-values*, respectively. Scanner specific and vulnerability specific truth-values form fuzzy sets. The assurance of a scanner based on the vulnerability it is able to detect can provide an assurance of quality in the vulnerability reports provided by various scanners.

Using the grading results, low performance scanners can be selectively upgraded or omitted and vulnerabilities with high damage potential can be identified and affected systems isolated. Additionally, computation of report confidence can also be carried out. For example, a low confidence level obtained while the vulnerability is detected would mean that there is a low likelihood that the vulnerability is actually present. On the other hand, a high confidence level implies that there is a high probability that the detected vulnerability will not be a false positive. The confidence level thus obtained is an assurance of the risk analysis for the various vulnerabilities that has been done on the particular website location.

In the next few sub-sections, we detail the development of the framework design based on the requirements and from the assertions made. The assertions form an integral part of the system that provide a basis for the subsequently proposed metrics.

A. Assertions made

Assertion 1:
Some vulnerabilities are more difficult to exploit than others.

Not all vulnerabilities are equally susceptible to exploitation and the potential damage that can be caused will also not be the same. Hence, a *vulnerability grade system* needs to be present to provide a better diagnosis of the various vulnerabilities that are detected by the web application scanners.

Assertion 2:
Web-based vulnerabilities can be classified into 4 types, namely:

i. Evolved Vulnerability:
If there is recorded detection for the vulnerability and there also exists at least one recorded exploitation method that is still usable.

ii. Dormant Vulnerability:
If all recorded exploitation methodologies can no longer be used and there is at least one recorded exploitation method.

iii. Relatively-new Vulnerability:
If there is no recorded detection but there exists at least one recorded exploitation method.

iv. New vulnerability:
If there is no recorded detection or exploitation method for the vulnerability.

By classifying vulnerabilities into specific types, it is possible to evaluate the capability of scanners as well in a better way. For example, a less sophisticated scanner would not be expected to find a new or even relatively-new vulnerability. This would provide a credibility rating for the scanner with specific scanner metrics defined in the sections to follow. The severity of the vulnerability can also be ascertained to a certain degree with this approach. For example, a dormant vulnerability resulting from a series of successful patches will have a lower severity than an evolved one.

Assertion 3:
The difficulty of detection of an evolved vulnerability is directly proportional to the difficulty of exploiting it.

The above assertion is influenced intuitively by the notion that if a complex algorithm and/or extended process were needed to detect the vulnerability, a proportionate effort would be required in effectively exploiting it. In other words, if the vulnerability can be easily detected or observable then the skill level / effort needed for exploitation is correspondingly less. Given the above assertions and the fact that not all scanners will be able to deal with a particular vulnerability with the same degree of effectiveness, scanning capability must be graded. The capability of a scanner to effectively detect a vulnerability is represented by an index allocated to it known as S_{GRADE} (*Scanner Grade*). A scanner with a higher S_{GRADE} is then better suited to detect the vulnerability than one with a lower S_{GRADE}.

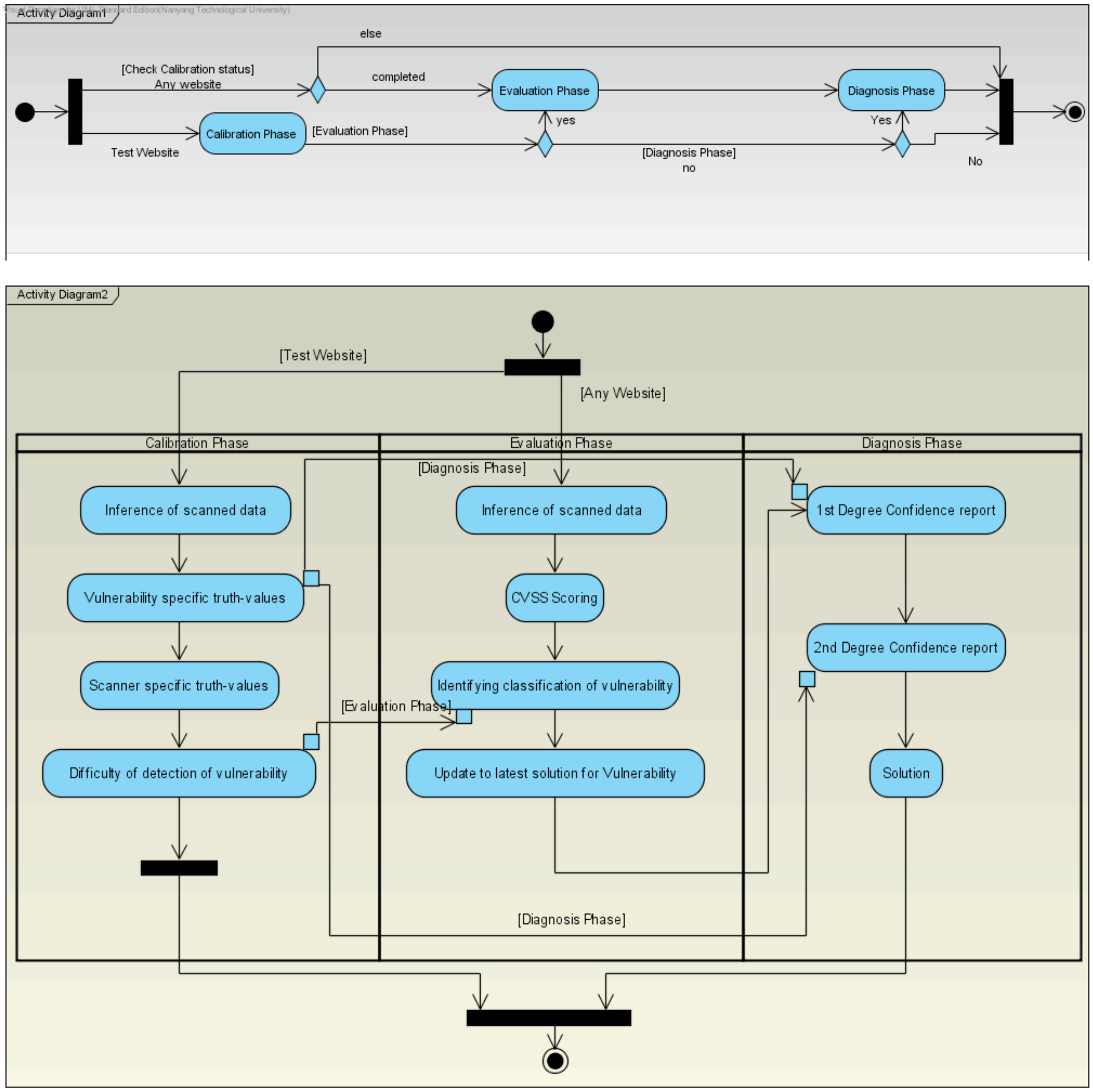

Activity Diagram1
else
[Check Calibration status]
Any website
completed
Evaluation Phase
Diagnosis Phase
Test Website
Calibration Phase
[Evaluation Phase]
yes
Yes
[Diagnosis Phase]
no
No
Activity Diagram2
[Test Website]
[Any Website]
Calibration Phase
Evaluation Phase
Diagnosis Phase
[Diagnosis Phase]
Inference of scanned data
Inference of scanned data
1st Degree Confidence report
Vulnerability specific truth-values
CVSS Scoring
2nd Degree Confidence report
Scanner specific truth-values
Identifying classification of vulnerability
[Evaluation Phase]
Difficulty of detection of vulnerability
Update to latest solution for Vulnerability
Solution
[Diagnosis Phase]

Figure 2: Overview of Framework Design

Assertion 4:
The impact level of a vulnerability (vulnerability-exploitability measure) is likely to vary with varying instances of detection and at varying locations.

It has been stated in assertion 1 as to how some vulnerabilities may be more difficult to exploit than others. It is also true that the same vulnerability may be present in a more exploitable location than others. Consider cross-site scripting at a site where the vulnerability may be exploitable directly. At a later time, perhaps after changes in site architecture, the same vulnerability may be exploitable only after a series of authorization pages. Hence, the exploitability of the cross-site scripting has changed for the same website at different instances of detection. Similarly, the exploitability of cross-site scripting may also vary due to platform differences at various sites.

Assertion 5:
Existence of one vulnerability may influence the prevalence of another.

Steps are usually taken to prevent some vulnerabilities in a system. Some other vulnerabilities may, however, still exist. These other vulnerabilities may directly or indirectly enable the avoided vulnerabilities to bypass the previous prevention schemes. An illustrative example is that of *cross-site request forgery* (CSRF) and *cross-site scripting* (XSS) [19]. If cross-site request forgery has been avoided by non-usage of JavaScript and secret tokens at each level (which is quite an effective methodology), the website is virtually protected from CSRF and typical scanners will also declare the same. However, if XSS has not been avoided, it can be used to get the tokens ahead of time and a hybrid use of CSRF can be realized which cannot be detected by the scanners.

Hence, we may define $R_{V1}(V2)$, where $R_{V1}(V2)$ is the likelihood of occurrence of vulnerability V1 when vulnerability V2 has occurred.

B. Vulnerability Specific Truth-Values

The various vulnerabilities have their own levels of difficulty as defined by the assertion 3. It is therefore a necessity to ascertain how scanners react to the various vulnerabilities. The scanners themselves need to be graded as well, which is described in the section V.C. After the computation of the vulnerability specific truth values and scanner specific truth values, the difficulty of detection of the vulnerability can be ascertained. The difficulty of detection is a useful estimation that can help in ascertaining the importance of detection thus ensuring a quality-based analysis of scanner detections.

Let the vulnerability specific truth values for vulnerability *j* be represented by {$V(j)_{TP}$, $V(j)_{TN}$, $V(j)_{FP}$& $V(j)_{FN}$}. These are also known as {True Positive, True Negative, False Positive and False Negative}, respectively. These truth values can be used to derive fuzzy metrics that would be useful in dealing with discrepancies among different scanners detecting vulnerabilities in the system. The vulnerability with a higher TP value can be used to determine the certainty of the vulnerability being present and a high TN value indicates a high probability that the vulnerability is absent.

The fuzzy classifiers are defined as,

$$VI(j)_T = \sum_{i=1}^{WS_T} I_{ij} \qquad \textbf{(1)}$$

$$VI_F(j)_T = \sum_{i=1}^{WS_T} F_{ij} \qquad \textbf{(2)}$$

$$V(j)_{TP} = \frac{\sum_{k=1}^{WS_T} \sum_{i=1}^{I_{kj}} d(v_{ij})}{VI(j)_T} = \frac{\sum_{i=1}^{VI(j)_T} d(v_{ij})}{VI(j)_T} \qquad \textbf{(3)}$$

$$V(j)_{FP} = \frac{\sum_{k=1}^{WS_T} \sum_{i=1}^{F_{kj}} d(v_{ij})}{VI_F(j)_T} = \frac{\sum_{i=1}^{VI_F(j)_T} d(v_{ij})}{VI_F(j)_T} \qquad \textbf{(4)}$$

$$V(j)_{TN} = 1 - V(j)_{FP} \qquad \textbf{(5)}$$

$$V(j)_{FN} = 1 - V(j)_{TP} \qquad \textbf{(6)}$$

Where,
$D\ (v_{ij})$={1 if instance of vulnerability j is detected at location *i* or
0 if instance of vulnerability *j* is not detected at location *i* }
I_{ij} is the number of instances of vulnerability *j* present at location *i* during calibration
F_{ij} is the number of instances of vulnerability *j* falsely detected at location *i* during calibration
$VI(j)_T$ Is the total number of instances of vulnerability j used for calibration
$VI_F(j)_T$ is the total number of instances of vulnerability *j* falsely detected during calibration

WS_T is the total number of test websites used for calibration

$V(j)_{TP}, V(j)_{TN}, V(j)_{FP} \& V(j)_{FN}$ are the vulnerability specific truth values which refer to the vulnerability specific true positive, true negative, false positive and false negative, respectively.

C. Scanner Specific Truth-Values

Different scanner's output data differ in content, format and organization [1, ICIMP 2009]. The data generated by scanners depends, to an extent, on the algorithm being used in the specific scanner. Some scanners with access to large databases are equipped to detect more classes of vulnerabilities. Others comprising lightweight computational modules provide basic diagnoses while several lie somewhere in between. For example, IBM AppScan and HP WebInspect are scanners with access to large databases while NStalker is associated with a relatively smaller database. It is hence necessary to first analyze and understand the scanning process as well as the capability of the scanner in order to derive the required metrics.

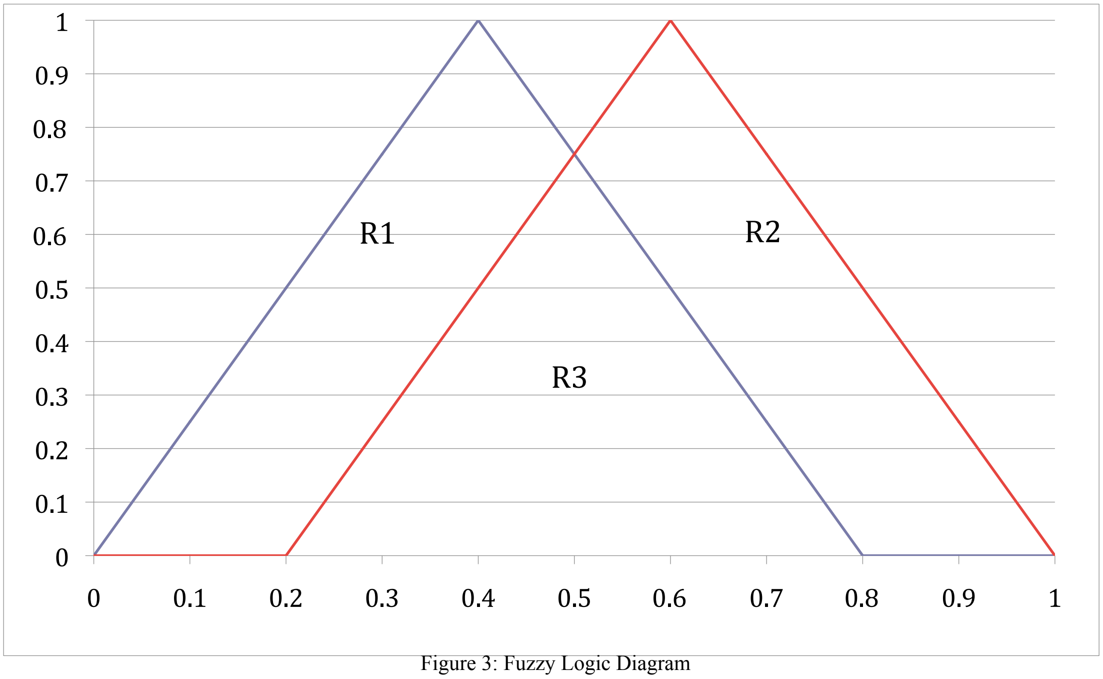


Figure 3: Fuzzy Logic Diagram

Let the scanner specific truth-values be represented by { S_{TP}, S_{TN}, S_{FP}, S_{FN} }. These are also known as {True Positive, True Negative, False Positive and False Negative}, respectively. These values form an important measure of the vulnerability detection capability of the scanner. Derived scanner metrics require calibration in order to grade the scanner prior to its usage for adequate quality assurance based on performance. Calibration of the detection capabilities of the scanner is performed with a sample of customized websites for positive or negative vulnerabilities thus reflecting the performance of a scanner with expected results providing a valid basis for a suitable quality check.

In the pre-calibration phase, the system would be unlikely to produce results with the levels of reliability expected by the user. The scanner metrics are defined with respect to a scanner's prediction capabilities. The prediction capabilities of the scanner are then calibrated against the expected prediction performance. Vulnerabilities are also calibrated in this phase and their fuzzy metrics are defined in section V (C.). It must be noted that the classification of vulnerabilities in the calibration phase influences the scoring by CVSS [7]. Once the *calibration phase* for the scanner is completed, the reports from the scanner can be processed for data in a more reliable manner. The fuzzy metrics for scanners are defined as follows:

$$U_T = \sum_{j=1}^{EV_T} VI(j)_T \qquad \textbf{(7)}$$

$$U_F = \sum_{j=1}^{EV_T} \sum_{i=1}^{WS_T} U(F_{ij}) \qquad \textbf{(8)}$$

$$S_{TP} = \frac{\sum_{j=1}^{U_T} V(j)_{TP}}{U_T} \tag{9}$$

$$S_{TN} = \frac{\sum_{j=1}^{U_F} V(j)_{TN}}{U_F} \tag{10}$$

$$S_{FP} = 1 - S_{TN} \tag{11}$$

$$S_{FN} = 1 - S_{TP} \tag{12}$$

Where,

U_T is the total number of unique vulnerabilities incorporated for calibration.

U_F is the total number of unique vulnerabilities falsely detected by scanners.

$U(F_{ij})$={1, if the instance is unique for that vulnerability || 0, otherwise}

EV_T is the total number of vulnerability instances evaluated

$V(j)_{TP}$ and $V(j)_{TN}$ are vulnerability specific truth values defined in section V.B

$VI(j)_T$ is the total number of instances of vulnerability j used for calibration

S_{TP}, S_{TN}, S_{FP}&S_{FN} are the scanner specific truth values.

Figure 3 shows the graphical representation of the fuzzy metrics for the scanner. A similar diagram can also be used for vulnerability fuzzy metrics. The region R1 represents the true positive S_{TP}, the region R_2 represents the true negative S_{TN} and the region R_3 is the combined space of false positive and false negative S_{FP} and S_{FN}.

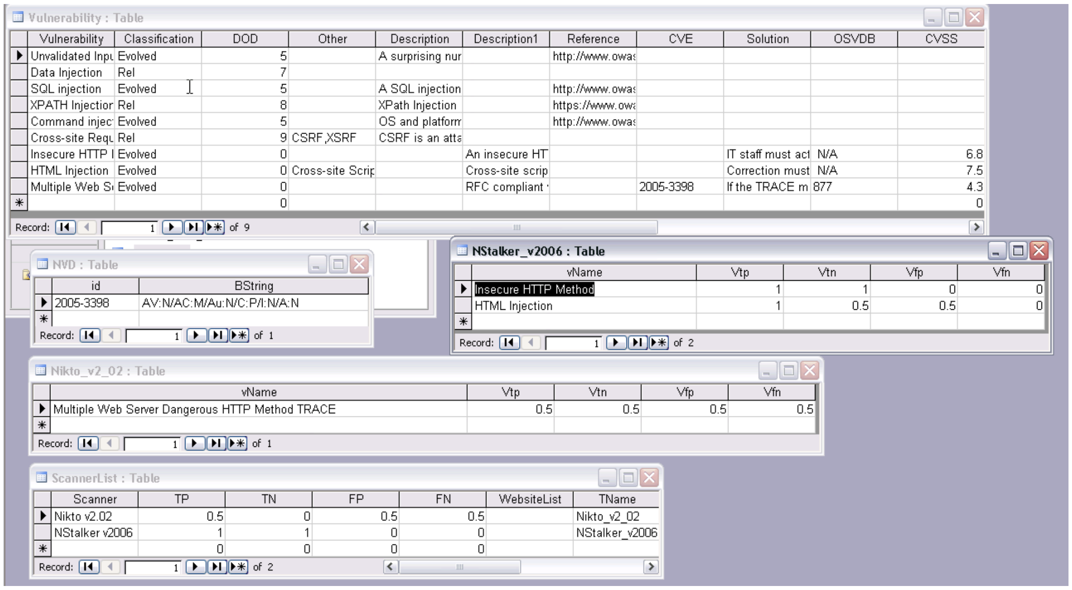
VI. Grading Systems

This section describes the design details of the two grading systems: the scanner grading system and the vulnerability grading system.

A. Scanner Grading System

The scanner grading system is used to grade the capability of a web application scanner. The scanner grading system makes use of a scanner database as well as the vulnerability databases [17]. The scanner database comprises a table list that is maintained for every graded scanner (see Figure 4). It contains information on the scanner specific truth values as well as the vulnerability specific truth values. The Scanner Grade, S_{GRADE}, may be computed for all web-based vulnerabilities listed in the vulnerability database.

The overall *sensitivity* and *specificity* of the scanners can be computed by using the equations, Eqn. 13 and Eqn. 14. Sensitivity is the percentage of correctly detected activities out of all true instances of a particular class, averaged over



Vulnerability : Table

	Vulnerability	Classification	DOD	Other	Description	Description1	Reference	CVE	Solution	OSVDB	CVSS
▶	Unvalidated Inpu	Evolved	5		A surprising nur		http://www.owas				
	Data Injection	Rel	7								
	SQL injection	Evolved	5		A SQL injection		http://www.owas				
	XPATH Injectior	Rel	8		XPath Injection		https://www.ow				
	Command injec	Evolved	5		OS and platforr		http://www.owas				
	Cross-site Requ	Rel	9	CSRF,XSRF	CSRF is an atta						
	Insecure HTTP I	Evolved	0			An insecure HT			IT staff must act	N/A	6.8
	HTML Injection	Evolved	0	Cross-site Scrip		Cross-site scrip			Correction must	N/A	7.5
	Multiple Web S	Evolved	0			RFC compliant		2005-3398	If the TRACE m	877	4.3
✱			0								0

NVD : Table

	id	BString
▶	2005-3398	AV:N/AC:M/Au:N/C:P/I:N/A:N
✱		

NStalker_v2006 : Table

	vName	Vtp	Vtn	Vfp	Vfn
▶	Insecure HTTP Method	1	1	0	0
	HTML Injection	1	0.5	0.5	0
✱					

Nikto_v2_02 : Table

	vName	Vtp	Vtn	Vfp	Vfn
▶	Multiple Web Server Dangerous HTTP Method TRACE	0.5	0.5	0.5	0.5
✱					

ScannerList : Table

	Scanner	TP	TN	FP	FN	WebsiteList	TName
▶	Nikto v2.02	0.5	0	0.5	0.5		Nikto_v2_02
	NStalker v2006	1	1	0	0		NStalker_v2006
✱		0	0	0	0		

Figure 4: Table List in Scanner Grading System

all activities. Specificity measures the proportion of correctly identified negative occurrences to all true negative occurrences. If a scanner is more sensitive, it has a greater chance of discovering the vulnerability. Similarly, if a scanner is more specific, it has a greater chance of discovering the absence of the vulnerability.

$$S_{Specificity} = \frac{S_{TP}}{S_{TP} + S_{FN}} \qquad \textbf{(13)}$$

$$S_{Specificity} = \frac{S_{TN}}{S_{TN} + S_{FP}} \qquad \textbf{(14)}$$

Where,
S_{TP}, S_{TN}, S_{FP}&S_{FN} are the scanner specific truth values defined in section 5.3
$S_{ensitivity}$ is the sensitivity measure of the scanner
$S_{Specificity}$ is the specificity measure of the scanner
However, a greater sensitivity could also mean greater probability of false positives for the scanner. Similarly, a higher specificity could mean there are a greater number of false negatives for the scanner. For a given scanner, the trade-off between sensitivity and specificity depends on the vulnerability and the web application being scanned.

B. Vulnerability Grading System:

As mentioned previously, the representation for the vulnerability specific truth values is also similar to Figure 3. The *vulnerability specific sensitivity* and *specificity* for a scanner are defined by Eqns. 15 and 16, respectively. We also define the *likelihood ratio* for both true positive and true negative results with the Eqns. 17 and 18, respectively.

$$V_{Sensitivity} = \frac{V_{TP}}{V_{TP} + V_{FN}} \qquad \textbf{(15)}$$

$$V_{Specifi\ city} = \frac{V_{TN}}{V_{TN} + V_{FP}} \qquad \textbf{(16)}$$

$$V_{LR+} = \frac{V_{TP}(V_{TN} + V_{FP})}{V_{FP}(V_{TP} + V_{FN})} \qquad \textbf{(17)}$$

$$V_{LR-} = \frac{V_{FN}(V_{TN} + V_{FP})}{V_{TN}(V_{TP} + V_{FN})} \qquad \textbf{(18)}$$

Where,
$V_{Sensitivity}$is the vulnerability specific sensitivity measure for the scanner
$V_{Specificity}$is the vulnerability specific specificity measure for the scanner
$V_{LR+}$is the likelihood ratio for positive detection
V_{LR-} is the likelihood ratio for negative detection

V_{LR+} in Eqn. 17 gives the likelihood ratio of the vulnerability to be present given the vulnerability specific truth values for the specified scanner. V_{LR-} in Eqn. 18 gives the likelihood ratio of the vulnerability to be absent given the vulnerability specific truth-values for the specified scanner. Combined with the scanner metrics, these could be used as a basis in predicting the levels of vulnerability present. However, this will hold absolutely true only for vulnerabilities that fall under the evolved-vulnerability category.

The vulnerability specific sensitivity and the specificity metrics can also be used to study the scanner's performance and behavioral characteristics with certain classes of vulnerabilities. If the scanner is more sensitive towards a specific vulnerability, it will exhibit better detection of the presence of that particular vulnerability and if it is more specific, it will be more able to detect the absence of the particular vulnerability.

A *positive vulnerability* means that the vulnerability is present in the website location at that instant and there is evidence to support it. A *negative vulnerability* means that the vulnerability is not present in the website and can be proved to a satisfactory level. Assertion 3 implies that some vulnerabilities may be more difficult to find and may generate false negatives. Similarly, some vulnerabilities may be more complex and can lead to the generation of false positives. Hence, it is important to grade each vulnerability to an adequate level.

Assertion 1 states the need to grade the various vulnerabilities with a vulnerability grading system. The system could create a list of known web-based vulnerabilities from the online vulnerability databases [15], classified into evolved, dormant, relatively-new or new categories as defined in Assertion 2.

Such a grading would also provide a better understanding of the vulnerability. The grading is a constantly changing one as the scanner algorithms may change over time with upgrades and there are also instances that the vulnerability definitions themselves may change [23].

The *difficulty of detection* of a vulnerability *j* is given by

$$D(j) = 1 - \frac{\sum_{s=1}^{S_n} (V(j)_{TN} + V(j)_{TP}) \Big/ 2}{S_n} \qquad \textbf{(19)}$$

Where,
D(j) is the difficulty of detection of vulnerability *j*
S_n is the number of scanners used
$V(j)_{TN}$ is the vulnerability specific true negative value for vulnerability *j* for a specific scanner
$V(j)_{TP}$ is the vulnerability specific true positive value for vulnerability *j* for a specific scanner
The higher the value of D(j), the more difficult it is to detect and the lower difficulty implies the easier detection by the scanner.

TABLE 1: DETECTION EXAMPLE 1

Vulnerability	Scanner1	Scanner2	Scanner3
Vul1	YES	YES	YES
Vul2	YES	NO	YES
Vul3	NO	NO	NO
Vul4	YES	NO	YES
! Vul1	NO	NO	NO
! Vul2	NO	NO	YES
! Vul3	NO	YES	NO
! Vul4	YES	YES	NO

In the table 1, "VulX" refers to the positive vulnerability and "! VulX" refers to the negative vulnerability. Table 1 has been formed with the assumption that all the scanners used in the calibration have been designed with their respective algorithms to detect the stated vulnerabilities.
Applying Eqn. 19 to table 1, we can get
D(Vul1)=0; D(Vul2)=0.333; D(Vul3)=0.666; D(Vul4)=0.5;
From the above table we can conclude that the D(Vul3)>D(Vul4)>D(Vul2)>D(Vul1)

VII. Computation of Scanner report Confidence

While it is important for the end-user to be able to infer from various diagnostic reports, it is also important to be able to gauge the confidence of the information within the report. Confidence levels are required to ascertain if the report can be trusted and the extent of this trust. There needs to be a confidence level associated with every vulnerability detected by the various scanners. The grading systems are used in the calculation of the confidence of the report.

It must be noted that not all scanners agree on the reports generated. Some scanner algorithms may be better suited to tackle some classes of vulnerability and they are effective against these but the same algorithm may be the reason for their weaker performance against other classes of vulnerabilities. Some scanners have access to a very comprehensive database while others suffer from inadequate ones. Hence, using the S_{GRADE} (defined in Assertion 3) we can specify the performance of the scanner with respect to a particular vulnerability. When the results of all scanners agree on that vulnerability, the confidence on the report will be higher than the confidence due to an individual scanner. This factor is also moderated when scanners have a conflict in results. The moderation, however, depends on the metric values assigned to the scanners by the scanner grading system.

The methodology for calculating the 1st and 2nd degree confidence is detailed as follows :

To find 1st Degree Confidence.

The 1st degree confidence report gives a report based on the various scanner reports on a website location. The various scanner and vulnerability specific truth-values are used along with the vulnerability reports to give a more useful inference to the user.

There are two types of indices that the 1st Degree Confidence report can have. They are the *positive index* and the *negative index*. The report first tries to calculate the positive index. This represents the possibility of the vulnerability being present. It shows how confident the user can be about the vulnerability being present in a scale ranging from (>0%) to 100%. Only if the condition given in step 7 of positive index is satisfied, the negative index is calculated. The negative index represents the possibility of vulnerability not being present. The computed output will contain only one of the two indices. If any scanner is known to lack the detection capability with regard to the vulnerability, then the scanner must be excluded from the computation.

Positive Index Computation:

1. Create the set A that contains the respective Vulnerability specific V_{TP} or V_{TN} values depending on whether the scanner detects the vulnerability or not.
2. Let a_0={largest V_{TP} value in A}. If there are no V_{TP} values calculate negative index.
3. $A \leftarrow - \; A - a_0$, Where – refers to set subtraction.
4. Rearrange the remaining elements in descending order.
5. $A=a_0 \pm a_1(1-a_0) \pm (a_2(1-(a_0 \pm a_1(1-a_0))) \pm \ldots\ldots\ldots\ldots (a_n(1-(a_{n-1}(\ldots(1-a_2(1-(a_0 \pm a_1(1-a_0))))\ldots))))$ where **(20)**
   □ a gives the 1st degree confidence report
   □ $a_i \in A \; \& \; (0 \leq a_i < 1)$

6. If a_i is a V_{TP} value, it is added otherwise, it is subtracted.
7. If $a \leq 0$ calculate negative index.
8. The resultant value will give the confidence level for the vulnerability being present from the scanner reports obtained.

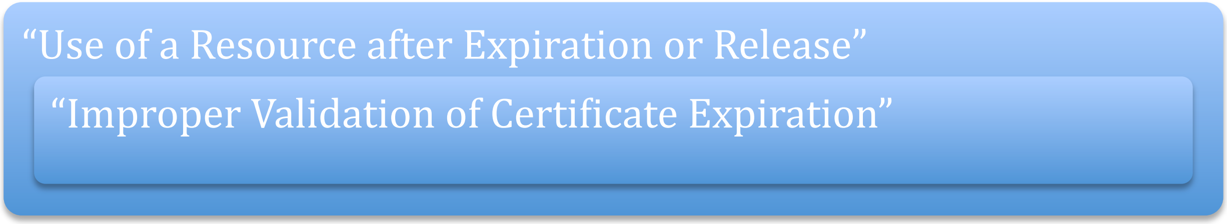

“Use of a Resource after Expiration or Release”
“Improper Validation of Certificate Expiration”

Figure 5. Relationship example

Negative Index Computation:

1. Create the set A that contains the respective Vulnerability specific V_{TP} or V_{TN} values depending on whether the scanner detects the vulnerability or not.
2. A0={largest V_{TN} value in A}
3. $A \leftarrow A - a_0$, Where – refers to set subtraction.
4. Rearrange the remaining elements in descending order.
5. $A=a_0 \pm a_1(1-a_0) \pm (a_2(1-(a_0 \pm a_1(1-a_0))) \pm \ldots\ldots\ldots\ldots (a_n(1-(a_{n-1}(\ldots(1-a_2(1-(a_0 \pm a_1(1-a_0))))\ldots))))$ where **(21)**
   - □ a gives the 1st degree confidence report
   - □ $a_i \in A \;\&\; (0 \leq a_i < 1)$
6. If a_i is a V_{TN} value, it is added otherwise, it is subtracted.
7. The resultant value will give the confidence level for the vulnerability being absent from the scanner reports obtained.

The value of a is the positive or negative index that is associated with the report. The general idea behind the derivation of a lies in the fact that the positive and negative indices can never be ~~a~~ 100%. While the rule in itself does not stop them from taking the value of 0%, it is that this value could appear. This is because being 0% requires both positive and negative indices to be 0 i.e. sufficient scanners to nullify the true detection and sufficient scanners to nullify the false detection. Hence, the value of a is restricted to be always less than 1.

To find 2nd Degree Confidence.

A 1st degree confidence report covers the direct existence of any vulnerability. However, a 2nd degree report is necessary to cover Assertion 5. The 2nd Degree Confidence covers the fact that one vulnerability can have a relationship to or influence the existence of another.

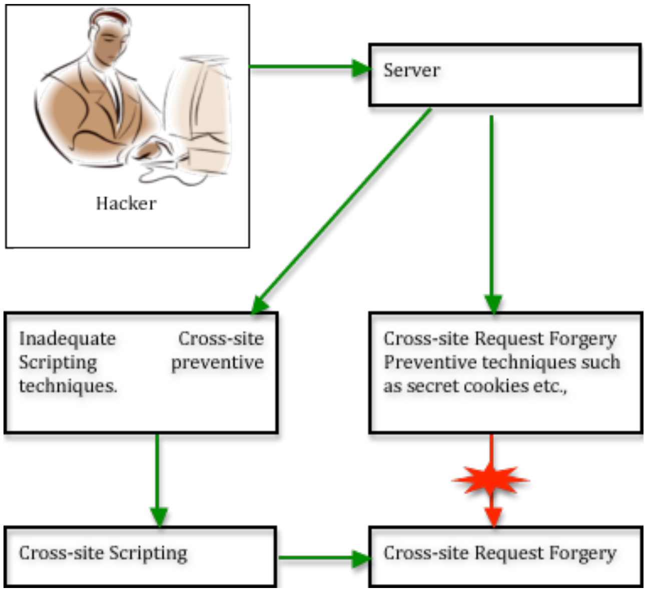

Hacker
Server
Inadequate Scripting techniques.
Cross-site preventive
Cross-site Request Forgery Preventive techniques such as secret cookies etc.,
Cross-site Scripting
Cross-site Request Forgery

Figure 6: XSS-CSRF Example

The 2nd Degree Confidence report is hence a result of combining the confidence reports with the Assertion 5. The relationship between two vulnerabilities is given by a value ranging between 0 and 1. Every vulnerability has a relationship of 1 with itself. The relationship with other vulnerabilities is generally less than 1. However, there could be exceptions as exemplified by Figure 5. If the vulnerability “Use of a Resource after Expiration or Release” is present, then the vulnerability “Improper Validation of Certificate Expiration” has a lower likelihood of presence but if the reverse situation occurs, the presence of the vulnerability “Improper Validation of Certificate Expiration” would mean the same (if not greater) likelihood of the presence of “Use of a Resource after Expiration or Release”. i.e. “Improper Validation of Certificate Expiration” has a relationship of 1 towards “Use of a Resource after Expiration or Release” but “Use of a Resource after Expiration or Release” has a relationship of

<1 towards “Improper Validation of Certificate Expiration”. Quality of the final result is enhanced by the use of these relations since they take more factors into account than a typical scanner-based risk analysis that are available currently [6].

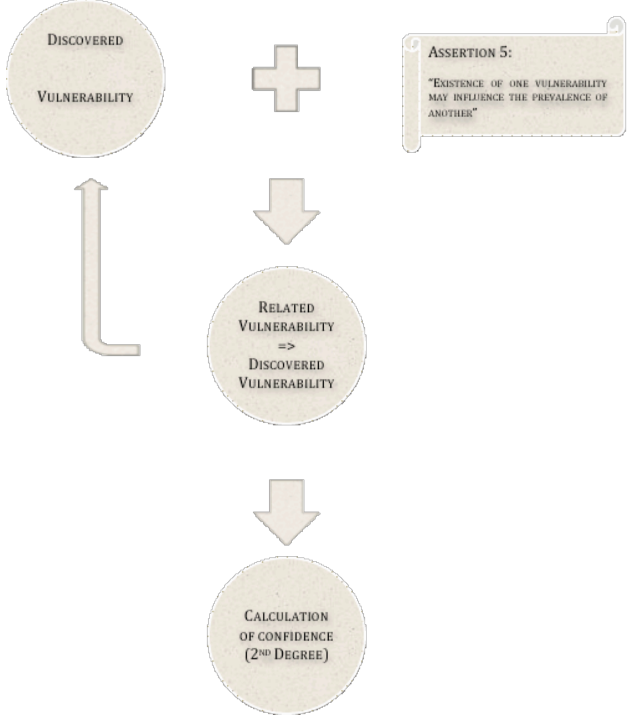


Figure 7: Calculation of 2nd degree report

From the Figure 5, the relation of “Failure to Follow Chain of Trust in Certificate Validation” to “Failure to Validate Certificate Expiration” can be defined as that of a subset-superset where the former is the superset and the latter is the subset. Such a relationship would also contain the fact that the absence of the former does not mean the absence of the latter. Another example is that of the Cross-site scripting and Cross-site request forgery.

The Cross-site request forgery (CSRF or XSRF) and Cross-site scripting (XSS) example is one that shows how influential the relationships are in a system. From Figure 6, it can be seen that even while CSRF is prevented for the hacker, if XSS is not prevented, the CSRF cannot be considered as an eliminated threat. It should also be noted that the reverse is not true.

The Figure 7 shows the calculation of the 2nd degree confidence report from the 1st degree report.

Let there be a list of related vulnerabilities for the vulnerability v. This list includes both the ones that are directly related and those that are indirectly related till a satisfactory set of related vulnerabilities is formed.

Let the a(v) be the 1st degree confidence of vulnerability v. If a(v) is a positive index value, then $a_0(v)=a(v)$. If a(v) is a negative index value, then $a_0(v)=0$.

The above step is the pre-operation step that needs to be done to ensure that the fact that the steps taken against one vulnerability do not reflect on the decrease in the 2nd degree confidence if the Assertion 5 relates to it through another vulnerability known to be present. The extent of confidence on the vulnerability’s presence, however, depends on the strength of the relationship value and the confidence level on the vulnerability relating to it. A chain of related vulnerabilities could be formed by relating one vulnerability to another that is related to another and so on.

The Figure 8 shows the relationship between various vulnerabilities obtained from a recent analysis.

$$\begin{aligned}&a_1(v) = a_0(v) + \\ &(1 - a_0(v))(R_v(v_1) * a_0(v_1)) + \\ &(1 - ((1 - a_0(v))(R_v(v_1) * a_0(v_1)))(R_v(v_2) * a_0(v_2))) + \\ &(1 - (1 - ((1 - a_0(v))(R_v(v_1) * a_0(v_1)))(R_v(v_2) * a_0(v_2))))(R_v(v_3) * a_0(v_3)) + ...\end{aligned} \qquad \mathbf{(22)}$$

Using the above equation, the $a_1(v)$ is calculated for the entire set of related vulnerabilities {R}. In case of a closed relation eg. v related to v1, v1 related to v2 and v2 related to v, then v is not considered related to v2 during calibration of v.

As the relationship keeps expanding eg.v related to v1, v1 related to v2 etc., the same procedure is repeated for $a_2(v)$ replacing all $a_0(v)$ with $a_1(v)$, then $a_3(v)$ and so on. It is repeated till $a_n(v)$ is calculated. This gives the 2nd degree confidence level.

The confidence levels have been found to be a useful criteria for evaluation of any given website. A remediation database is being designed to give a suitable remediation based on the confidence level. Hence a more accurate remediation can be provided to the user.

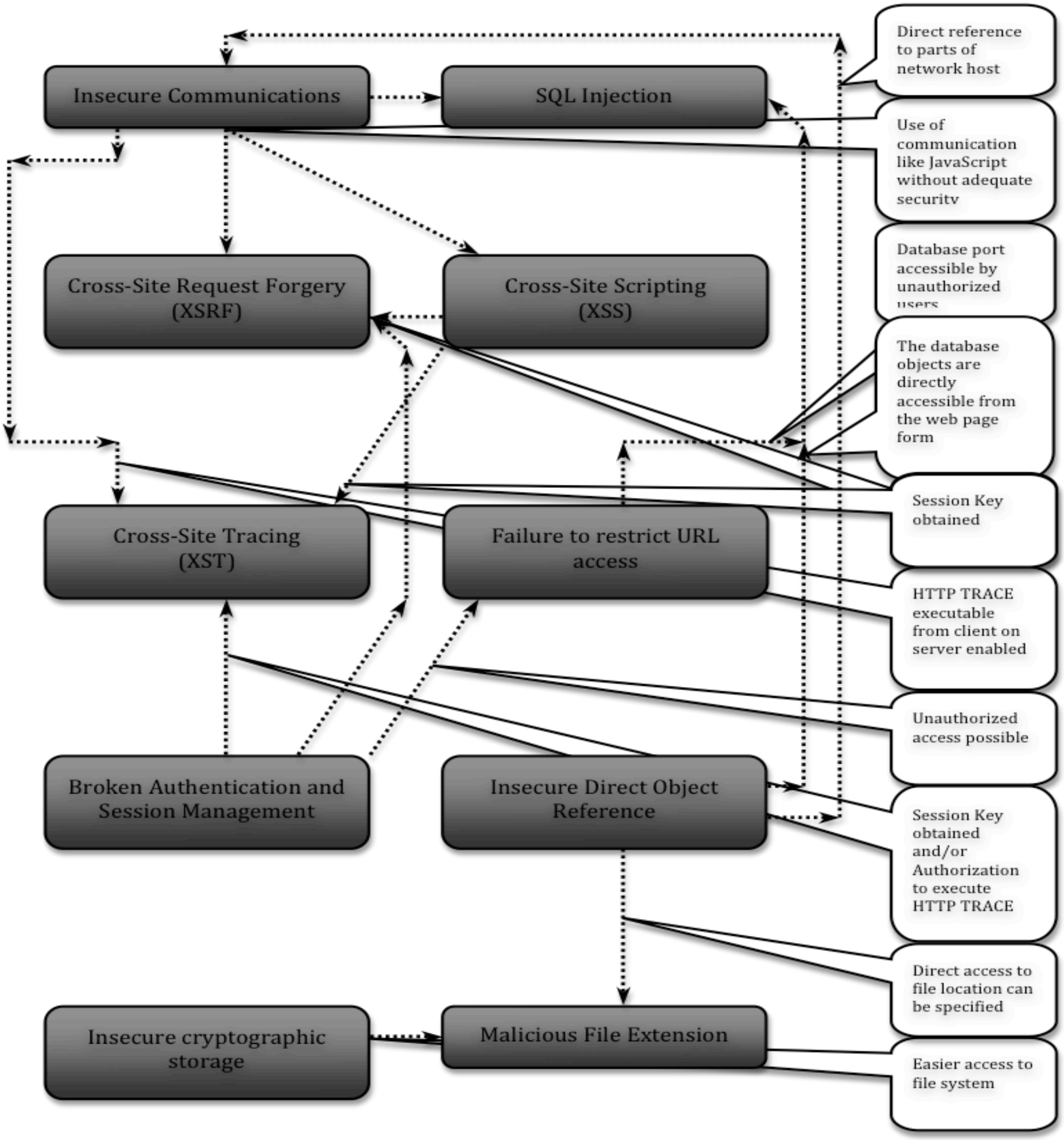

Insecure Communications
SQL Injection
Cross-Site Request Forgery (XSRF)
Cross-Site Scripting (XSS)
Cross-Site Tracing (XST)
Failure to restrict URL access
Broken Authentication and Session Management
Insecure Direct Object Reference
Insecure cryptographic storage
Malicious File Extension
Direct reference to parts of network host
Use of communication like JavaScript without adequate security
Database port accessible by unauthorized users
The database objects are directly accessible from the web page form
Session Key obtained
HTTP TRACE executable from client on server enabled
Unauthorized access possible
Session Key obtained and/or Authorization to execute HTTP TRACE
Direct access to file location can be specified
Easier access to file system

Figure 8: Analysis of relationship between vulnerabilities

VIII. Illustrative Example

Let there be 3 scanners s_1, s_2, s_3. Hence S_n=3.
Let there be 5 vulnerabilities detected and stored in the database v_1 v_2 v_3 v_4 v_5.

TABLE 2: S-V TABLE

	s_1		s_2		s_3	
	Instance1	Instance 2	Instance1	Instance 2	Instance1	Instance 2
v_1	Y	Y	Y	Y	Y	Y
v_2	Y	Y	Y	Y	Y	N
v_3	Y	Y	N	N	Y	Y
v_4	Y	Y	Y	N	N	N
v_5	Y	N	N	N	N	Y

TABLE 3: S-!V TABLE

	s_1		s_2		s_3	
	Instance1	Instance 2	Instance1	Instance 2	Instance1	Instance 2
! v_1	N	N	N	N	N	N
! v_2	N	N	N	N	N	Y
! v_3	N	N	Y	Y	N	N
! v_4	N	N	N	Y	Y	Y
! v_5	N	Y	Y	Y	Y	Y

Table 2 shows the detection for instances where the vulnerability is known to be present. Table 3 shows the detection where the vulnerability is known to be absent.

TABLE 4: VULNERABILITY SPECIFIC TRUTH-VALUES

	v_1	v_2	v_3	v_4	v_5
s_1	{.9 .9 .1 .1}	{.92 .8 .08 .2}	{.8 .7 .2 .3}	{.8 .7 .2 .3}	{.7 .6 .3 .4}
s_2	{.8 .7 .2 .3}	{.7 .6 .3 .4}	{.7 .6 .3 .4}	{.6 .6 .4 .4}	{.5 .5 .5 .5}
s_3	{.7 .6 .3 .4}	{.8 .71 .2 .3}	{.6 .6 .4 .4}	{.9 .8 .1 .2}	{.7 .6 .3 .4}

The Vulnerability specific truth-values in table 4 are represented as {V_{TP}, V_{TN}, V_{FN}, V_{FP}}

$D(v_1)$=1-(6/6)=0
$D(v_2)$=1-(5/6)=1/6=0.1667
$D(v_3)$=1-(2/3)=1/3=0.3333
$D(v_4)$=1-(1/2)=1/2=0.5
$D(v_5)$=1-(1/4)=3/4=0.75

Hence $D(v_5)>D(v_4)>D(v_3)>D(v_2)>D(v_1)$

1st degree confidence.

In this example let us consider the 2nd instance of v_2 from the table 2.

The results from the 3 scanners are {Y Y N}
Hence A={.92 .71 .7} and a_0=.92
Hence A={.71 .7}
a=.92-(.08(.71))+(1-.(92-(.08(.71))))(.7)=0.95896
In other words, we can be 95.896% certain that the result is true. This could mean there is high need for appropriate remediation.

2nd degree confidence.

TABLE 5: R-V TABLE

	v_1	v_2	v_3	v_4	v_5
R_{v1}	-	-	-	-	-
R_{v2}	.08	-	-	-	-
R_{v3}	-	-	-	.05	.04
R_{v4}	-	-	-	-	.09
R_{v5}	-	-	.1	-	-

The Table 5 shows the relationship between the various vulnerabilities and the suitable relationship values.

Let us consider the 2nd instance in the Table 2 and calculate the confidence for vulnerability v_2. Since only v_1 is related to v_2 (There are no indirect relations as v_1 is not related to any other vulnerability).

We know that for 2nd instance in Table 2,
$a_0(v_1)$ =.987{Y YY , all scanners have detected the vulnerability as positive}
$a_0(v)$ =.95896
$a_1(v)$ =.95896+.040506=.9994

Hence, the 2nd degree confidence report would suggest a possibility of 99.94% for the vulnerability's occurrence. Since $a_1(v)>a_0(v)$, it implies that other vulnerabilities are also present and must be rectified to rectify this vulnerability.

Let us consider another example using the 1st instance from table 2.

For the vulnerability v_5, A={Y N N}

1st Degree Confidence,
$a_0(v_5)$ =.7-(.3(.6))-((1-(.7-(.3(.6))))(.5)) =0.28

It can be observed that when there is only one scanner supporting the vulnerability, the confidence of report in the presence of the vulnerability also drops to a great extent. In the case shown above, the confidence report still remains in the positive index.

2nd Degree Confidence,

The vulnerability v_5 is related to v_3, which in turn is related to v_4 and v_5, and v_4 is related to v_5. The 1st degree confidences for these vulnerabilities can be given for the 1st instance of table 2 as

$a_0(v_3) \quad =.8+(.2(.6))-(1-(.8+(.2(.6))))(.6) \quad =0.872$
$a_0(v_4) \quad =.8-(.2(.8))+(1-(.8-(.2(.8))))(.6) \quad =0.856$

$v=v_5$

$a_0(v) \quad =0.28$
$a_1(v) \quad =a_0(v)+(1-a_0(v))(0.1(a_0(v_3)))=0.28+(0.72)(0.1(0.872))=0.343$
$a_2(v) \quad =a_1(v)+(1-a_1(v))(0.1(a_1(v_3)))$

$a_1(v_3)$ can be given by,
$a_1(v_3) \quad =a_0(v_3)+(1-a_0(v_3))(.05(a_0(v_4))\underline{+(1-(1-a_0(v_3))(.05(a_0(v_4)))(0.04(a_0(v_5))}$

but, since v=v5, the underlined portion of the equation above cannot be considered

hence, a1(v3) becomes
$=0.872+(.128)(.05(0.856))$
$=0.877$

therefore,
$a_2(v) \quad =0.343+0.657(0.1(0.877)) = 0.401$

Hence the 2nd Degree Confidence report shows 40.1% confidence in the presence of vulnerability v5 compared to the 28% confidence showed by the 1st degree confidence report.

IX. Conclusion

This research has enabled the improved risk analysis of web-based vulnerabilities. While several scanners are available to detect the vulnerabilities, their varying algorithms and proprietary nature makes it difficult to ascertain if the vulnerabilities found by them is true or false. The methodology used in this paper is a practical approach designed to work in spite of the proprietary nature of the algorithms while still being able to grade the various scanners. The variable nature of a vulnerability is also accounted for and the proposed methodology uses fuzzy-based classification and estimation metrics solve this problem. Another problem is the lack of detection of vulnerabilities whose presence is also influenced by other vulnerabilities. The proposed methodology is an effective one to tackle such problems as well. Five assertions have been defined to help establish the theoretical aspects of the approach. By using relationship between vulnerabilities given by assertion 5, the 2nd degree confidence report can be used to tackle such a problem. The confidence reports have been able to provide the user with valuable information and this has been tested with a successful implementation of the system. The confidence reports also provide a greater reliability of results than that of individual scanner reports. The open issues faced in this research include the need to grade every scanner for every vulnerability using test sites, which can be a very tedious process. The methodology of finding the relationship between vulnerabilities is still in progress. The development of an inference engine to compute the diagnosis is also in progress. The remediation database to accurately provide the remediation based on confidence level is being expanded to a sizable level of vulnerabilities.

References

[1] D. Subramanian, H.T. Le, P.K.K. Loh, “Fuzzy Heuristic Design For Diagnosis Of Web-Based Vulnerabilities”, *The Fourth International Conference on Internet Monitoring and Protection* (ICIMP), May 2009, Venice/Mestre, Italy.

[2] H. T. Le and P. K. K. Loh, "Unified Approach to Vulnerability Analysis of Web Applications," in *International Electronic Conference on Computer Science. AIP Conference Proceedings*, Volume 1060, pp. 155-159 (2008)

[3] H.-T. Le and P. K. K. Loh, "Realizing Web Application Vulnerability Analysis via AVDL," in *10th International Conference on Enterprise Information Systems (ICEIS 2008)*, Barcelona, Spain, 2008, pp. 259-265.

[4] H. T. Le and P. K. K. Loh, "Evaluating AVDL Descriptions for Web Application Vulnerability Analysis," in *IEEE International Conference on Intelligence and Security Informatics 2008 (IEEE ISI 2008)*, Taipei, Taiwan, 2008, pp. 279-281.

[5] Jonathan Gomez and Dipankar Dasgupta, “Evolving Fuzzy Classifiers for Intrusion Detection”, *Proceedings of the 3rd Annual IEEE Information Assurance Workshop*, New Orleans, Louisiana: June, 2002.

[6] Larry Suto, “Analyzing the Effectiveness and Coverage of Web Application Security Scanners”, White Paper, October 2007, Publisher: Strategic Data Command, http://www.stratdat.com/webscan.pdf

[7] P. Mell, K. Scarfone, and S. Romanosky, "The Common Vulnerability Scoring System (CVSS) and its applicability to federal agency systems," NIST Interagency Report, NIST-IR7435, pg. 1-20, August 2007, http://csrc.nist.gov/publications/nistir/ir7435/NISTIR-7435.pdf

[8] Quals, “Vulnerability Management for Dummies”, *Copyright © 2008 by John Wiley & Sons Ltd*, Chichester, West Sussex, England.

[9] Ramaswamy Chandramouli, Tim Grace, Rick Kuhn and Susan Landau, “Emerging Standards: Common Vulnerability Scoring System”, IEEE Security & Privacy, vol. 4, no. 6, pp. 85-89, Nov/Dec, 2006.

[10] David Minnen, Tracy Westeyn, Thad Starner, Jamie A. Ward and Paul Lukowicz, “Performance Metrics and Evaluation Issues for Continuous Activity Recognition”, Performance Metrics for Intelligent Systems (PerMis’06), Gaithersburg, Maryland, United States of America, August 21-23, 2006, www.cc.gatech.edu/~dminn/papers/minnen-permis2006.pdf

[11] Ambareen Siraj, Susan M. Bridges, Rayford B. Vaughn, “Fuzzy Cognitive Maps for Decision Support in an Intelligent Intrusion Detection System”, *Proceedings of the Joint 9th International Fuzzy Systems Association World Congress and the 20th North American Fuzzy Information Processing Society International Conference on Fuzziness and Soft Computing in the New Millennium*, vol. 4, pg. 2165-2170, July 2001, Vancouver, Canada,

[12] Jeffrey R. Jones, “Estimating Software Vulnerabilities”, *IEEE Security & Privacy*, Volume 5, Issue 4, pg. 28 – 32, July-Aug 2007.

[13] Omar H. Alhazmi, Yashwant K. Malaiya, "Quantitative Vulnerability Assessment of Systems Software", *Proceedings of the Annual Reliability and Maintainability Symposium*, pg. 615-620, Jan. 2005.

[14] P. Wang, K.-M. Chao, C.-C. Lo, C.-L. Huang, and M. Younas, "A fuzzy outranking approach in risk analysis of web service security ", *Cluster Computing*, vol. 10, pp. 47-55, 2007.

[15] S. T. Halkidis, A. Chatzigeorgiou, and G. Stephanides, "Quantitative Evaluation of Systems with Security Patterns Using a Fuzzy Approach", *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*: Springer Berlin / Heidelberg, 2006, pp. 554-564.

[16] Kenneth L. Ingham, Anil Somayaji, John Burge and Stephanie Forrest, "Learning DFA representations of HTTP for protecting web applications", *Computer Networks: The International Journal of Computer and Telecommunications Networking*, Volume 51 , Issue 5, pg. 1239-1255 , April 2007

URLs

[17] CERT/CC Statistics 1988 – 2006, http://www.cert.org/stats/

[18] National Vulnerability Database (NVD) Statistics, http://nvd.nist.gov/statistics.cfm

[19] DHS National Security Division, NIST, "Web Application Vulnerability Scanners", https://samate.nist.gov/index.php/Web_Application_Vulnerability_Scanners

[20] Common Vulnerabilities and Exposures (CVE), http://cve.mitre.org/

[21] Jeremiah Grossman, "WhiteHat Website Security Statistics Report", October 2007, Publisher: WhiteHat Security (United States of America). https://whitehatsec.market2lead.com/go/whitehatsec/WPStatsreport_100107

[22] Jeremiah Grossman, "WhiteHat Website Security Statistics Report", August 2008, Publisher: WhiteHat Security (United States of America). https://whitehatsec.market2lead.com/go/whitehatsec/WPstats0808

[23] Jeremiah Grossman, "WhiteHat presentation on XSRF", Publisher: WhiteHat Security (United States of America) https://whitehatsec.webex.com/whitehatsec/nbrshared.php?action=playback&recordID=21578512&recordKey=2E8BF7FFE53556F277FD706294A7E3ED86F81580F46B5A0DDC7345881C9B224C

[24] N-Stalker®, "Overview of N-Stalker Reports", http://nstalker.com/products/development/report-details, Retrieved on 20th July 2009

[25] Accunetix web application Security, "In depth checking for SQL Injection, Cross Site Scripting (XSS) and Other Vulnerabilities", http://www.acunetix.com/vulnerability-scanner/sql-injection-ft.htm, Retrieved on 20th July 2009

[26] Wikipedia, XSS, http://en.wikipedia.org/wiki/Cross-site_scripting, Retrievedon 20th July 2009.

Security Analysis of Private Data Enquiries in Erlang

Florian Kammüller
Technische Universität Berlin
Software Engineering Group
flokam@cs.tu-berlin.de

Reiner Kammüller
Universität Siegen
Fakultät für Elektrotechnik und Informatik
reiner.kammuller@gmail.com

Abstract

Privacy is an issue of increasing concern to the Internet user. To ensure the continued success of distributed information systems, a reliable information flow must be established in certified but immediately evident ways. We begin with basic consideration of the privacy problem in the general setting of database enquiries. From there, we develop a simple solution, which we illustrate with a simple implementation in the programming language Erlang. We first provide an informal security analysis that is then developed into a formal definition of a type system for noninterference.

Index Terms

Privacy; distribution; noninterference; type systems

1. Introduction

Privacy has become an important issue in public e-business. In order to protect their customers, commercial services have to provide electronic privacy, which is approximated by anonymity using pseudonyms. However, it has been long known that chaining transactions quickly reveals the identities behind pseudonyms. Even more crucially, applications that appear secure from a superficial point of view may well contain numerous covert channels. Some of these covert channels – the ones inherent in the logic of programs – can be identified by a painstaking information flow analysis [10]. Such an analysis verifies a formal notion of security over different data domains, so-called noninterference [12], for all possible control flow of programs. Even without this classical but cumbersome method, some kind of formal language would appear to be necessary for a thorough analysis of security risks.

Several formal languages have been proposed to encode privacy policies. The Platform for Privacy Preferences (P3P) is just one example of a language that enables enterprises to communicate their privacy policies to customers. The customers may then decide whether they are willing to accept a policy prerequisite for their database enquiry. Apparently, even with means such as P3P, it is not easy to determine whether in-house enforcement policies meet their published P3P privacy promises [4].

In this paper, we first provide a simple formal specification of an obvious requirement for such privacy promises illustrating that it is virtually impossible to expect such policies to work. From this, we devise a simple idea of a different database enquiry that achieves privacy. We illustrate this solution using a prototype in the parallel programming language Erlang (Section 2). Efficiency is the price to pay for the privacy gained. We further illustrate how parallelism in Erlang helps to overcome this drawback. We then justify our claim with an informal security argument (Section 3). Once given the intuition how such an analysis of information flow works we introduce the formal approach to noninterference. We further define a type system that enables the static analysis of Erlang programs for covert channels (Section 4). Finally, we briefly contrast our Erlang solution for private data searches with alternative approaches using active objects or Java, and offer our conclusions (Section 5). This paper is an extended version of an earlier conference paper [19]: the formal approach to security by an original noninterference type system for Erlang contained in Section 4 is novel.

1.1. Privacy Policies

The enforcement of privacy policies within an enterprise constitutes an interesting problem in itself. However, if we ignore for a moment the actual implementation issue and try to establish a precise requirement specification for some of the problems involved, we can identify *data retention* as conflicting with privacy. By retention, we mean the requirement that user data provided for the identification of services only be retained a specified period after which the data must no longer be stored in the enterprise's database.

Formally, we can identify two operations *copy* and *delete* simply denoting that a data item is copied at the enterprises site and that it is deleted in order to regain some privacy. We assume the following algebraic properties of *copy*, *delete*, and *run*, a process representing all possible behaviours.

$$\begin{aligned} copy; delete &= id \\ copy; run &= run; copy \end{aligned}$$

Using a specification formalism like CSP [14], we could now specify what is meant by the fact that a system P *does not retain data* d for any alphabet A of possible system events, as follows.

$$copy(d); run(A \setminus copy(d)); delete(d) \sqsubseteq P$$

Here, the refinement order relation $\sqsubseteq$ constrains the behaviour of P in that the specification *spec* on the left-hand side is only implemented by such processes P that implement a behaviour contained in *spec*.

The expression *run(A* $\setminus$ *copy(d))* specifies that between a *copy(d)* and the corresponding action *delete(d)* any sequence of events of A may happen, except another copy of d. The specification thus prohibits excessive copies and hence unauthorized retention of data d.

The interesting question is whether we can guarantee such a behaviour. In principle, the answer is yes if we can observe every sequence of actions in a server of which we require a service. Pondering this for a moment, we realize that the above retention specification is unrealistic for real-world scenarios: no service will lay open all its internal action traces.

Starting from this discouraging – but highly compelling insight – we develop a different type of database enquiry that differs from the usual service architecture model. Instead of disclosing our incentives, i.e. private data, we perform the kernel action on the data offered by a service ourselves. Clearly, there will be a loss of efficiency but we will gain security. Since the data we wish to keep private is contained in our kernel service action, the service provider has no access to it.

There are cryptographic schemes addressing similar problems. The most general, oblivious transfer, by Rabin [30], follows ideas similar to the original “Conjugate coding” by Stephen Wiesner now so popular through quantum cryptography. The scheme of private information retrieval [7] is closely related. This scheme abandons perfect secrecy for the sake of efficiency – the solution protects against attackers bound by complexity theory.

The approach we investigate here is the only one that guarantees privacy in an information theoretic sense but is deemed “practically unacceptable” because of the communication overhead [7]. We illustrate this approach in Erlang and show how massive parallelization may be used to minimize the effort.

2. An Erlang Implementation of Database Enquiries

A database enquiry is a service that is usually provided by a server through the transfer of a search key to the server, e.g. Google. In the respective service the server performs a search action on the data, e.g. the Internet, that is in its data domain. Unfortunately, this efficient standard solution implies that we must trust the server not to make unauthorized copies of our search key, i.e. the private data we wish to keep confidential. For example, we might need to input our name, address and some incentive in order to find the required services in our neighbourhood.

Instead of disclosing our personal information, we can demand access to some larger relevant data domain and perform the selection, i.e. the search corresponding to our profile or key, in our private secure domain. We will illustrate this type of database enquiry on a concrete implementation in the parallel programming language Erlang [1]. We begin with a short introduction to Erlang.

2.1. Erlang

The programming platform Erlang/OTP provides the infrastructure for programming open distributed telecommunication (OTP) systems. The language Erlang [1] was developed by the Ericsson corporation to address the complexity of developing large-scale programs within a concurrent and distributed setting. The platform Erlang/OTP consists of the functional language Erlang – with support for communication and concurrency – and the OTP middleware.

The most important features of Erlang include the following.

- Erlang variables are immutable: their value is assigned once only; no multiple assignments are allowed.
- Erlang processes do not share memory space; interaction is through explicit message passing.
- Erlang’s process creation speed is much faster than the operating system’s processes, much like thread creation [1][Section 8.4].

The programming style of Erlang resembles that of the ML language [28]. Recursive functions may be defined in a fairly intuitive way. For example, the factorial function is defined as follows.

```
fact 0 -> 1;
fact N -> N * fact(N-1).
```

Processes may be created by the `spawn` command, which takes the processes’ function and initial arguments as parameters. The value of a `spawn` command is the process identifier `Pid` of the created process. Message passing between parallel processes is, for sending, simply written as `Pid ! message` – in our example the process identifier `Pid`. Reception of messages in processes is organized through a mailbox in each process that can be read by the receive command. Using pattern matching, receive-statements can be written concisely and elegantly. The main data types are (untyped) lists and records, e.g. {green, apple}. Any lower-case name is interpreted as a constant, and higher-case names are variables. These various language features are used below when considering our database enquiry.

2.2. A Simple Database Enquiry

To simplify matters, we assume that the database is a file of already structured data. We do so to focus our attention on the communication necessary for the enquiry, leaving out the complexity of a realistic data analysis. In brief, the basic database enquiry program implements a server providing the database and our privacy-aware client that orders the data and performs a search on it. We explicitly model the server to provide a basis for the subsequent security analysis. To model a real world scenario, we provide simple programs for these two components. Later, we will see how we can improve the system through parallelization to enhance performance.

The server listens on a port for the opening of a socket and accepts the socket. After accepting it, the server closes the listening socket which does not affect the existing connection but merely prevents new connections. The Erlang package gen_tcp optimally supports the implementation of such distributed systems based on the tcp-protocol. We omit some parameters so as not to overload the exposition. The complete program code can be downloaded from the authors' website [1].

```
start_server() ->
  {ok, Listen} = gen_tcp:listen(2345, ...),
  {ok, Socket} = gen_tcp:accept(Listen),
  gen_tcp:close(Listen),
  loop(Socket).
```

The loop procedure repeatedly reads data units from the database accessible to the server. To facilitate the example, databases are simply represented as files. The socket is opened and closed by the client. The server opens the database, represented by the file specified by the client, and delegates processing of the stream transfer to the send_stream procedure.

```
loop(Socket) ->
   receive
     {tcp, Socket, FileB} ->
        FileS = binary_to_term(FileB),
        {ok, S} = file:open(FileS, read),
        ok = send_stream(Socket, S),
        loop(Socket);
     {tcp_closed, Socket} -> ok
   end.
```

The data is sent by the procedure send_stream to the socket in a repeated read action from the opened file stream S until end of file eof is reached.

```
send_stream(Socket, S) ->
  case io:read(S, '') of
    {ok, X} ->
       gen_tcp:send(Socket, term_to_binary(X)),
       send_stream(Socket, S);
    eof     ->
       file:close(S),
       gen_tcp:send(Socket, term_to_binary(eof)),
       ok
  end.
```

The client now opens the socket and transmits the database we wish to investigate, represented by a file. Here, we use a generic name host, representing some actual hostname. The actual reception of the file's contents is delegated to the procedure client_receive. The search results are returned by this procedure as a result list Res and are immediately output.

```
client_eval(Key, FileS) ->
 {ok, Socket} =
   gen_tcp:connect("host", 2345, ...),
 ok = gen_tcp:send(Socket, term_to_binary(FileS)),
 {eof, Res} = client_receive(Key,Socket,self(),[]),
 io:format("Client result: ~p~n", Res),
 gen_tcp:close(Socket).
```

The database's contents arrive at the client and are immediately analyzed corresponding to the search key Key. The actual data analysis is, for clarity's sake, reduced to a simple pattern matching on the received data items. Only matching contents are assembled in the result list Res.

```
client_receive(Key, Socket, From, Res) ->
 receive
  {tcp, Socket, Bin} ->
    Val = binary_to_term(Bin),
     case Val of
      eof      -> From! {eof, Res};
      {Key, X} ->
       client_receive(Key, Socket, From, [X|Res]);
      Any      ->
       client_receive(Key, Socket, From, Res)
    end
 end.
```

Two processes, one for the server and one for the client, can now be started independently by compiling the code presented above on two separate sites running Erlang. Invoking the function start_server() on the first, the server's site, while calling client_eval(key,"file.dat") on the client site has the following effect on the latter

```
  Client result: "Ottostr 38, 10999 Berlin"
  ok
```

where the key was drugstore and file.dat contains, amongst other arbitrarily structured data, an item

```
  {drugstore, "Ottostr 38, 10999 Berlin"}.
```

[1] http://www.swt.cs.tu-berlin.de/∼flokam/research

The server site only reports ok after successful termination of the process.

2.3. Efficiency by Parallelization

The simple client server introduced in the previous section represents the desired security solution but it is clearly not efficient because all data has to be transferred from the server to the client before the actual selection takes place. Generally, security does not come for free, so we can see this as the price to be paid. However, the communication overhead may constitute a crucial bottleneck in an application. One of the strong points of Erlang is the possibility to create a large number of parallel processes. To show that our approach scales up to realistic application scenarios, we present below an extension to the previous basic program which significantly enhances performance. In fact, this extension is a standard way of using Erlang. We therefore only show the extensions to the basic program presented in the previous section to explain the principle, but also go on to discuss some important practical issues.

The main clue to parallelizing the server is to start a new parallel process in `start_server` whenever a new connection is provided by a client through `gen_tcp:accept`. Note, that the listening socket is, unlike the sequential server, not closed down as we accept new connections.

```
start_par_server() ->
  {ok, Listen} = gen_tcp:listen(...)
  spawn(fun() -> par_connect(Listen) end).

par_connect(Listen) ->
  {ok, Socket} = gen_tcp:accept(Listen),
  spawn(fun() -> par_connect(Listen) end),
  loop(Socket).

loop(..) -> % as above
```

On the client site, we use the same principle to make parallel client processes each communicating with a parallel server.

```
par_client(Key, FFile) ->
  {ok, S} = file:open(FFile, read),
  ok = client_par_eval(Key, S).

client_par_eval(Key, S) ->
  case io:read(S, '') of
    {ok,FileS} ->
        spawn(fun() -> client_par_eval(Key,S) end),
        client_eval(Key, FileS);
    eof -> file:close(S),
           ok
  end.
```

The input of the file names of the files to be searched is provided by an input file `FFile` on the client site. The gradual selection of new source files for a goal-directed search may be integrated (see Section 5).

This parallel server can potentially create thousands of connections. Performance is thus significantly enhanced, although clearly the bandwidth of the communication channels is strained. For a more sophisticated implementation, we can limit the maximum number of simultaneous connections by simply keeping count of new connections and finished ones.

3. Informal Security Analysis

3.1. Security Assumptions

A security analysis starts with a two-sided model comprising (a) the attacker and (b) the security policy, or security goals. We cannot achieve 100% security because (a) there always is the all-powerful attacker and (b) we cannot generally achieve all security goals for all involved parties because they may conflict. Usually, when investigating privacy, we use a multilateral security model [36] that enables consideration of differing protection goals of several involved parties.

Nevertheless, we analyze the privacy of the client using a typical multi-level security model (MLS) [9] because we are, in this paper, only interested in the privacy of the client's data. We therefore assume, for the security policy, that the user – or, in our case, the client process – has a higher security level than the server side, the potential attacker. Let this security level be H, or high, for the client, and L, or low, for the server. We further extend the security policy by assuming that the local host is a secure domain, i.e. that its data and internal communication are secure. All other communication channels outside the client, and all data on the server, is assumed to be visible to the attacker.

3.2. Information Flow Security

The first to formalize information flow in a program were the Dennings [10]. The most natural way to formalize confidentiality as a property of information flow have been Goguen and Meseguer by their notion of noninterference [12]. There are quite a few different definitions of noninterference [32], mainly because it is a relation over behaviours of programs (it it sometimes characterized as a bisimulation property). Thus, the underlying computation model – leading to different notions of behaviour – results in different notions of noninterference. Without giving a formal introduction to this notion, we attempt to provide a basic understanding of it. We adopt a state-based view: program behaviour is viewed as a transition between vectors of variable values.

The basis of noninterference is a relation of *indistinguishability* of program states based on a similar relation on the program variables: high variables are all indistinguishable, but low variables are only if they have equal values. Informally, the indistinguishability between states during a program run is defined extensionally over the indistinguishability of its components, the state variables.

Given an indistinguishability relation on program states, we can say that noninterference is defined as *low*-indistinguishability. In other words, given a security policy that assigns high and low to all data variables, a program is non-interfering iff any two program runs remain *low*-indistinguishable throughout the program behaviour if they have been so from the start.

The important implication of noninterference is that the attacker, who can only read low values, is thus unable to learn anything about the values of high variables, even if he can observe different runs of the same program on different – but indistinguishable – data.

To show noninterference with respect to a given security policy in practical terms, we have to analyze all control flows of a program and ensure that there are no information flows from high to low variables. In practice, this process is often supported by a static analysis with specialized noninterference type systems [18], [32] (see Section 4).

3.3. Informal Security Analysis of Privacy-Enhancing Database Search

Although we do not yet intend to provide a formal analysis according to some notion of noninterference here, we already wish to use its essential idea in an informal argument. Let our security policy be an assignment that assigns high to the variables `Key`, `Any`, and `Res`. All other data may be assigned low; most of the variables, like `Socket`, `S`, and `FileS`, must be low because they have to be communicated between the insecure server and the confidential client. To show noninterference, we have to analyze all control flows in our program and exclude all explicit and implicit information flows from `Key`, `Any` and `Res` to any other (low) variable.

An explicit flow is either given by an assignment from one variable to another, which is impossible in Erlang as it is functional (all variables are only assigned once), or it is given by a function call, whereby a value can then be assigned as the initial value of the receiving process. The `Key` variable is passed on from `client_eval` to `client_receive`, and from `client_receive` again to itself in two separate recursive calls. In both invocations, `Key` is again assigned by pattern matching to the variable `Key`, which is also high, so there is no illegal explicit flow there. These are all explicit flows from `Key` to any other variable.

An implicit information flow is given when the control flow can branch, e.g. at an `if` statement: according to the value of the first variable, the tested variable, a second variable in one of the branches receives a value, depending on the value of the first. Again, such an implicit flow should not lead from a high to a low variable. The only possible branching of the control flow is the `case` statement in `client_receive`. Here, there are implicit flows from `Key` to `Any` and `Res`: depending whether `Val` matches `Key` one of the two "non-eof" branches is selected. Consequently, `X` – containing matching data – is added to the result list `Res` if the `Key` match is successful otherwise nothing is added to `Res`. Therefore, `Res` has to be marked as H as well. In addition, the variable `Any` must be H because – if it were L – we could work out `Key`: the difference between the original file contents and those matched with `Any` gives just those data items containing `Key` as first element: Bingo!

The variable `Any` is not used in any further function calls, so there are no explicit flows from it to any other variable. Considering, finally, the variable `Res`, we see that, here too, there are no flows from it to any low variable, neither explicit nor implicit. The final output of the value of receive by the `io:format` call must be considered secure because it happens inside the secure domain of the client and has no effect on other low variables. To summarize, the privacy-enhancing database search is non-interfering with respect to our security policy.

4. Formal Security Analysis

In this section, we use the ideas already provided in the previous section in the informal security analysis and make them formal. In detail, we present the syntax and semantics of Erlang – more precisely, a small subset of the original Erlang language, Core Erlang, sufficient for many applications, including ours. Then, we provide a novel type system that encodes the legal information flows and define a notion of indistinguishability. Finally, we show that programs that are accepted by the type system are secure. This proof implies that programs, for which a type can be inferred according to our type system and over an initial security policy, do not leak information.

In order to define formal semantics, it is always advisable to use a reduced language set cutting out, on one side, syntactic sugar, while still enabling, on the other side, the full power of the language. There have been a few attempts to define a Core Erlang language that serves exactly this purpose, e.g. [6]. Several other papers, in particular those concentrating on formal aspects of the Erlang language, e.g. [17], [25], define semantics also in a more formal way. This is a prerequisite for a formal analysis.

One way to take advantage of these earlier works is to use a rigorous translation from Erlang to some formal calculus and to exploit means for a security analysis. In [19] we proposed such a strategy for a formal analysis: use a translation of Erlang to the π-calculus [25] to represent our application in a calculus that is more easily accessible to a formal analysis. This would enable the use of existing formalizations of noninterference for the π-calculus [15] to demonstrate information flow security. However, after some consideration of information flow analysis in the π-calculus we decided to follow a different but much simpler approach based on *language based noninterference* by Volpano, Smith, and Irvine [35], [34], [33].

In this section, we base our presentation on the syntax and semantics for Erlang following Huch [17], present, then, an original noninterference type system for this language, to prove that a well-typed program does not contain illegal flows. We finally illustrate experimentally the use of our Erlang noninterference type system by showing how it is used to check our program in a process of inferring security types based on some initial security policy.

4.1. Syntax of Core Erlang

In the presentation of the syntax and semantics of Core Erlang we directly model the asynchronous communication and the message queues of Erlang. This is in difference to some models that base the formal models of Erlang on operational models of the π-calculus or CCS, unnecessarily complicating the situation by mapping the asynchronous communication of Erlang to synchronisation in π or CCS, respectively.

The syntax of Core Erlang is defined as follows where c denotes constructors and ϕ may denote any built-in or user defined function or constructor.

$$\begin{array}{rcl} p & ::= & f\texttt{(}X_1,\ldots,X_n\texttt{)}\ \texttt{->}\ e\texttt{.} \mid p\ p \\ e & ::= & \phi\texttt{(}e_1,\ldots,e_n\texttt{)} \mid X \mid \mathit{pat}\texttt{=}e \mid \texttt{self} \mid \\ & & e_1\texttt{,}e_2 \mid e_1\texttt{!}e_2 \mid \texttt{case}\ e\ \texttt{of}\ m\ \texttt{end} \\ & & \texttt{receive}\ m\ \texttt{end} \mid \texttt{spawn (}f\texttt{,}e\texttt{)} \\ m & ::= & \mathit{pat}_1\texttt{->}\ e_1\texttt{;}\ldots\texttt{;}\mathit{pat}_n\texttt{->}\ e_n \\ \mathit{pat} & ::= & c\texttt{(}\mathit{pat}_1\texttt{,}\ldots\texttt{,}\mathit{pat}_n\texttt{)} \mid X \end{array}$$

These are the known Erlang constructors. To express the semantics we need to define also constructors for pattern matching but these we treat differently as they are only needed as an "internal representation" of the most naturally expressed pattern matching expressions used in (Core) Erlang. We distinguish three different cases for matching `match`, `casematch`, and `queuematch` for simple patterns, patterns in case statements and in queues, respectively. The semantics of these patterns is given in the Appendix.

4.2. Core Erlang Semantics

Next, we introduce the operational semantics of the language Core Erlang. Usually, that is, in all the major introductory texts, e.g. [2], [1], but also in scientific descriptions of Core Erlang, e.g. [6], [25] the semantics of Erlang is just informally described. For our purposes, this is not sufficient. Hence we need to re-engineer a formal semantics. We do so closely following Huch [17] but significantly simplifying this original contribution as we do not need the same level of detail.

The semantics, we present, follows the idea of a *structural operational semantics* which means that we define a reduction relation $\longrightarrow_{\mathit{Erl}}$ that gives rules for all cases of possible syntax structures of Core Erlang programs. The reduction represents possible one-step evaluations of an Erlang program. The parallel program state is represented as a set of Erlang processes Π where each single process term is a triple (p, t, q) where p is the process identifier of this process, t is the sequential Erlang term representing the process in its current state of computation, and q is the current message queue of the process. The entire operational semantics is presented in Table 1. We will explain the meaning of these rules, including special notation we use in the following explanations.

We use reduction contexts E [11] to have a succinct representation of the semantic rules and enable determining a reduction strategy. Defining the reduction contexts as follows, we define the operational semantics as a leftmost innermost operational semantics.

$$\begin{array}{rcl} E & ::= & [] \mid \phi\texttt{(}v_1,\ldots,v_i,E,e_{i+2},\ldots,e_n\texttt{)} \mid E\texttt{,}e \mid \mathit{pat}\texttt{=}E \\ & & \texttt{spawn(}f\texttt{,}E\texttt{)} \mid E\texttt{!}e \mid v\texttt{!}E \mid \texttt{case}\ E\ \texttt{of}\ m\ \texttt{end} \end{array}$$

These reduction contexts are used in the rules to identify where a reduction may take place. The nonterminal [] is called the "hole" and marks the point of the next computation. We write $E[e]$ for the context E where the hole is replaced by the term e that is reduced next. In each of the reduction rules we reduce the set of current Erlang processes Π. We introduce $\Pi\dot{\cup}(p,t,q)$ as a map represented by triples: $\Pi\dot{\cup}(p,t,q) = \Pi \cup (p,t,q)$ if $\forall\ x,y.\ (p,x,y) \notin \Pi$ else Π.

The sequential reduction rules define the evaluation inside Erlang processes. The first one specifies that a sequence of terms, juxtapositioned by comma, $v\texttt{,}e$ reduces to the second argument, if the first one is a reduced value v. Next, a sequential term can be evaluated by replacing the semantics F_A of a predefined function F by its definition. The constant `self` may be replaced by the process identifier p representing the current process. Note, that – by using `self` and assigning this semantics to it – we do not need an explicit recursion in the semantics, instead iteration is mapped to invocation of processes. Finally, the sequential rules define that a function can be replaced by its function body.

The next set of three rules defines how matchings are evaluated; the functional definitions of the matching function are contained in the Appendix. Their definitions are quite complex and for the current context it suffices to anticipate the intuitive meaning of the matching results. A simple `match` leads to a pointwise match ρ of all variables in v according to pattern *pat* and can be applied to the entire sequential program $E[v]$. Similarly, a list of matches in a `casematch` results in a matching case i and a match ρ that can be applied in a `case` clause to evaluate the

SEQUENTIAL REDUCTION

$$\frac{}{\Pi\dot{\cup}(p, E[v\texttt{,}\ \ e], q) \longrightarrow_{Erl} \Pi\dot{\cup}(p, E[e], q)} \qquad \frac{F \textit{ predefined function}}{\Pi\dot{\cup}(p, E[F(v_1, \ldots, v_n)], q) \longrightarrow_{Erl} \Pi\dot{\cup}(p, E[F_A(v_1, \ldots, v_n)], q)}$$

$$\frac{}{\Pi\dot{\cup}(p, E[\texttt{self}], q) \longrightarrow_{Erl} \Pi\dot{\cup}(p, E[p], q)} \qquad \frac{f\texttt{(}X_1\texttt{,}\ \ldots\texttt{,}X_n\texttt{)}\ \texttt{->}\ e. \in \textit{program}}{\Pi\dot{\cup}(p, E[f\texttt{(}v_1\texttt{,}\ \ldots\texttt{,}v_n\texttt{)}, q) \longrightarrow_{Erl} \Pi\dot{\cup}(p, E[e[v_1/X_1\texttt{,}\ \ldots\texttt{,}v_m/X_m]], q)}$$

MATCHING

$$\frac{\texttt{match}(\textit{pat,v}) = \rho}{\Pi\dot{\cup}(p, E[\textit{pat} = v], q) \longrightarrow_{Erl} \Pi\dot{\cup}(p, \rho(E[v]), q)} \qquad \frac{\texttt{casematch}((\textit{pat}_1, \ldots, \textit{pat}_m), v) = (i, \rho)}{\Pi\dot{\cup}(p, E[\texttt{case}\ v\ \texttt{of}\ \textit{pat}_1\ \texttt{->}\ e_1\texttt{;}\ \ldots\texttt{;}\textit{pat}_m\ \texttt{->}\ e_m\ \texttt{end}], q) \longrightarrow_{Erl} \Pi, (p, \rho(E[e_i]), q)}$$

$$\frac{\texttt{queuematch}((\textit{pat}_1, \ldots, \textit{pat}_m), (v_1, \ldots, v_u)) = (i, j, \rho)}{\begin{array}{l}\Pi\dot{\cup}(p, E[\texttt{receive}\ \textit{pat}_1\ \texttt{->}\ e_1\texttt{;}\ \ldots\texttt{;}\ \textit{pat}_m\texttt{->}\ e_m\ \texttt{end}], (v_1, \ldots, v_j, \ldots, v_u)) \\ \longrightarrow_{Erl} \Pi\dot{\cup}(p, \rho(E[e_i]), (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_u))\end{array}}$$

CONCURRENT REDUCTION

$$\frac{f\texttt{(}X_1\texttt{,}\ \ldots\texttt{,}X_n\texttt{)}\ \texttt{->}\ e. \in \textit{program} \text{ and } p'\text{a new pid}}{\begin{array}{l}\Pi\dot{\cup}(p, E[\texttt{spawn(}f, [v_1, \ldots, v_n]\texttt{)}], q) \\ \longrightarrow_{Erl} \Pi\dot{\cup}(p, E[p'], q), (p', e[v_1/X_1, \ldots, v_n/X_n], ())\end{array}} \qquad \frac{v_1 = p' \in \textit{Pid}}{\Pi\dot{\cup}(p, E[v_1\texttt{!}v_2, q), (p', e, q') \longrightarrow_{Erl} \Pi\dot{\cup}(p, E[v_2], q), (p', e, q'\texttt{@}[v_2])}$$

Table 1. Operational semantics of Core Erlang

case construct while simultaneously replacing the matching values in e_i. The queuematch, finally, produces – besides the matching – the selection of the right clause i in a receive construct and the selection of the message v_j in the message queue of the process which is subsequently removed from it.

For the parallel semantics, we define the semantic rules CONCURRENT REDUCTION that entail a rule for spawning a new process and one for message dispatch with ! between processes where @ denotes list append.

4.3. Security Type System

Given the operational semantics of the Core Erlang language we can now define a type system assigning security types $\tau \in \{L, H\}$ to Core Erlang program terms. This will then enable to prove that there are no illicit information flows in well-typed Core Erlang programs. However, this statement will be dependent on the proper behaviour of the program according to the semantics. As Erlang is not a strongly typed language we would need to first provide a classical type system and prove type safety, i.e. progress and preservation for this classical type system. As this would go well beyond the scope of this paper and, moreover, such classical type systems are provided by others, e.g. [20], [27], we will just use these results without further introduction by implicitly assuming the additional hypothesis, "program p is classically well-typed".

Our security type system is a language based type system assigning L (low) and H (high) security types to terms of Core Erlang programs. We want to assign these classes to the parameters and results of Erlang processes in order to avoid explicit flows from H to L that may happen when a H process replies to a L process with ! or when a H process spawns a process that is of class L passing H values as parameters. To forbid such flows, we can guard the corresponding program terms by type constraints to expressions, like process identifiers, and to parameters. In addition, we use the subtype relation $\leq$, i.e. $L \leq H$. To type entire terms – like a function that is being spawned – we introduce the type constructor *cmd*. Now, programs can have type H *cmd* or L *cmd*; intuitively, a program of type H *cmd* cannot transmit information to L processes. For example, if a function has type H *cmd* it can only be applied to parameters of type H. The case of a function $f : H$ *cmd* applied to an L parameter is possible because – due to subtyping $L \leq H$ – the parameter is also of type H. How do we exclude the forbidden case $f : L$ *cmd* and parameter of type H despite subtyping? In our type system, the type τ *cmd* is neither contravariant – as in other type systems [33], [5] – nor covariant; it is simply not related. Still, the flow from L to H is enabled – as pointed out above, and the flow from H to L disabled because an L *cmd* cannot be upgraded (as it would be the case with covariance). However, as Erlang is a functional language – in contrast to the simple imperative languages considered in [33], [5] – we need to be able to use arbitrary expressions as arguments. That is, terms can be supplied as arguments to functions or sent to processes. To accomodate the necessary transformation of terms of type τ *cmd* to argument terms of type τ we add a type transformation rule (C-VAL).

The typing of identifiers, like variables and simple terms, and processes is encoded in *type maps* γ for identifiers and Γ_{proc} assigning a security type to each process. More precisely, we introduce different categories of types; the base types τ are used to assign security types to variables and constants, and the constructed types τ *cmd* to assign types to terms.

$$\begin{array}{rcl} \tau & ::= & L \mid H \\ \varsigma & ::= & \tau \mid \tau\ cmd \\ \gamma & : & Id \Rightarrow \tau \\ \Gamma_{proc} & : & Pid \Rightarrow \tau\ cmd \end{array}$$

The typing rules define inductively the typing relation $\Gamma_{proc},\ \gamma \vdash t : \rho$ where Γ_{proc}, γ are *type environments*, maps assigning types to base terms under which the actual typing $t : \rho$ is valid. Table 2 summarizes the entire type system; it is explained in the following in detail. In the type rules, we use the type variable ς as a "meta"-variable, i.e. $\varsigma \in \{L, H, L\ cmd, H\ cmd\}$. We further use the following operator $\overset{\circ}{\cup}$ that enables extension of type maps according to equalities induced by pattern matches,

$$\tau \overset{\circ}{\cup} \rho = \tau \cup (\tau \circ \rho)$$

where the operator $\circ$ is relational composition. Rules IDENT and PIDENT state that the typings encoded in the type maps Γ_{proc} and γ can be transformed into typings with $\vdash$. Simple terms may have some fixed type assigned to it – as is encoded in rule SIMPLE TERMS. Together, the first three rules provide the means to input a security policy to the type analysis. The following three rules COMPOSE, FUNCTION, and APPLICATION encode that arguments of suitable type can be plugged into the term constructors for composition, function definition and application of Erlang. The next three rules PATTERN, CASE, RECEIVE all deal with pattern matching. The assumption set ρ assigns variables to their matched terms which is included by $\overset{\circ}{\cup}$ into the type maps. If the patterns have suitable type τ, the statement using the pattern can be typed correspondingly by τ *cmd* because it will not contain a subterm lower than τ. The rules SPAWN and SEND control the communication: SPAWN demands that the arguments to a `spawn` conform to the security bound of the spawned function and SEND states that a process v_1 of level τ *cmd* can receive messages of type τ (or higher, because of rule SUBTYPE). The rule CONFIGURATION is the main rule that lifts the typing of the rules to the level of a configuration by dividing the typing to single process terms. Finally the rules BASE, REFL, C-VAL, and SUBTYPE determine the ordering on the set of types, transform *cmd*-terms to simple terms to allow terms as parameters to functions and as messages, and enable subsumption, i.e. elements of a type are also elements of its supertypes.

4.4. Indistinguishability

We provide key properties that will be used in the subsequent section to prove that well-typed programs have the noninterference property. In addition, we use this section to introduce the notion of indistinguishability that is the core of the definition of noninterference.

Lemma 4.1 (Confinement): Let $\vdash \Pi : \Gamma_{proc}$ and $(p, t, q) \in \Pi$ with $\Gamma_{proc}(p) = Hcmd$. Then, $\Gamma_{proc}, \gamma \vdash t : H\ cmd$ and for all subterms t_0 and corresponding contexts E with $t = E[t_0]$ we have that $\Gamma_{proc}, \gamma \vdash t_0 : H\ cmd$, i.e. all subexpressions are H as well.

Proof: By induction over the typing rules on the structure of t. □

The following theorem is a sanity check; it shows that the types are preserved by the semantics. If they were not preserved any properties encoded in the types would be destroyed.

Theorem 1 (Subject Reduction):

$$\vdash \Pi : \Gamma_{proc} \wedge \Pi \longrightarrow_{Erl} \Pi' \Rightarrow \exists\ \Gamma'_{proc}.\ \vdash \Pi' : \Gamma'_{proc}$$

where $\Gamma_{proc} \subseteq \Gamma'_{proc}$.

Proof: The proof is by induction and a straightforward, albeit long, case analysis. □

Next, we introduce a notion of equivalence of configurations that will enable the definition of noninterference. As already explained in the previous section when introducing noninterference informally, the intuition is that an attacker cannot see any difference of H values when regarding L values. This equivalence needs to be shown over arbitrary program runs. Hence, we need to define what it means for two program states to be indistinguishable for the attacker. As we have a dynamic set of processes, represented by Pid, that needs to be compared we use a technique seen in [3] that uses typed bijections to that end.

Definition 4.2 (Typed Bijection): A typed bijection is a finite partial function σ on process identifiers p such that

$$\forall p : \mathrm{dom}(\Pi).\ \vdash p : T \Rightarrow\ \vdash \sigma(p) : T$$

(where T is given by $\Gamma_{proc}(p)$).

The intuition behind typed bijections is that $\mathrm{dom}(\sigma)$ designates all those processes that are, or have been, visible to the attacker. We cannot assume the names in different runs of programs, even for low elements, to be the same. Hence, we relate those names via a pair of bijections. These bijections are typed because they relate processes that are all of type L.

The following definition of indistinguishability uses the typed bijection in this sense. The intuitive relationship between type L and membership in $\mathrm{dom}(\sigma)$ is only later made formal by an invariant. We believe that this invariant decisively ameliorates the exposition of the proofs and the understanding of the model (compare with the proofs in Banerjee and Naumann's paper [3]).

IDENT
$$\frac{\gamma(x) = \varsigma}{\Gamma_{\mathit{proc}}, \gamma \vdash x : \varsigma}$$

PIDENT
$$\frac{\Gamma_{\mathit{proc}}(p) = \tau\ \mathit{cmd}}{\Gamma_{\mathit{proc}}, \gamma \vdash p : \tau\ \mathit{cmd}}$$

SIMPLE TERMS
$$\Gamma_{\mathit{proc}}, \gamma \vdash e : \tau$$

COMPOSE
$$\frac{\Gamma_{\mathit{proc}}, \gamma \vdash c_1 : \tau\ \mathit{cmd}\ ,\ \Gamma_{\mathit{proc}}, \gamma \vdash c_2 : \tau\ \mathit{cmd}}{\Gamma_{\mathit{proc}}, \gamma \vdash c_1\texttt{,}\ c_2 : \tau\ \mathit{cmd}}$$

FUNCTION
$$\frac{\Gamma_{\mathit{proc}}, \gamma \cup \{X_1 : \tau, \ldots, X_n : \tau\} \vdash e : \tau\ \mathit{cmd},}{\Gamma_{\mathit{proc}}, \gamma \vdash f\texttt{(}X_1\texttt{,}\ldots\texttt{,}X_n\texttt{)}\ \texttt{->}\ e. : \tau\ \mathit{cmd}}$$

APPLICATION
$$\frac{\Gamma_{\mathit{proc}}, \gamma \vdash f\texttt{(}X_1\texttt{,}\ldots\texttt{,}X_n\texttt{)}\ \texttt{->}\ e. : \tau\ \mathit{cmd},\ \forall\ i.\ \Gamma_{\mathit{proc}}, \gamma \vdash v_i : \tau}{\Gamma_{\mathit{proc}}, \gamma \vdash e[v_1/X_1, \ldots, v_n/X_n] : \tau\ \mathit{cmd}}$$

PATTERN
$$\frac{\Gamma_{\mathit{proc}}, \gamma \vdash v : \tau,\ \Gamma_{\mathit{proc}}, \gamma \overset{\circ}{\cup} \texttt{match}(\mathit{pat}, v) \vdash \mathit{pat} : \tau\ \mathit{cmd}}{\Gamma_{\mathit{proc}}, \gamma \vdash \mathit{pat}\texttt{=}v : \tau\ \mathit{cmd}}$$

CASE
$$\frac{\begin{array}{c}\Gamma_{\mathit{proc}}, \gamma \vdash v : \tau,\ \texttt{casematch}((\mathit{pat}_1, \ldots, \mathit{pat}_m), v) = (i, \rho),\\ \Gamma_{\mathit{proc}}, \gamma \overset{\circ}{\cup} \rho \vdash \mathit{pat}_i\texttt{=}v : \tau\ \mathit{cmd},\ \Gamma_{\mathit{proc}}, \gamma \overset{\circ}{\cup} \rho \vdash e_i : \tau\ \mathit{cmd},\ \forall i \leq m\end{array}}{\Gamma_{\mathit{proc}}, \gamma \vdash \texttt{case}\ v\ \texttt{of}\ \mathit{pat}_1\ \texttt{->}\ e_1\texttt{;}\ldots\texttt{;}\mathit{pat}_m\ \texttt{->}\ e_m\ \texttt{end} : \tau\ \mathit{cmd}}$$

RECEIVE
$$\frac{\begin{array}{c}\texttt{queuematch}((\mathit{pat}_1, \ldots, \mathit{pat}_m), (v_1, \ldots, v_u)) = (i, j, \rho),\\ \Gamma_{\mathit{proc}}, \gamma \overset{\circ}{\cup} \rho \vdash \mathit{pat}_i\texttt{=}v_j : \tau\ \mathit{cmd},\ \Gamma_{\mathit{proc}}, \gamma \overset{\circ}{\cup} \rho \vdash v_j : \tau,\ \Gamma_{\mathit{proc}}, \gamma \overset{\circ}{\cup} \rho \vdash e_i : \tau\ \mathit{cmd},\ \forall i \leq m\end{array}}{\Gamma_{\mathit{proc}}, \gamma \vdash \texttt{receive}\ \mathit{pat}_1\ \texttt{->}\ e_1\ \texttt{;}\ldots\texttt{;}\mathit{pat}_m\texttt{->}\ e_m\ \texttt{end} : \tau\ \mathit{cmd}}$$

SEND
$$\frac{\Gamma_{\mathit{proc}}(v_1) = \tau\ \mathit{cmd},\ \Gamma_{\mathit{proc}}, \gamma \vdash v_2 : \tau}{\Gamma_{\mathit{proc}}, \gamma \vdash v_1\texttt{!}v_2 : \tau\ \mathit{cmd}}$$

SPAWN
$$\frac{\Gamma_{\mathit{proc}},\ \gamma \vdash f\texttt{(}X_1\texttt{,}\ldots\texttt{,}X_n\texttt{)}\ \texttt{->}\ e. : \tau\ \mathit{cmd},\ \forall\ i.\ \Gamma_{\mathit{proc}},\ \gamma \vdash v_i : \tau}{\Gamma_{\mathit{proc}},\ \gamma \vdash \texttt{spawn(}f, [v_1, \ldots, v_n]\texttt{)} : \tau\ \mathit{cmd}}$$

CONFIGURATION
$$\frac{\forall\ (p, t, q) \in \Pi.\ \Gamma_{\mathit{proc}}, \varnothing \vdash t : \Gamma_{\mathit{proc}}(p)}{\vdash \Pi : \Gamma_{\mathit{proc}}}$$

BASE
$$L \leq H$$

REFL
$$\varsigma \leq \varsigma$$

C-VAL
$$\frac{\Gamma_{\mathit{proc}}, \gamma \vdash e : \tau\ \mathit{cmd}}{\Gamma_{\mathit{proc}}, \gamma \vdash e : \tau}$$

SUBTYPE
$$\frac{\Gamma_{\mathit{proc}}, \gamma \vdash p : \varsigma, \varsigma \leq \varsigma'}{\Gamma_{\mathit{proc}}, \gamma \vdash p : \varsigma'}$$

Table 2. Security type system for Core Erlang

Definition 4.3 (Indistinguishability): An indistinguishability relation is a heterogeneous relation $\sim_\sigma$, parameterized by a typed isomorphisms σ whose differently typed subrelations are as follows.

$$\begin{array}{rcl}
t \sim_\sigma t' & \equiv & t_\sigma = t' \\
p \sim_\sigma p' & \equiv & \sigma(p) = p' \\
(p, t, q) \sim_\sigma & \equiv & p \sim_\sigma p' \wedge t \sim_\sigma t' \wedge \\
(p', t', q') & & \forall\, i. q\#i \sim_\sigma q'\#i \\
 & & \mathrm{dom}(\sigma) \subseteq \mathrm{dom}(\Pi_0) \wedge \\
\Pi_0 \sim_\sigma \Pi_1 & \equiv & \mathrm{ran}(\sigma) \subseteq \mathrm{dom}(\Pi_1) \wedge \\
 & & \forall\, p, p'.\ p \sim_\sigma p' \Rightarrow \Pi_0(p) \sim_\sigma \Pi_1(p')
\end{array}$$

The above exploits the convention that equations involving partial functions are interpreted as false when the function is undefined. Hence, $p \sim_\sigma p'$ only if $(p, p') \in \sigma$, otherwise $\sigma(p) = \mathit{false}$.

The H part of the program is not relevant for L-indistinguishability and thus not recorded at all in the corresponding typed bijections. "H-indistinguishability" thus corresponds intuitively to "indistinguishability not defined".

The partial bijection approach is an elegant concept for specification but technically proofs become cluttered with technical detail. We explicitly mark the correspondence between type L and the domain of the isomorphism σ. The following invariant specifies this correspondence.

Definition 4.4 (Invariant):

$$p \in \mathrm{dom}(\sigma) \quad \equiv \quad \Gamma_{\mathit{proc}}(p) = L\ \mathit{cmd}$$

We write *invariant*(σ) if the configurations are clear from context.

The invariant immediately transfers to the range of σ because it is a typed bijection.

Corollary 4.5: If the invariant holds we also have the following equivalence.

$$p \in \mathrm{ran}(\sigma) \quad \equiv \quad \Gamma_{\mathit{proc}}(p) = L\ \mathit{cmd}$$

Note, that the Invariant only *specifies* this correspondence. The invariant is a tool to clarify the proof of noninterference. Its validity for given typings and pairs of configurations has to be established.

Lemma 4.6 (Initial Invariant): Given two indistinguishable initial configurations Π_0, Π'_0 that are well-typed, the isomorphism σ can be constructed such that the invariant holds. These initial configurations are then indistinguishable with respect to σ, i.e. $\Pi_0 \sim_\sigma \Pi'_0$

Note, that if the initial configurations were not indistinguishable, their types could be different in which case the existence of a pair of isomorphisms fulfilling the invariant would be impossible.

4.5. A Well-typed Program is Secure

After these preparations we are able to prove the main theorem. Well-typed programs are secure. This property is a so-called *bisimulation* property, that is, a property over different runs of a program showing that a relation is preserved by the evaluation. This property will be the indistinguishability relation. The essence of the property is that the evaluation does not change the L-indistinguishability, therefore the attacker cannot learn more by observing the program running if he could not learn anything from start.

We first prove a theorem that assumes the invariant to hold and then shows that indistinguishability is preserved. The main result is a simple corollary from this: as we can chose σ to verify the invariant for initial configurations, all reachable configurations are secure. We prove a strong version of bisimulation in which the second transition is $\longrightarrow^{01}_{\mathit{Erl}} = \mathit{id} \cup \longrightarrow_{\mathit{Erl}}$ and not just $\longrightarrow^{*}_{\mathit{Erl}}$ (as, for example in Volpano and Smith's work on noninterference of a simple multi-threaded while language [33]).

Theorem 2 (Noninterference): Let Π_0 and Π_1 be configurations such that $\Pi_0 \sim_\sigma \Pi_1$, $\vdash \Pi : \Gamma_{\mathit{proc}}$ and $\vdash \Pi' : \Gamma_{\mathit{proc}}$, and the Invariant holds. If $\Pi_0 \longrightarrow_{\mathit{Erl}} \Pi'_0$ then there exists Π'_1 such that $\Pi_1 \longrightarrow^{01}_{\mathit{Erl}} \Pi'_1$ and $\Pi'_0 \sim_{\sigma'} \Pi'_1$ such that $\sigma \subseteq \sigma'$, and the invariant remains valid for σ'.

Proof: We proceed by case analysis and induction over the semantics $\longrightarrow_{\mathit{Erl}}$. In each case, we define a new σ' based on the existing σ for which the invariant holds by assumption. The case analysis hinges on $p \in \mathrm{dom}(\sigma)$ rather than L and H as in classical proofs, e.g. [33] (however, it is important to keep in mind that this predicate corresponds to H/L-typing in form of the proof invariant).

The **high** case is proved once for all cases of the semantic reduction. Let $p_0 \in \mathrm{dom}(\Pi_0)$ and $p_0 \notin \mathrm{dom}(\sigma)$ with $\Pi_0(p_0) \neq \Pi'_0(p_0)$, i.e. this process has been reduced. Let $\sigma' = \sigma$ and $\Pi'_1 = \Pi_1$. Then, $\Pi'_0 \sim_\sigma \Pi'_1$ because $\mathrm{dom}(\sigma' = \sigma) \subseteq \mathrm{dom}(\Pi_0) \subseteq \mathrm{dom}(\Pi'_0)$. The new process that may have been introduced – in case the reduction was with rule SPAWN – is H since, by the Invariant, from $p_0 \notin \mathrm{dom}(\sigma)$ follows that $\Gamma_{\mathit{proc}}(p_0) = H\ \mathit{cmd}$. In turn, by preservation, the new process has type $H\ \mathit{cmd}$ whereby the Invariant remains valid and indistinguishability as well.

The other case $p \in \mathrm{dom}(\Pi_0)$ such that $p \in \mathrm{dom}(\sigma)$ and $\Pi_0(p_0) \neq \Pi'_0(p_0)$ entails the **low** cases which are proved case by case following the semantics. Generally, we know – since $p \in \mathrm{dom}(\sigma)$ – that $\Gamma_{\mathit{proc}}(p) = L$ and that $\sigma(p_0) = p_1$ for some $p_1 \in \mathrm{dom}(\Pi_1)$. Furthermore, $\Gamma_{\mathit{proc}_\sigma}(p_1) = L$ because σ preserves types. We show the case for `spawn` as it is one of the more interesting cases where a new process is added and consequently changes appear. The other cases are very similar and are omitted.

Case (SPAWN). Let f be defined in the program, $\Pi_0(p_0) = (p, E[\mathtt{spawn(}f, [v_1, \ldots, v_n]\mathtt{)}], q)$, and p'_0 a new pid, then $\Pi'_0(p_0) = (p_0, E[p'_0], q_0)$ and $\Pi'_0(p'_0) = (p'_0, e[v_1/X_1, \ldots, v_n/X_n], ())$ according to the semantic rule SPAWN. Since we consider $p_0 \in \mathrm{dom}(\sigma)$, which is, due to the Invariant, equivalent to $\Gamma_{\mathit{proc}}(p_0) = L$, thus by rule CONFIGURATION $\gamma(t) = L\ \mathit{cmd}$, and, consequently, by confinement, we have that $\gamma(\mathtt{spawn(}f, [v_1, \ldots, v_n]\mathtt{)}) = L\ \mathit{cmd}$. Since $p_0 \in \mathrm{dom}(\sigma)$, there is a unique p_1 with $p_0 \sim_\sigma p_1$ and by assumption $\Pi_0(p_0) \sim_\sigma \Pi_1(p_1)$, hence, for $\Pi_1(p_1) = (p_1, t_1, q_1)$, we have $t_0 \sim_\sigma t_1$ and thus, by definition of indistinguishability, $t_1 = E_\sigma[\mathtt{spawn(}f, [v_{1_\sigma}, \ldots, v_{n_\sigma}]\mathtt{)})]$. We can select, $\Pi'_1 = \Pi_1 \dot\cup (p_1, E[p'_1], q_1) \dot\cup (p'_1, e[v_1/X_1, \ldots, v_n/X_n], ())$ where p'_1 is a fresh pid. Then $\Pi_1 \longrightarrow_{\mathit{Erl}} \Pi'_1$ according to rule SPAWN as well. According to preservation, the successor configurations are well-typed with types $\Gamma_{\mathit{proc}} = \Gamma_{\mathit{proc}} \dot\cup (p_0, L\ \mathit{cmd})$ by typing rule SPAWN. The new processes p'_0 and p'_1 have type $L\ \mathit{cmd}$ by confinement and rule CONFIGURATION; The new isomorphism $\sigma' := \sigma \cup \{(p'_0, p'_1)\}$, whereby the invariant remains valid. Finally, we see that $\Pi'_0 \sim_\sigma \Pi'_1$. □

Corollary 4.7 (Reachable Configurations Security): Let Π_0 and Π_1 be configurations reachable from some initial indistinguishable configurations then there exist σ such that $\Pi_0 \sim_\sigma \Pi_1$.
The corollary follows by induction over $\longrightarrow_{\mathit{Erl}}$ from Lemma 4.6 and the Noninterference Theorem.

4.6. Experimental Evaluation

The security type system we have constructed for the Erlang language shall guarantee statically that there are no illegal information flows from H to L. How is this achieved? We finally want to illustrate here the use of the concepts developed in the current section by some simple experiments. Going back to the informal security analysis from where we started in Section 3.3, we reconsider the argument we pointed out there and show up how the Erlang type system we presented in Section 4.3 enables checking whether a given security policy is valid for the privacy concerns of our data base enquiry program.

Let us reconsider the critical case of the informal analysis. There, we found out that – if we want to keep the variable `Key` private, i.e. H – then `Res` and `Any` (at least) have to be assigned H by the security policy as well. Technically, this is realized by setting up the variable assignment γ as mapping those variables to H. As we have seen before, `Any` and `Res` had to be set to H as a consequence of an implicit flow in the program – both variables depend on `Key`. To test our Erlang noninterference type system let us see what would happen if we were to set these two variables to L.

Type inference is a process of iteratively applying the typing rules from Table 2 starting from the initial assignment given by the security policy, here

$$\gamma_{\mathit{init}} = \{\mathtt{Key} \mapsto H, \mathtt{Val} \mapsto L, \mathtt{Any} \mapsto L, \mathtt{Res} \mapsto L\}$$

reconstructing the entire program thereby constructing the security types – which in our case should be impossible.

Indeed, at some point during this reconstruction, we need to give a type for the crucial case statement in `client_eval`. The only applicable typing rule CASE imposes in its conclusion – below the line – that the case construct can only have a type τ *cmd* if for all matches this type τ *cmd* may be constructed for all variables and the branches e_i. However, as `Key` is contained in $\{$`Key, X`$\}$ this pattern pat_2 has type H *cmd* (rule MATCH). Hence, the branch e_2 after the `Key` clause must be of type H *cmd* which can only be inferred by rules SIMPLE TERMS, FUNCTION, APPLICATION if `Res` already has type H. Equally, `Any` must be H as well. These two constraints are *not* fulfilled in our initial assumption γ_{init}. This leads to a failure of type inference with our Erlang noninterference type system; the program is rejected for the security policy γ_{init}. A straightforward implementation of the type rules would automatically detect the error in the security policy. In fact, the constraints elaborated above by the walk-through of the type rules can only be fulfilled if the initial assignment of γ_{init} already marks `Res` and `Any` as H – corresponding to our informal security consideration in Section 3.3.

Interestingly, also `Val` – as variable v at the head position of the case statement – must be H according to rule CASE. This is an additional requirement that we have not established informally in Section 3.3. In our case, it is not necessary but, in general, the head position of a case statement can also be a H variable – then forcing the other match variables to be H as well. However, this additional requirement does obstruct neither the functioning of the program nor the correctness of the type analysis. It is a well-known problem that static security analysis is often too restrictive; in fact it must be so because it is provable that an accurate noninterference analysis is undecidable [34].

To ascertain that this over-functioning of the type system does not go too far (i.e. typing everything as H), let us consider just one more variable. To function well according to our global security policy – i.e. the server is public and only the client is our trusted component – the files, that are on the server and on which we want to perform our search and variables, e.g. `Bin`, that are related to those files, need to be assigned L in the security policy γ_{init}. Now, can `Bin` be L? This seems surprising as it is used in close context, i.e. direct pattern matching, with `Val` of which we have just seen that it needs to be H. The crucial pattern matching in the `receive` statement, i.e. term `Val = binary_to_term(Bin)` represents a *legal* information flow from L to H. This is reflected in our type system as follows. As `Bin` is of type L it is – according to rule SUBTYPE – also of type H. Hence, by rule PATTERN, the corresponding pattern matching `Val = binary_to_term(Bin)` can be typed H *cmd* because `Bin` can be typed H and `Val` can be typed H *cmd* : if we assume `Val` initially – in γ_{init} – to be of type H *cmd*, it also is of type H because of rule C-VAL – thus complying to the requirements of the CASE rule seen before.[2]

The proof of type safety and the Noninterference theorem generally guarantee what we have just illustrated by a walk-through of the crucial parts of our program. Any program that is accepted by the Erlang noninterference type system will not leak information from variables designated as H in the security policy. If we want to keep, like in our example, some variable `Key` private, the type system shows us which other parts of the program need to be kept confidential as well.

5. Alternative Approaches and Conclusions

In this final section, we briefly consider alternative approaches and discuss the solution. The approach we have presented – based on the simple idea of running the security-sensitive part of the service on the client site – corresponds to the concepts of common web service implementations like Javascript, Active Server Pages and Java Applets. However, other concepts for web services, like CGI-Scripts or Java Servlets run on the server site. Compared to the presented, secure, client-sited approach, running the entire service remotely is clearly more efficient. An open question is, whether we can provide a *secure* solution to this more efficient way of running our security sensitive key-application on a remote site. We illustrate this alternative approach next using a calculus for active objects. We then sum up relevant work in the Java sector, before we reach our general conclusions.

5.1. Implementation with Active Objects

By adding the concept of objects, familiar from object-oriented programming, to the existing concepts of parallelism and distribution as in a functional – and hence relatively safe – language like Erlang, we provide functional active objects through our new calculus ASP$_{\text{fun}}$ [16]. Active objects allow confidentiality by encapsulating local data. In contrast to data locality in a process, the idea of data encapsulation is stronger because the data is an inherent part of an *object*. Such objects are *activated* as a whole entity and become active objects. We can thus remotely activate such objects without the risk of disclosing the confidential data contained in the object.

Although we could also remotely start an Erlang function, we still have to transmit with this activation (or `spawn`) command the initial data for the process, in our case the key we wish to keep secret. Since we cannot assume that the communication channels are secure [31], confidentiality of the key, i.e. privacy, cannot be established.

We can implement a database search using active objects by starting an active object that acts as a kind of gateway

[2] Flows from H to L cannot be produced: if a left hand side is a pattern of type L *cmd* this has no relation to H *cmd*.

process. Given confinement of data in an active object, we also achieve privacy. It is beyond the scope of this paper to properly introduce ASP$_{\mathrm{fun}}$, but, as it is a simple and concise calculus, we still use it here to concretize this idea of a gateway object. We now give the – actually very short ASP$_{\mathrm{fun}}$ program – with just a very brief and informal explanation of its functionality.

Let Δ be the remote database we wish to search. The following short ASP$_{\mathrm{fun}}$ program, when run on a client, activates an object containing a search method σ and the key κ on a server.

```
Active([s = σ, k = κ]).s(Δ)
```

The `Active` command creates a new activity that contains our gateway object `[s = σ, k = κ]` as active object. The call of `s` with Δ as parameter to the then remotely active object – notated in the object-oriented fashion as `.s(Δ)` – initiates the search. The result of this method call is returned to the caller, here the client.

A security consideration for this program, or for ASP$_{\mathrm{fun}}$ in general, must assume that active objects are guaranteed to be confined. In other words, the contents of an active object can only be accessed through calls of corresponding methods. This is the principle of language based security. Although, in principle, any malicious remote run-time system can crash our active object and get access to its contents, language standards can guarantee that this will not happen. An example for such language standards is bytecode verification in virtual machines. Additionally, we need to ensure that the invocation of an active object, through the `Active` command, is secure. This means that the transfer of the object to a remote site is done in a secure fashion. Given such security measures, we can apply a security policy that permits only active objects with equal or higher security levels to call methods to an active object, thereby guaranteeing the exclusion of illicit flows.

5.2. Java Privacy Solutions

The server sided approach using active objects with ASP$_{\mathrm{fun}}$ we have illustrated in the previous section is also realized for the programming language Java. Andrew Myers has provided in his PhD thesis [22] an approach called Decentralized Label Model (DLM) that enables a rôle based approach to enforcing security in programs. The main idea in DLM is to have explicit labels in the program modelling the actors that have access to labelled program parts. Myers and Liskow augmented the DLM model with the idea of information flow control as described in the papers [23], [24]. Further works by Myers have been mostly practically oriented, foundations only considered later. Initially, he implemented a Java tool package called JFlow that implements the DLM and information flow control [21]. The extensions given by the DLM to the standard noninterference notions lead to undecidable typing. That is, typing cannot be completely performed at compile time but has to be partly done at run-time – which is costly and risky. From the point of view of privacy policies, an interesting experiment is the paper by Hayati and Abadi [13] in which they sketch how privacy policies can be expressed using the DLM. Hayati and Abadi use the Jif framework [26]– the new version of JFlow – to outline the Jif signatures to encode purpose and retention, two important notions for privacy policies. In more recent work, still along the same lines, Zheng and Myers [37] have gone even further in exploring the possibilities to dynamically assign security labels while still trying to arrive at static information flow control. A recent implementation of these concepts is given by the framework called Sif – for Servlet Information Flow – which basically represents a new version of Jif, i.e. Jif 3.0, but has an additional layer of Java servlets that focuses thus on the support of web-services with server-sided applications.

The impressive amount of work by Myers and his team pushes the idea of static analysis of information flow with type systems very far. The DLM concept enables even multilateral security modelling but causes troubles with decidability. However, being centred around the Java world, the work additionally has to struggle with concurrent access to data. Using active objects, as sketched on a conceptual level in the previous section, the clear separation of separate data spaces grants a simpler and more natural use of distribution in web-services.

5.3. Conclusions

Triggered by public demand, privacy in social networks – in particular internet based ones – increasingly attracts scientific interest. Leading European projects, like Primelife [29] and its predecessor Prime, bundle scientific and industrial competence to explore privacy in future networks and services. Various techniques in computing have been applied to tackle the related challenges, for example, artificial intelligence and unsupervised learning [8].

We have presented an approach to a privacy-enhancing data base enquiry that solves the problem by not disclosing the search key but by performing the search itself. The concept has been demonstrated by a simple implementation in Erlang; it's feasibility has been achieved through Erlang's high scale parallelism. A final security consideration shows, informally and formally, that the security goal of keeping the key data confidential, i.e. private, is achieved.

As briefly mentioned in Section 2.3, a sensible extension of our parallel search program is to evaluate in each step the results obtained in the previous step with respect to new goals, i.e. file names, to continue the search. This is a simple enough extension that merely needs to identify new file names, or more generally Internet sites, to continue the database enquiry. However, this kind of goal-directed search raises new security issues. An observer could infer some information about the key from the way we continue our search because we select the file names from the matched search results. To overcome this, we have to cover up our search and load, as before, all possible files referenced in the previous round. However, for the sake of efficiency, we would only really analyze those that we know to be interesting. Here, again, a new covert channel opens up, that is not visible in the data flow based model we use in Section 3: an attacker could distinguish our two ways of treating incoming files in this "cover-up" analysis by measuring the times between new demands for connections with `gen_tcp:send(..., FileS)`.

The informal security argument we have shown is not invalidated by this consideration. In the simple implementation, we have not used information from the files. There simply is no dependency between the files being downloaded. Consequently, we cannot lose any information from them. A similar analysis of the extended "sensible" search would reveal this illegal information flow because we would use information from files analyzed in previous rounds to determine the new `FileS`.

Our informal security analysis is first informally introduced and then developed into a formal notion of information flow for Erlang. We develop a security type system and prove that when a program is well-typed according to this type system then there are no (undesired) covert channels. This approach is widely accepted for language based security analysis; a concise overview of type-based information flow control is given by Sabelfeld and Myers [32]. The special challenge we have been facing here for Erlang is that we have to address concurrency. Earlier papers on security type systems [33], [5] for concurrent systems already point out the following difficulty: the termination of processes is observable by other processes. The motivating example [5] shows that this problem is due to concurrency of shared data. By contrast – as Erlang processes have strictly separate data spaces – these incriminating examples cannot be reproduced in our case; since active objects in ASP_{fun} also have separate data spaces this security problem is neither relevant there.

The construction of a security type system for Erlang and proof of noninterference for this type system – presented in this paper – provide a general security measure that can straightforwardly be implemented in a static type checker for Erlang security. Extending these ideas to active objects is current research.

Acknowledgements. We would like to thank Jeff Sanders for helping us to get the initial understanding and the anonymous referees for constructive criticism and valuable links to related approaches.

References

[1] J. Armstrong. *Programming Erlang – Software for a Concurrent World.* The Pragmatic Bookshelf, 2007.

[2] J. Armstrong, M. Williams, and R. Virding. *Concurrent Programming in Erlang.* Prentice-Hall, Englewood Cliffs, NJ, 1993.

[3] A. Banerjee and D. A. Naumann. Stack-based Access Control for Secure Information Flow. *Journal of Functional Programming* **15**(2), 2003.

[4] A. Barth and J. C. Mitchell. Enterprise Privacy Promises and Enforcement. *WITS'05.* ACM 2005.

[5] G. Boudol, I. Castellani. Noninterference for Concurrent Programs. *ICALP'01.* LNCS:**2076**, Springer, 2001.

[6] R. Carlsson. An introduction to Core Erlang. *Proceedings of the PLI'01 Erlang Workshop*, 2001.

[7] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan. Private Information Retrieval. *Journal of the ACM*, **45**(6): 965–982, 1998.

[8] G. Choral, E. Armengol,A. Fornells, and E. Golobardes. Data Security Analysis Using Unsupervised Learning and Explanations. *Innovations in Hybrid Intelligent Systems.* Advances in Soft Computing (ASC), **44**: 112–119, 2007.

[9] D. Denning. A Lattice Model of Secure Information Flow. *Communications of the ACM*, **19**(5): 236–242, 1976.

[10] D. E. Denning and P. J. Denning. Certification of programs for secure information flow. *Communication of the ACM*, **20**(7), 1977.

[11] M. Felleisen, D. P. Friedmann, E. Kohlbecker, and B. Dulba. A syntactic theory of sequential control. *Theoretical Computer Science,* **52**(3):205–237, 1987.

[12] J. Goguen and J. Meseguer. Security Policies and Security Models. *Symposium on Security and Privacy, SOSP'82*, pages 11–22. IEEE Computer Society Press, 1982.

[13] K. Hayati and M. Abadi. Language-based Enforcement of Privacy Policies. *Privacy Enhancing Techniques.* **3432**:302–313, 2005.

[14] C. A. R. Hoare. *Communicating Sequential Processes.* Prentice-Hall, 1984.

[15] M. Hennessy. The security pi-calculus and non-interference. *The Journal of Logic and Algebraic Programming.* **63**:3–34. Elsevier 2005.

[16] L. Henrio and F. Kammüller. Functional Active Objects: Typing and Formalisation. *8th International Workshop on the Foundations of Coordination Languages and Software Architectures, FOCLASA'09. Satellite to ICALP'09.* Proceedings to appear in ENTCS, Elsevier, 2009. Also invited for journal publication in Science of Computer Programming, Elsevier.

[17] F. Huch. Verification of Erlang Programs using Abstract Interpretation and Model Checking. *ACM SIGPLAN International Conference on Functional Programming (ICFP '99).* ACM Sigplan Notices, 34(9):261-272, 1999.

[18] F. Kammüller. Formalizing Non-Interference for A Small Bytecode-Language in Coq. *Formal Aspects of Computing*: **20**(3):259–275. Springer, 2008.

[19] F. Kammüller and R. Kammüller. Enhancing Privacy Implementations of Database Enquiries. *The Fourth International Conference on Internet Monitoring and Protection.* IEEE Computer Press, 2009.

[20] S. Marlow and P. Wadler. A practical subtyping system for Erlang. *ACM SIGPLAN Notices*, **32**(8):136–149, 1997.

[21] A. C. Myers. JFlow: Practical mostly-static information flow control. *26th ACM Symposium on Principles of Programming Languages, POPL'99.*

[22] A. C. Myers. *Mostly-Static Decentralized Information Flow Control*. PhD thesis, MIT, Cambridge, 1999.

[23] A. C. Myers and B. Liskov. Complete, safe information flow with decentralized labels. *IEEE Symposium on Security and Privacy*. 1998.

[24] A. C. Myers and B. Liskov. Protecting Privacy using the decentralized label model. *ACM Transactions on Software Engineering and Methodology*. **9**:410–442, 2000.

[25] T. Noll and C. K. Roy. Modeling Erlang in the Pi-Calculus. *Proceedings of the 2005 ACM SIGPLAN workshop on Erlang, Erlang'05.* ACM Press, 2005.

[26] A. C. Myers, L. Zheng, S. Zdancewic, S. Chong, N. Nystrom. *Jif: Java information flow.* Software release at `http://www.cs.cornell.edu/jif`. 2001.

[27] S.-O. Nyström. A soft typing system for Erlang. *Proceedings of the 2005 ACM SIGPLAN workshop on Erlang, Erlang'03.* ACM Press, 2003.

[28] L. Paulson. *ML for the Working Programmer*. Cambridge University Press, 1995.

[29] Primelife. Bringing sustainable privacy and identity management to future networks and services. EU research project in the 7th FP. `http://www.primelife.eu/`, 2009.

[30] M. O. Rabin. How to exchange secrets by oblivious transfer. *TR-81, Aiken CL, Harvard University*, 1981.

[31] K. Rikitake and K. Nakao. Application Security of Erlang Concurrent Systems. *Computer Security Symposium, CSS'08*. Okinawa, 2008.

[32] A. Sabelfeld and A. C. Myers. Language-Based Information-Flow Security. *Selected Areas in Communications*, **21**:5–19. IEEE 2003.

[33] G. Smith and D. Volpano. Secure Information Flow in a Multi-threaded Imperative Language. *POPL'98.* ACM 1998.

[34] D. Volpano and G. Smith. A Type-Based Approach to Program Security. *TAPSOFT'97.* LNCS **1214**, Springer 1996.

[35] D. Volpano, G. Smith, and C. Irvine. A Sound Type System for Secure Flow Analysis. *Journal of Computer Security*, **4**(3): 167–187, 1996.

[36] G. Wolf and A. Pfitzmann. Properties of Protection Goals and Their Integration into a User Interface. *Computer Networks*, **32**:(685–699), 2000.

[37] L. Zheng and A. C. Myers. Dynamic Security Labels and Static Information Flow Control. *International Journal of Information Security*. **6**(2–3), Springer, 2007.

Appendix

In the rules for pattern matching, `case`, and `receive`, we use the functions `match`, `casematch`, and `queuematch` for modelling the pattern matching mechanism of Erlang. This mechanism is more complicated than those of other programming languages because of non-linear patterns with multiple occurrences of the same variable.

$$\begin{array}{l}
\texttt{match}(X,t) := [t/X] \\
\texttt{match}(c(pat_1,\ldots,pat_n,c(v_1,\ldots,v_n)) := \\
\qquad \texttt{match}(pat_1,v_1) \cup \ldots \cup \texttt{match}(pat_n,v_n) \\
\texttt{match}(\_,\_) := \text{Fail, otherwise}
\end{array}$$

Two derived substitutions can only be unified if the overlapping parts are identical. Otherwise the matching fails.

$$\begin{array}{lcl}
\text{Fail} \cup \sigma & := & \text{Fail} \\
\sigma \cup \text{Fail} & := & \text{Fail} \\
\sigma_1 \cup \sigma_2 & := & \left\{ \begin{array}{l} \sigma_1 \cup \sigma_2,\ \text{if}\ \forall\, X \in (\mathrm{dom}(\sigma_1) \cap \mathrm{dom}(\sigma_2)). \\ \qquad\qquad \sigma_1(X) = \sigma_2(X) \\ \text{Fail, otherwise} \end{array} \right.
\end{array}$$

The function `case` evaluates to the expression corresponding to the first pattern matching a given value. The function `casematch` returns a tuple containing the number of the first matching pattern and the corresponding substitution, or Fail, if none of the patterns matches.

$$\begin{array}{l}
\qquad \texttt{casematch}((pat_1,\ldots,pat_n),v) \\
= \left\{ \begin{array}{l} (i,\rho),\ \text{if}\ \texttt{match}(pat_i,v) = \rho\ \text{and} \\ \qquad\qquad \texttt{match}(pat_j,v) = \text{Fail}\ \forall\, j < i \\ \text{Fail},\ \text{otherwise} \end{array} \right.
\end{array}$$

The constructor `receive` has the same behaviour but all values in the queue have to be considered. In Erlang, a pattern is successively matched against all values in the queue before the next pattern is matched. This is implemented in the function `queuematch` which returns the match and in addition the position of the queue value that matches.

$$\begin{array}{l}
\texttt{queuematch}((pat_1,\ldots,pat_n),(v_1,\ldots,v_n)) \\
= \left\{ \begin{array}{l} (i,j,p),\ \text{if}\ \ \texttt{match}(pat_i,v_j) = \rho\ \text{and} \\ \qquad\qquad \texttt{match}(pat_i,v_k) = \text{Fail}\ \forall\, k < j\ \text{and} \\ \qquad\qquad \texttt{match}(pat_l,v_h) = \text{Fail}\ \forall\, l < i,\ h \le n \\ \text{Fail},\ \text{otherwise} \end{array} \right.
\end{array}$$

An Integrated System for Intelligence, Surveillance and Reconnaissance (ISR)

Barbara Essendorfer, Eduardo Monari, Heiko Wanning
Fraunhofer IITB
{essendorfer, monari, wanning}@iitb.fhg.de

Abstract—**Connecting systems that are responsible for gathering information of any kind (e.g., images (optical/infra red/radar), video streams, vehicle tracks, etc.), processing stations (merging images series, stabilizing videos, etc.) and exploitation systems in a large (e.g., multinational) environment is a difficult task. Currently those stations tend to be operating independently irrespective of the vast amount of other systems available.**

As cooperation between different nations and different operating entities (civil and military) is increasingly important the sharing of information to generate a common awareness is vital.

This paper describes a solution towards this goal by introducing a system capable of dealing with the ingestion and distribution of data and the fusion perspectives/possibilities that arise when a multitude of stations report a single event from many viewpoints.

Keywords: client-server architecture, sensor/ exploitation/ information system, information/ data fusion, ISR

1. Introduction

ISR (Intelligence, Surveillance and Reconnaissance) in civil and military environments is often accomplished using separated stove-piped systems, therefore data accumulated from multiple heterogeneous sensors can't be shared and the capabilities to fulfill missions are limited. Services, like data/information collection and analysis, are not performed by the best but by the only system available, resource management is therefore suboptimal. To overcome this problem a concept to share data has been developed and implemented: A Coalition Shared Data (CSD) Server is used to store standardized data. It allows the dissemination according to user requirements, network and security settings. Tasking elements, sensor and exploitation systems store relevant metadata and products in common (standardized) formats in a shared database. The data model and interfaces of the database are based on established military standards (STANAGs). Instead of "only" collecting and distributing data an integrated data processing approach also should evaluate the incoming information, preselect interesting events for the human operator and, by using fusion algorithms, summarize complementary information and sort out redundant data to reduce the amount of information. This fusion aspect will be shown exemplary in the domain of image and video fusion.

The CSD concept was deployed in a rather restricted environment where legacy systems are in place that have to be reliable and secure. These systems are developed by different nations and by different companies. The challenge in such a context is "real" interoperability. To achieve this some constraints were put externally on properties of the final architecture (e.g. using STANAGs for the description of data, meta data and the client-server communication "language"), that our solution had no way to circumvent. Therefore the result and described work here is the output of those design decisions giving in our opinion the best possible solution solving the given problem under the mentioned restrictions. The design and implementation phases were and still are a long continuing process.

This paper is based on a conference paper that focused on border surveillance [1]. Here not only civil applications are taken into account, but the full aspect of civil and military ISR.

It is structured as follows: in Section 2 the task of information sharing within ISR is introduced and described. In the following section, the requirements for standardized data formats, which are highly important for information sharing, are discussed. Next, in Section 4, our system architecture for integrated ISR systems is presented. Finally, exemplary deployments of the developed system, which have been used in different military exercises and demonstrations on civil security are described.

2. Information Sharing within ISR

ISR are different aspects of information gathering with human and technical sources to enhance situation awareness. This happens on different organizational and administrative levels.

Within (civil) surveillance on a local level, immediate measures have to be found to enable a quick and efficient reaction to an imminent danger. On a regional level relevant information has to be shared to be able to exercise precautionary measures and avoid an escalation of a crisis. If the local authority is not able to handle the crisis, reinforcement has to be provided. If a crisis affects more than one region or even country then national and transnational decisions have to be made and information has to be shared.

Another (military) differentiation is information gathering on a tactical, functional or strategical level. On a tactical level decisions have to be made within operations based on information on current events and immediate decisions on an appropriate reaction have to be made (short term decisions). On a functional level situation awareness is generated to enable the planning of current military operations (mid term decisions). On a strategical level situation awareness is generated with security-relevant information to enable decision makers (on a political and military level) to predict long term developments in critical areas.

The use of a mix of sensor and information systems is key to adequate situation awareness on the different levels.

2.1 Integrated ISR Systems

Within an integrated ISR system, disparate technologies that complement one another are installed, the interaction of the data output is essential.

An integrated ISR system consists of sensors (technical systems and humans), exploitation systems (that might also be deployed as situational awareness displays) and external information systems.

In Figure 1 a critical area, e.g. a border is monitored by a range of different sensor types. Those sensors deliver data to a surveillance unit (SU). However, the areas that are monitored intersect and data that is of interest for one surveillance unit may also be of interest for adjacent units. Our architecture allows the necessary data sharing and accommodation of additional information from external systems resulting in enhanced situation awareness.

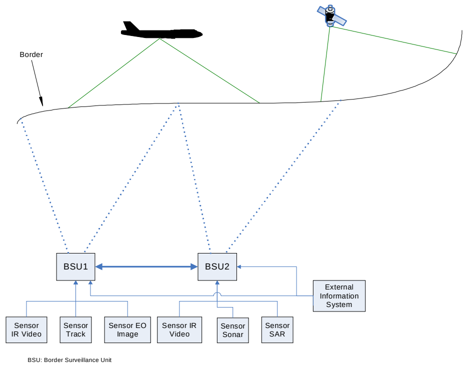


Figure 1: Surveillance System

2.1.1 Sensor Systems

Sensor systems normally consist of the sensor and a ground station that does the primary data processing and possibly some exploitation. Combined sensor systems that consist of different sensors might use some sensors as triggers for others and only the secondary information is passed on to an "outside" exploitation system. Depending on the sensor type and the processing a proprietary (raw) data stream may be created. To observe different types of critical areas it is necessary to make use of different sensor types with differing ranges and tasks [2]:

- Long-range surveillance conducted by space borne and airborne systems is of interest for an all-weather and 24 hour detection of threats that harm a wide area (e.g., oil slicks that indicate an attack on the environment and/or on nations resource supply). The sensors can deliver all kinds of imagery such as IR (infrared), EO (electro optical) and SAR (synthetic aperture radar) as well as motion imagery (video), SIGINT (signal intelligence) or radar data.
- Airborne sensors, including the use of balloons or zeppelins can be used for medium-range surveillance.
- Ground-based or seaborne sensors are mainly used for short-range surveillance. Real time information can be provided on critical areas, objects and people. Seaborne sensors can be installed above (cameras, radar) or under water (e.g., sonar, metal detection).

The display of sensor data in a common picture only makes sense if the operator/analyst is able to interpret that information correctly. Raw sensor data have to be interpreted by specialists. Therefore sensor data are only provided on system, local or at the most the regional level.

2.1.2 Exploitation Systems

Exploitation Systems are used for the exploitation of preproduced data. Exploitation can be done in different contexts and can be specific to the system, data type, area or task. For exploitation systems that work on products that are produced from multiple sensors it is important that data are available in an inter-coordinated data format (see section 4.2 for how it was implemented in our case). Exploited data normally already contain more enhanced information. Similar to the sensor data it has to be integrated adequately into a common picture. This type of information is of interest for upper decision bodies. Still some special expertise is needed to read and decide upon it. On a national level only the result of an analysis would be provided. The main effort on this level would be to fuse information from different sources. (more on fusion in section 3.3)

2.1.3 Information Systems

Information systems are relevant for the rating/ evaluation of derived data and information. Weather data can give essential advice which product sources are of interest in certain circumstances, systems such as the “Schengen Information System” (SIS) provide data on detected persons or goods and databases/information services freely available on the Internet can provide background information for all kinds of questions (provided it is understood that this is in general low grade intelligence). Public information sources are subject of data protection and the usage of this data has to be legally defined across borders. The system type, structure, language and concepts that are used within those systems differ from region to region and nation to nation. This is why an integration and combined usage of such system information is extremely complicated. Apart from legal regulations, aspects of intelligent data discovery and sharing are subject to research and discussion here.

2.2 Situation Awareness

Situation awareness means that threats and suspicious behavior have to be perceived, the threat has to be understood and an appropriate reaction has to be performed [3]. To perceive threats, products from different sources (information systems, sensors, exploitation systems) have to be available. The data has to be accessible with respect to time and location of the product as well as to other decision-relevant (e.g., urgency) information. Relevant sources of knowledge should be incorporated. Integrated systems achieve enhanced situation awareness by developing a common picture of the tasked area. To support analysts, operators and decision makers it is important to integrate the correct i.e. temporally relevant information in this common picture in a user-friendly manner [4].

2.3 Information Fusion

In general, the usage of multiple and heterogeneous sensors increases the possibilities and functionalities of the superior system and therefore improve the quality of the surveillance task. Especially for security applications (i.e. monitoring) state of the art multi-sensor-systems enable the surveillance of large areas with a reduced amount of manpower.

On the other hand, large sensor networks produce a huge amount of information that the analysts and operators have to monitor. Therefore, automatic data processing and information fusion is an essential part of large distributed and multimodal surveillance systems. The highest benefit of implementing automatic image processing is achieved when the automatism attracts the operator’s attention on relevant events and situations only. Hereby, integrated data processing approaches evaluate the incoming information and preselect interesting events for the human operator. Motion detection algorithms for example lead to a reduction of the operator’s workload by generating indications of movement automatically and therefore allow the human operator to focus on areas of interest only.

This clearly can be solved more robustly, when sensors of different modalities are incorporated (e.g., video cameras with a visual spectrum during the day and infrared spectrum at night). On the other hand, the usage of different sensor modalities requires the implementation of more complex schemes for information fusion. Implementing theses schemes will result in a higher level of information quality, since automatic sensor data processing and information fusion summarize complementary information and reduce the amount of redundant data.

3. Architectural Aspects of Information Sharing

As stated above, the surveillance of borders or areas of interest requires the co-ordination of many different agencies each with their own personnel, systems, assets and equipment.

The task of information sharing in a time sensitive domain places requirements on the architecture. To be able to share data and information some aspects have to be taken into account:

Information has to be reliable and secure: To be able to make the right decisions and react appropriately, information has to be reliable. Access to classified data has to be limited to entitled agencies and persons. To ensure the security of the data transfer and the protection against cyber attacks encryption and authentication techniques must be implemented. Also, user roles that include access levels can be used to restrict data access to authorized persons only. In addition it allows the display of only the relevant data to the right personnel (analyst, decision maker on a regional/national level etc.) avoiding information overload.

Information has to be provided in time and at the right place: An adequate data transmission network is required (i.e. distributed architecture and standardized mechanisms to access data). At the first step databases at different locations only synchronize their meta data (“video clip from location x, time y, sensor z, showing i”) and not the full data (the video file itself) as this would mean shifting unnecessary data loads through out the network causing congestions.

To integrate information systems with different semantics, intelligent methods of information retrieval have to be established. The annotation of information with metadata and ontologies enables semantic interoperability between the systems.

To be able to access data in time at the right place data transmission between more than two systems has to be enabled. This leads to a distributed architecture and standardized mechanisms to access data. Databases at different locations have to be able to synchronize their information, without synchronizing the data as this would mean to shift unnecessary data loads through the network.

To grant an easy and adaptable data access standardized data formats and data sharing mechanisms should be used. The data formats have to provide the right type of information which implies a detailed analysis of the application domain.

3.1 Data formats for the ISR domain

The previously described particularities of the domain make an adequate handling of the data necessary.

Surveillance has to be weather, season and daytime independent. Sensor systems have to consider the various landscapes and it has to be possible to detect all kinds of threats. Thus different sensor types providing different data types have to be deployed.

A mix of sensors has to be implemented. To survey an area at night time thermal sensors like infrared (passive or active) have to be installed, contrariwise these sensors are not built for hot weather. Metal detectors have to be deployed to register weapons and for gas or explosives olfactory sensors are of interest.

To get an overall picture and assign surveillance products to an area and put them in a chronological context as well as to fuse data, it is important to provide metadata with the product. The requirements at that point depend on the overall architecture and the display, exploitation and fusion capabilities. Information on chronological and areal allocation of the product, the source and the coverage of the sensor as well as the product type and size should be mandatory. If the data is confidential congruent metadata has to be defined.

Standardization agreements exist in the military domain as well as in the commercial world. NATO standards are of interest for ISR because of the dual use within military and civil ISR. Nations with heterogeneous information and sensor technology need to combine their efforts and achieve a common situational awareness. Most of the information is time sensitive and in both domains there are areas of graded interest.

In the commercial world there are a number of standardization agreements but most of them are less binding and the commitment to those standards is dependent on the application domain.

- ***Military standards***

For the storage and dissemination of digital data STANAG (Standardization Agreement) 4559 NSILI (NATO Standard ISR (Intelligence, Surveillance, Reconnaissance) Library Interface) is the standard interface for querying and accessing heterogeneous product libraries maintained by various nations. “The interface provides electronic search and retrieval capabilities for distributed users to find products from distributed libraries in support of, but not limited to, rapid mission planning and operation, strategic analysis, and intelligent battlefield preparation. The overall goal is for the users, who may be intelligence analysts, imagery analysts, cartographers, mission planners, simulations and operational users from NATO countries, to have timely access to distributed ISR information...” [5].

For data types like image, video, radar or tracks standardized formats exist and are in use (e.g., STANAG 4545 [6], STANAG 4609 [7] and STANAG 5516 [8]). For secondary information like the textual analysis of surveillance products report standards are defined.

- ***Commercial Standards***

The OpenGIS® Catalogue Service [9] defined by the OGC (Open GIS Consortium) is a standard for data dissemination that concentrates on geospatial data, related services and resources. It was not designed for the surveillance area, but could be adapted. The functionalities are similar to the ones defined in the STANAG 4559 [5].

For digital image conservation or video compression there are many standards available. The usage depends on user and domain needs. For video the codices and container formats defined by the MPEG (Moving Pictures Experts Group) consortium are among the most popular ones. Here standards for video compression (MPEG-2/ MPEG-4) and the management of corresponding audio and collateral data (MPEG-4) as well as the handling of metadata (MPEG-7) are defined. For tracks the ASTERIX (**A**ll Purpose **ST**ructured **E**urocontrol Su**R**veillance **I**nformation E**x**change) Standard that is defined by Eurocontrol [11] is of interest. In the maritime sector the NMEA (National Marine Electronics Association) defined a standard that handles navigation data [10].

It is necessary to consider using “the best of both worlds”.

3.2 Data transmission

The choice which medium (wired ethernet/ wireless) to use depends on: the number of connected sensors/sensor ground stations, their location in space, their mobility and the kind of transmitted data. Generally, if the required overall data rate is very high, wireless communications might have problems before a landline’s limit would be reached. Also, security aspects deserve closer attention when using a wireless connection, since it is easier for the man in the middle to listen in, provide false information or disrupt communication. The usage of cables which can be shielded much easier than the complete area of radar coverage bears a smaller risk in this regard.

For the implementation an architecture was chosen with wired connections wherever possible, only a few sensors are connected via a radio link due to the vast distances covered and the prohibition to lay cables everywhere.

4. An architecture for ISR

As ISR systems need to consist of many components it is necessary to build a flexible and adaptive architecture. Once the system components (sensors, exploitation and information systems) are decided upon, it is necessary to establish a way of connecting them. This chapter will detail the “how” of data transfer, what network layout to use. It will explain what happens to the data, once they have left the originating sensor, and in what way they arrive at the desired destination (i.e. the analyst).

For the integrated sensors and information sources converters have to be developed that translate the incoming data into a common data format. As can be seen in Figure 2 different proprietary data formats are converted into a standardized one, so that inquirers do not have to know the type of source (e.g., sonar, radar or infrared) the information is coming from, but can focus on the information itself. The architecture described here is based on these formats (resp. STANAG 4559) [5].

The standardized data is then transferred over the network to a local data server. Connected to the same network are exploitation systems taking in a filtered set of the provided information (depending on the tasking). By fusion and analysis they generate new additional information (e.g. reports) that is also stored in the data server(s). Situational awareness systems are able to display selected intelligence and can ask for additional information from sensors, exploitation or information systems to support decision makers. A detailed description of data dissemination with the shared database is explained within 4.1.

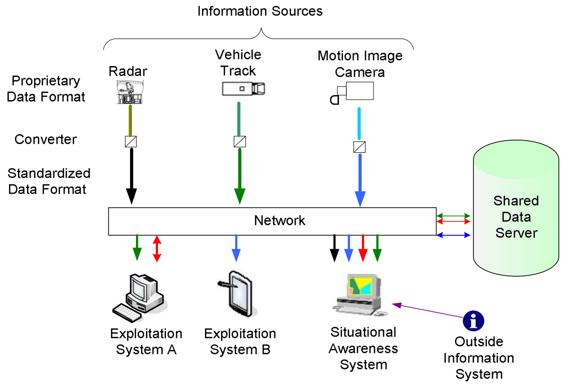


Figure 2: Information sharing within a local architecture

4.1 The Shared Database

The middleware CORBA for the client-server communication was chosen due to its providing transparency in regards to object location, language, and implementation and its established and widespread proven stability and performance. The clients access the server using interfaces for the ingestion or retrieval of data (see Figure 3). If a product is stored in the database, the client hands over the metadata and the product itself to the database via the CreationMgr. To retrieve a product the client asks for the metadata via the CatalogMgr and is able to order the product if it is of interest for the user. Within the metadata all (for the domain) relevant aspects of the product are defined and queryable. Those parameters could be for example: location, time, speed, size, friend/foe, weather condition, certainty of the info, product type.

There are two different ways of retrieving data:

- With the interactive query the user searches the database on aspects of the product (e.g. time, space, product type). A summary of the results is presented in list form on the screen and the relevant data sets can be marked for download in the next step.
- With the subscription method the data query is transferred to the server once and continues running for a set time interval whereby the client is automatically notified whenever new incoming data sets fit the query parameters.

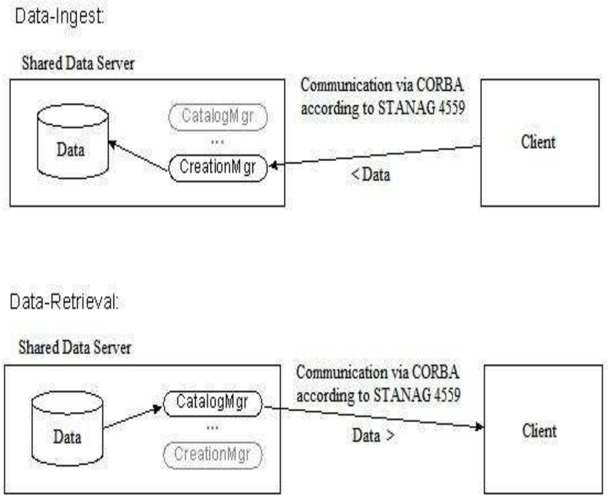


Figure 3: Data ingest and retrieval with the client

Synchronization of information is necessary to be able to share information along different levels and locations. To distribute information from the local network to all interconnected nodes, thereby reducing bandwidth usage and increasing stability, a decentralized setup is used. Here (see Figure 4) the regional servers/local network hubs (in white) collect the data from their directly connected information sources (in black), respectively the converters.

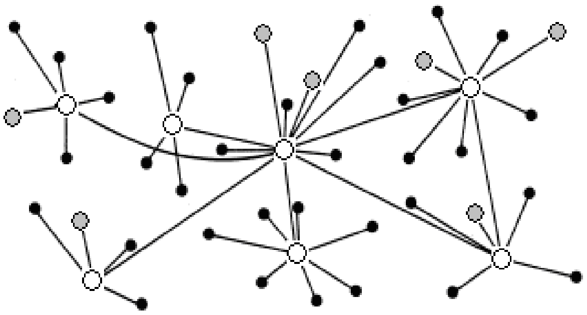


Figure 4: Decentralized network

Note, that sometimes (preferably always) several possible connections between regional servers exist, so that connection loss on one link does not mean the cut off of data sharing. This is especially important, since the connections are of varying quality and reliability (ranging from fluctuating small band wireless to high-speed cable).

The regional servers synchronize their respective metadata content among each other in times of otherwise low network usage but in a set minimum interval using the mentioned subscription method whereby server A plays the role of client when receiving data from server B. The original product data (e.g., images, movie clips of possibly high data volume) are kept only on the originating server. When a situation awareness system (in grey) wants to collect data from different sources scattered throughout the whole network, it uses its client to send a query to its local server which delivers the results by looking only at its own data base. Only when one or more of the results are specifically requested, a connection to the originating server is initiated where the potentially large files (e.g. videos) are transported across the network.

This setup has the advantage of reducing the amount of data traffic on the overall network, because bigger data packets are only transferred on demand, not every time they are created. Also, the local servers serve as a backup in case of network or server failures.. That “complete reliance upon a single point is not always required” [12] was the idea that started the ARPANET (Advanced Research Projects Agency Network) in the 1960s, which later evolved into the Internet.

This system is fully scalable, i.e. the size of each ‘subnet’ including its server can be understood to be either belonging to a certain sensor type in a small region or at the other extreme they could also consist of all sensors of a nation participating in a European network. Also, several layers of subnets are possible.

In addition to accumulation, storage and distribution of data, other tasks of the servers can be the processing and generation of new data out of the received information. Examples are:

- Data fusion (combine the information obtained from a daylight and infrared camera pointing at the same location resulting in new knowledge unobtainable from each single sensor individually)
- Data clarification/extraction (motion imagery processing like image stabilization, mosaicing, object tracking, noise reduction)
- Object recognition (e.g., 3-D object recognition using several 2-D images [13])

To allow hierarchical information access according to different levels of authority further data fields (metadata) are necessary for each product.

On the server – querying client connection side user roles have to be supported, i.e. clients need an account specifying their access rights and statuses. For restricted information, a login and user password has to be supplied before each connection to guarantee the authenticity and the right of access. Additionally, the usage of certificates raises the security of the system to a satisfactory level.

To circumvent unauthorized access to sensitive data from a third party and ensure a secure transmission the data transfer should be encrypted (for example by using a private key infrastructure (PKI)). Using secure lines alone only saves one from intruders outside the network of permitted nodes, but not everyone within the network should be able to read all data from everyone and everywhere. A further complicating situation arises when “friends become foes” within coalitions: a partner who you share data with in one area might be a competitor/adversary in another area or at another time. A more detailed analysis of this “Dynamic Coalition Problem (DCP)” with an emphasis on the military domain can be found in [14].

4.2 Image and Video Processing

In surveillance systems, image and video processing modules are incorporated for three basic reasons: The first one concerns the human machine interface, more precisely the enhancement of video stream visualization. The second one concerns automatic or semi-automatic scene understanding to disburden the operator or analyst. The third one concerns aspects of efficient data storage, e.g. for recording the happenings for offline reconnaissance.

Visualization of sensor data streams may be improved by diverse methods of automatic image processing. The purpose thereby is, to manipulate the video streams such that the video material is optimized for the perception of a human observer, e.g., realized by automatically optimizing contrast or color values of the images.

Another way of improving visualization requires some basic methods of automatic scene understanding: Highlighting moving or otherwise relevant objects or persons in the video streams help the human observer to focus his mind on relevant situations. The same procedure can be implemented, when aspects of data storage are addressed: Instead of recording all sensor data all the time, the recording functionality can be switched off, whenever the scene does not show any (relevant) alteration.

More complex forms of automatic scene understanding can be established in order to generate alarms that call the observers attention to an abnormal situation. Since this kind of high-level situation awareness highly depends on the specification of the specific abnormality and – in addition – presumes more scene information, these methods are less portable to scenarios other than specified.

A system was built that includes a module for (semi-) automatic video processing and sensor data fusion. The module is able to process heterogeneous video sensors (e. g. EO, IR) from different video providers and redistribute extracted information to all participants using the shared database capabilities. Additionally, the image processing module uses sensor information provided by non-video sensors (e. g. sonar, radar, etc.) for automatic control of movable video sensors.

For interaction with the integrated algorithms, a user interface for image processing and analysis was created. The so-called Image Processing Unit has access to all available video streams in the sensor network, is able to receive alarms from all (video and non-video) sensor systems over the shared database and is additionally able to request stored imagery data for offline processing of historic video data from all video providers.

The Image Processing Unit acts as a fusion module between different video sensors and non-video data sources providing three basic processing modes: single-sensor video processing, multi-sensor video fusion and video-sensor remote control (by non-video sensors). In “single-sensor video processing”-mode the Image Processing Module provides generic sensor independent algorithms, applicable to all video sources available in the network. Using the GUI of the Image Processing Module the user is able to select an arbitrary video sensor provided by any participating subsystem and process the video data by miscellaneous algorithms, like “automatic motion detection”, “video-based moving target indicator”, “multi-object-tracking” or “real-time video mosaicing”. The processed video streams are sent back to the sensor network (to provide results to all participants) and important information (detections / alarms etc.) is additionally archived in the database. Of course, if an automatic image processing algorithm detects suspicious behavior it generates alarms like any other subsystem, and broadcasts the information using the common track data and the shared database.

4.3 The Fusion Perspectives

As motivated in 2.3, automatic image processing by means of analyzing distributed and multimodal sensor-data, requires the implementation of sensor fusion techniques, which can be solved by various types of fusion techniques.

In the context of (semantic) sensor data analysis, information fusion can be characterized under several aspects. One of these aspects focuses on the level of information aggregation, where the fusion takes place: “early fusion” indicates that information is fused before semantic descriptions are derived. This might be realized by extracting (abstract) features from each of the sensors, which then are concatenated to a common description, i.e. in terms of a high dimensional vector. The analyzing step then is solved on the common description and results in a multimodal description of events, situations etc.

“Late fusion” on the contrary indicates, that semantic concepts are derived from unimodal sensor data and the fusion step (that often coincides with a kind of decision or detection) is implemented on these semantic descriptions.

Both techniques are applicable in surveillance systems. While early fusion is predestined to be applied on sensor data, that is acquired within a locally limited area where redundant information has to be expected, late fusion techniques should be preferred, when events have to be evaluated in a more global manner.

Another method of sensor fusion, that is relevant for surveillance tasks, is to assign each type of sensor a sensor specific task, e.g., using PIR motion detection sensors as trigger for detailed automatic or semiautomatic analysis in EO- or IR-cameras. This approach is even more efficient, when the cameras are mounted on pan/tilt units: Triggered by non-video sensors cameras can be sighted towards the detected event. This kind of active vision not only results in reduced computational effort, but also allows the reduction of the amount of sensors, that have to be installed. Moreover, this method provides an interaction-free view on the point of interest in a high image resolution.

Finally, the direct fusion of video streams for visualization is a worthwhile technique in the context of surveillance systems: When the number of video streams that have to be displayed in order to monitor a scene is high, multiple sensor streams can be fused into one common video stream. This is even more profitable, when cameras of different modalities are directed onto the same location of the scene, since there exist image fusion techniques that accentuate the sensor specific information. E.g., for two cameras, one of them in the visual and the other one in the thermal spectrum, these approaches deliver images, that show visible structures and thermal structures as overlays.

5. Deployment of the Architecture

The previously described architecture was implemented in several trials and exercises. The core component for data dissemination always was a shared database. Data fusion was implemented on differentiating levels within the different projects.

5.1 MAJIIC

The primary driver for the Coalition Shared Data (CSD) Server was the project of MAJIIC (Multi-Sensor Aero-Ground Joint ISR Interoperability Coalition) [15]. During the last conflicts that German Bundeswehr and allied forces were involved in ISTAR Data (Intelligence, Surveillance, Target Acquisition, and Reconnaissance) could not be shared and exploited among the involved forces because of technical and operational problems which resulted in a loss of human lives and material.

To avoid this in the future the multinational project was introduced. The aim was to strengthen and prove the processes, methods and applications that support interoperability. Interoperability is enforced by standardized data dissemination.

The sensor data that is processed within the sensor workstations in proprietary formats is transformed into standardized formats and shared over standardized interfaces.

The concept passed its first full-blown test during a major NATO exercise in Norway, Bold Avenger/Trial Quest 2007 [16], which included real-time maneuvers by several thousand air and ground forces. During the exercise joint ISR interoperability was demonstrated in a „Live environment with a multi-sensor, multi-service geographically dispersed set-up".

5.2 Common Shield

In 2008 the concept was successfully tested during the common Bundeswehr experiment Common Shield and NATO DAT (Defence against terrorism) experiments Technology of ISTAR against Terrorism, Critical Infrastructure Protection, and Harbour Protection Trial [17]. The aim of the Common Shield exercise was to test C2 (Command and Control) processes in a NEC (network enabled capability) environment with integrated ISR and C2 systems. The Common Shield architecture integrated sensor systems, exploitation capabilities, situation awareness tools, common operational picture displays, and C2 systems provided by 27 different producers. Amongst the collection assets there were airborne imaging sensors and ground based imaging sensors; but also ground based radar systems providing MTI and chemical sensors capable to detect explosives as well as sea-based surface and sub-surface sensors. Exploitation systems provided capabilities to exploit still and motion imagery, MTI data, to fuse alarms generated by the chemical explosive detection sensors with imagery, and to fuse alarms and tracks generated by the sea-borne sensors with imagery. The seamless data and information exchange of all the sensor data and exploitation products with a real-time update of the common operational picture was enabled by the employment of a series of CSDs with the capability of storage, query, subscribe and retrieve, and automatic real-time synchronization of metadata.

5.3 SOBCAH

An exemplary implementation of this architecture was realized and successfully demonstrated in the European Project SOBCAH [18]. Within a demo at the harbor of Genoa different threat scenarios were exercised. The joint observation of land- and sea-borders with a variety of sensors, among them sonar, radar, video (IR and EO), container and car tracking systems and motion detectors, was tested. The information retrieved from all these sensors was stored in the SOBCAH Shared Database (SSD) that was designed upon the described architecture principles. A situation awareness system subscribed (via a client) to data stored in the SSD and was able to display all relevant information for local decision bodies. As it was possible to store relevant data forensic analysis at a later point in time was possible as well.

The demo showed successfully that data from all kinds of different sensors can be integrated into one system where in a timely manner, i.e. without long delays in time due to the large amount of data a common ground picture of a situation can be extracted.

6. Conclusion and Future Work

Within MAJIIC detailed processes to share information within the ISR domain have been developed, implemented and tested. The focus of this work was on multinational information sharing within a coalition with (mainly) satellite and airborne IMINT sources. The usage of an architecture as previously defined proved to be applicable for the information exchange between different nations.

Within the Bundeswehr experiment Common Shield adaptability of that architecture to different sensor types and surveillance needs was successfully tested. Although the adaptation of some of the STANAGS was necessary it was possible to integrate new systems in this type of architecture relatively easily.

Within SOBCAH the usage of such an architecture in a civil environment was demonstrated. Data fusion techniques as described above were of great use to help the operator focus on relevant events.

For future work within CIMIC the seamless integration such an architecture with the means of converters and services should be further developed.

Standardized mechanisms of data dissemination in civil and military security applications should be enforced as this enables an agile plug and protect system. Sensor and information systems can be integrated and thus different sources of information can be fused if necessary.

The integration of other data/information sources (e.g. human intelligence, electronic intelligence) in such an architecture has to be evaluated and planned. To be able to cooperate on a semantic level the mapping of data models and ontologies from the different domains (civil and military, ISR and C2 etc.) has to be enforced and integrated in an intelligent situation awareness system.

7. References

[1] Essendorfer, B., Monari, E., Wanning, H.(2009). An Integrated System for Border Surveillance. GlobeNet ICN 2009, International Conference on Networks. 28.02.-05.03.2009. Cancun, Mexico.

[2] Jaeger, T., Hoese, A., Oppermann, K. Transatlantische Beziehungen. Sicherheit- Wirtschaft- Oeffentlichkeit. Verlag für Sozialwissenschaften. 2005

[3] Endsley, M. R. Situation awareness global assessment technique (SAGAT). Proceedings of the National Aerospace and Electronics Conference (NAECON). (New York: IEEE), 789-795. 1998

[4] Endsley, M.R., Garland, D.J. Situation Awareness Analysis and Management. Lawrence Erlbaum Associates. 2000

[5] STANAG 4559 NATO Standard ISR Library Interface. Edition 2. http://www.nato.int/structur/AC/224/standard/4575/ag4_4575_E_ed2_nu.pdf. 27.07.2009

[6] STANAG 4545 NATO Secondary Imagery Format (NSIF). http://www.nato.int/structur/AC/224/standard/4545/4545_documents/4545_ed1_amd1.pdf. 27.07.2009

[7] STANAG 4609 NATO Digital Motion Imagery Format. http://www.nato.int/structur/AC/224/standard/4609/4609_documents/4609Eed01.pdf. 27.07.2009

[8] STANAG 5516 Tactical Data Exchange- Link 16.

[9] OGC (2005). Open GIS Consortium. OGC catalogue service specification. http://portal.opengeospatial.org/files/?artifact_id=5929&version=2. 27.07.2009

[10] NMEA 0183 Interface Standard. NMEA 0183 Interface Standard

[11] Eurocontrol standard document for surveillance data exchange. Part 1. All Purpose Structured Eurocontrol Surveillance Information Exchange (ASTERIX). http://www.eurocontrol.int/asterix/gallery/content/public/documents/pt1ed129.pdf. 27.07.2009

[12] Baran, P. (1964) On Distributed Communications Network. IEEE Transactions on Communications [legacy, pre - 1988], 12, 1

[13] Dutta Roy, S., Chaudhury, S. and Banerjee, S. (2004). Active Recognition through Next View Planning: A Survey. Pattern Recognition, 37, 3, pp. 429 – 446

[14] Phillips, C. E., Ting, T.C., Demurjian, S. A. (2002). Information sharing and security in dynamic coalitions, Proceedings of the seventh ACM symposium on Access control models and technologies, June 03-04, 2002, Monterey, California, USA

[15] MAJIIC (2007). http://www.nato-otan.org/docu/update/2007/pdf/majic.pdf. 27.07.2009

[16] Trial Quest (2007). Key NATO reconnaissance technology passes major test. www.nato.int/docu/update/2007/12-december/e1210d.html.27.07.2009

[17] Stockfisch, D. (2008): Common Shield 08. TechDemo 08/Harbor Protection Trials. Strategie & Technik, October 2008

[18] SOBCAH. Surveillance of Borders, Coastlines and Harbors, http://ec.europa.eu/enterprise/security/doc/project_flyers_2006/766-06_sobcah.pdf.27.07.2009

Design Patterns for a Systemic Privacy Protection

Kajetan Dolinar, Jan Porekar, Aleksej Jerman Blažič
Security Technology Competence Centre (SETCCE)
Tehnološki park 21, SI-1000 Ljubljana, Europe
kajetan.dolinar@setcce.si, jan.porekar@setcce.si, aljosa.jerman-blazic@setcce.si

Abstract

This paper shows that existing privacy enhancing technologies and the state-of-the-art in research on the field of privacy protection has grew to a considerable maturity up to date, yet privacy protection regulation disregards these advancements and remains in vague terms. Contemporary social situation with regards to privacy protection entails serious arguments why this disparity should rather soon be overcome. It is further shown how this disparity could be overcome by a collection of privacy protection patterns which include technical solutions as well as social models and can be combined into a systemic privacy protection framework that could be declared on the level of regulation itself in much more detailed and concrete terms than today.

1. Introduction

There have been many advancements in privacy enhancing technologies up to date. We are witnessing innovations in area of identity management solutions, trust management, privacy policy negotiation and trust negotiation, access control and many more. We know the legal and social context for privacy protection: there are known court cases and public affairs. Yet there is still a wide gap between the technologies on one edge of the gap and having them properly integrated into the legal and social paradigm of privacy protection on the other edge.

There is a lot more to privacy protection than only technologies. Technologies themselves are inefficient without a general consensus on how they should be used in a proper way. Privacy cannot be protected without a complete social support in terms of regulation, successful prosecution and business interest as well as public awareness, social studies and public education. Technology should be complemented with these expert areas to provide a systemic framework for privacy protection in society as a whole.

One of the most urgent problems is that business does not have enough incentives to invest in privacy enhancing technologies [1]. A possible reason for this may be that a formal institution of technical patterns and social models is missing where the technologies and the social structures would be combined in a congruent system with at least theoretic proof of working. Having such an institute would allow legislation on a much more concrete basis than today; it would make possible to define exact procedures for privacy protection in every data collection or processing. This in turn would force data controllers and data processors to invest into privacy enhancing technologies and thus give privacy enhancing technologies market value.[1]

This paper presents one of the possible ways how a formal institution of technical patterns and social models can be established. The approach presented here describes each of the patterns and models formally in terms of privacy protection patterns. The whole idea is referred to as *privacy protection cycle*. Section 2 elucidates the situation on field of privacy protection from the point of view of current European legislation and public privacy protection issues as witnessed up to date. In the Section 3 the current state-of-the-art in privacy enhancing technologies is presented. Section 4 gives an overview of the privacy protection patterns and Section 5 shows how they work inside the privacy protection cycle. Paper ends with conclusions in Section 6. Additional support for arguments in the following sections can be found in appendix A.

2. Legal and Social Context for Privacy Protection

This section uncovers the contemporary state in the advancements of legal frameworks and contemplates about the evolving situation in public affairs regarding privacy protection; different types of problematic situations are exposed in order to corroborate or defy the efficiency of regulation or

[1] Data controller is the party which determines the purposes and means of the processing of personal data while data processor means the party which actually processes the data on behalf of the data controller.

to provide requirements for various privacy protection patterns.

2.1. Legal Synopsis

Most of the countries in the world have their own privacy protection legal acts. This article will focus mainly on European regulation on data protection defined by European Directive 95/46/EC [2] as it summarizes all important data protection principles:

- Principle of fair and lawful processing (Article 6(1), letter a): *"Any processing of personal data should be carried out in a fair and lawful way with respect to the data subjects.[2]"*

- Finality principle or Limitation principle (Article 6(1), letter b): *"Personal data must be collected for specified, explicit and legitimate purposes and may not be further processed in a way incompatible with those purposes."*

- Data minimisation principle or Proportionality principle (Article 6(1), letter c): *"Processing of personal data should be limited to data that are adequate, relevant, and not excessive."*

- Time minimization principle (Article 6(1), letter e): *"Data should be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed."*

- Notification principle (Articles 10, 11): *"Data controller or his representative has to identify himself to the data subject and notify data subject about the personal data being processed, stored or further disclosed to any third party."*

- Principle of data subject consent (Article 2, letters a, h; Article 7, letter a; Article 8): *"User consent is required for a legitimate data processing by any data controller. The user consent is defined as 'any freely given specific and informed indication by which the data subject signifies his agreement to personal data relating to him being processed.'"*

- Principle of right to access personal data (Article 12): *"Data subject has right to access (and rectify) the data collected about him and to be informed about the intended processing and the logic behind the intended processing of the data."*

- Principle of right to object processing of personal data (Article 14): *"Data subject has right to object processing of personal data (subject to certain constraints: compare articles 7, letters e and f)."*

Additional to those principles many other legal principles indirectly related to data protection can be found in other legal acts. This paper will not include an overview of those legal instruments. For the purpose of this paper the above reduced list is enough to give the exposure of the level of detail provided by the regulation: generally the regulation does not prescribe methods or technologies for data protection, it merely provides the general principles.

2.2. Problematic Social Situations

A lot of situations and affairs are known to have happened up to date with a considerable importance for privacy protection. In the following we provide a succinct[3] overview of the global state of affairs and the most critical issues with respect to privacy protection. This summary has been produced from a notably more exhaustive source provided by European FP6 SWAMI project [3]; all the footnotes in this section are reproduced from this source.

Problem Situation 1 – Working from home, monitoring of employees. What is the dividing line between working environment and private / home environment? All the privacy protection principles are related to this problem; moreover, article 8 of European Convention of Human Rights [4] protects the private home. The problem is that a workplace might be situated in a private home and that also a typical workplace is used, to a lesser extent, for private purposes.[4]

Problem Situation 2 – Digital rights management. How much it interferes with privacy of an individual? Proportionality principle obliges policy-makers to consider alternative, less infringing ways of protecting intellectual property rights, reconciling them with privacy rights.

Problem Situation 3 – ID theft. Identity thefts are reality![5] This is against Principle of fair and lawful processing, Notification principle, Principle of data subject consent, and Principle of right to object processing of personal data. Automated payments make it easy to spend money quickly. Distance contracts do not offer the same guarantees, trust and confidence as in physical commerce. It is known that

[2] Data subject is a legal or natural person to which the data refer.

[3] Succinct here means that there are nine situations complex enough to elaborate on several tenths of pages but we have to compress them to fit a few; moreover, it is in purpose of clarity that every text that would be superfluous to the exact definition of the situation, yet important to support our arguments, is placed in a footnote. Thus, the reader is asked to forgive the abundance of footnotes on this and the following pages; a reader may skip them without much harm or read them in appendix A.

victims of identity theft have great difficulties recovering from the consequences.[6]

Problem Situation 4 – Data laundering. Companies are paying a lot of money for personal and group profiles and there are market actors in position to sell them.[7] This is clearly against data protection principles. This phenomenon is known as 'data laundering'. Similar to money laundering, data laundering aims to make illegally obtained personal data look as if they were obtained legally, so that they can be used to target customers.[8]

Problem Situation 5 – Personal profiling. Personal profiling is reality. A lot of people do not realise how much personal information they are constantly giving out.[9] Also this activity is illegal with respect to data protection principles.

Problem Situation 6 – Inadequate profiling. People are victims of an inadequate profiling based on false data or processing.[10] This violates the Principle of fair and lawful processing and is a great motivator for Notification principle, Principle of data subject consent, Principle of right to access personal data, and Principle of right to object processing of personal data. However, it is far from truth that people would always be given those rights.[11]

Problem Situation 7 – Disproportional request for personal information. It is not a rare case that a data controller requires information in extent disproportional to the purposes of business. There is not much case law in which one can find out what "proportional" data processing is and what is not. However, disproportionate data collection happens and is prohibited by Principle of fair and lawful processing, Finality principle, and Proportionality principle.

Problem Situation 8 – Spyware and personal preferences. Spyware is a frequent way to steal data and intrude privacy.[12] The use of spyware programs (installing and spying) constitutes a number of criminal offences according to the Cybercrime Convention [5]: illegal access (Article 2) and illegal interception (Article 3) when there is, in the latter case, an interception, without right, made by technical means, of non-public transmissions of computer data to, from or within a computer system.

Problem Situation 9 – Advertising and spam. Spam not only takes time and provokes irritation, but also can influence and infringe someones private world. European Directive 2000/31/EC [6] and European Directive 2002/58/EC [7] both contain provisions on (unsolicited) commercial communication, but do not really seem to have the desired effect. This situation also defies Principle of data subject consent and Principle of right to object processing of personal data.

2.3. Final Notices

Not much can be done to prevent authorities or business to know data of people. But it is relatively simple to provide an individual with a set of (potentially false) identities which can vouch for important properties of that individual such as the individual is employed, has a regular health insurance, is of appropriate age, receives such and such income, etc. The identities should provide the data along with certificate material proving the data are accurate, however the identities themselves could be completely pseudonymous. The feasibility of such a technology is not a question (cf. Section 4, Virtual Identity pattern); what is missing is that the legislation in the first place should legalize such identities and clearly define transactions where they are allowed.

Companies that process personal data acquired from third parties are bound by the rules of data protection. It might be a good safeguard to oblige those companies to check where the information they buy comes from and if it has been lawfully acquired. Similar obligations could be imposed like those on banks to control money laundering. This is important for Sticky Policies, Privacy Audit Trail, and Access Control privacy protection patterns.

Speaking of the problem of disproportionate data collection it is very difficult to define what is "proportional". On the one hand, there are too many diverse situations in which processing takes place, so that one particular situation might require more data processing for one reason or another. This motivates the idea of Privacy Policy Negotiation pattern.

3. Privacy Enhancing Technologies up to Date

Privacy enhancing technologies come in a variety of different kinds of solutions. This section will provide a quick overview through existing state-of-the-art and will try hard to be as broad as possible. This is important as there will be a need for references to the real solutions when defining patterns in Section 4. Privacy enhancing technologies can be divided into six categories:

- privacy preferences and policy languages,
- trust & reputation systems,
- trust & privacy policy negotiation,
- identity management,
- data conservation,

- control of processing,

We will briefly touch every category in the following paragraphs.

Privacy preferences and policy languages. There are systems which enable users to check how their privacy preferences relate to privacy policies of a data controller. An example is Privacy Bird [8] developed at the AT&T, a P3P user agent as a browser helper object for the Internet Explorer 5.01, 5.5, and 6.0 web browsers on Microsoft Windows 98/2000/ME/NT/XP operating systems. The Platform for Privacy Preferences (P3P) [9] enables organizations to express their privacy practices in a standard XML format that can be retrieved and interpreted automatically by user agents. P3P policies can encode contact information for the legal entity making the representation of privacy practices in a policy, enumerate the types of data or data elements collected, and explain how the data will be used. P3P Preference Exchange Language (APPEL) [10] complements P3P Specification by providing a language for specifying users preferences regarding P3P policies.

Several other policy languages have been devised for expressing privacy preferences: The Enterprise Privacy Authorization Language (EPAL) [11] is an interoperability language for exchanging privacy policy in a structured format between applications and / or enterprises, structured in XML. Another XML based language for expressing access control policies is XACML [12], which stands for eXtensible Access Control Markup Language. It complements SAML [13] which is in purpose of conveying information on authorization, authentication, and related attributes in an XML formatted assertions.

Besides encoding of simple policy rules there has been a lot of research made in knowledge representation systems enabling support for enriched semantic processing such as Description Logic [14], a family of languages on the level of the first order predicate logic with extensions for knowledge representation, some of them are well known ontology systems such as OWL - DL [15], KAON [16] and KAON2 [17]. The use of ontologies for representing important notions of privacy protection has also been researched [18][19].

The potential for machine reasoning and semantic processing over policies has been researched by PRIAM project [20][21]. The authors contemplate on theorem provers, systems that parse expressions of a (first order) logic language and process them in order to derive formulas of logical truths. There are many ways how this can be done such as method of analytic tableaux [22] or resolution with unification [23]. There are many known theorem provers such as Otter [24] or Coq [25]. One of the methods modern machine reasoners often use is superposition and term rewriting which takes a hypothesis formula, parses it and replaces subformulas by the rules of inference – either the classical formal logic rules or the rules representing facts about the actual domain of discourse referred to as *background knowledge* – as long as there is some rule possible to apply.[26][27] This method is especially suited to machine reasoning systems for evaluating ontologies such as already mentioned KAON and KAON2 systems, FaCT++ [28], Pellet [29], DIG [30] or JENA [31].

Trust & reputation systems. There are various technologies enabling trust in some way. Technologies based on PKI and cryptography have already been known for a long time; they are used to certify authorizations, identity, or other traits by means of cryptography and systems of global trust. The reference to those technologies will tacitly be assumed throughout the paper where questions about integrity or confidentiality will rise.

However, there is another branch of research and technologies which treats trust in a substantial way, tapping into the very social phenomenon of trust and reputation. There have been many attempts to formalize this important aspect of social interaction [32][33][34] and many a model studied of how trust and reputation are induced in people and society [34][35][36][37]. Some of those models are successfully used by today's leading electronic market actors such as Amazon.com or eBay for evaluating how good or bad individuals perform in transactions.

The notion of trust is most often modelled as a real number on interval from 0 to 1 inclusively, 0 meaning no trust and 1 meaning full trust. It is disputable, though, how relevant such a trust assessment is in a given situation with respect to what is to be trusted: trust should be assessed for every different kind of relation in separate. A separate number should be used for assessing how trustworthy a person is as a business partner, and a separate for how trustworthy that person is as an expert in his or her field; these two relations can each imply quite a different trust situation. One number cannot capture such a multidimensional problem. Even harder is to make a model for evaluation of trust based on individual assessments as there can always be individuals that will introduce false opinions in the system; however, recent investigations show that trust calculation models can be built which preserve high level of fidelity in communities where as much as 80% of individuals give false trust assessments [37]. Despite the weaknesses of trust management systems there are many more good reasons why such systems should be used (cf. Privacy Policy Negotiation and Trust & Reputation Evaluation System privacy protection patterns in Section 4).

Trust and privacy policy negotiation. Another way of how trust can be established between anonymous actors is by the so called *trust negotiation* whereby two negotiators exchange identifying or certifying information in order to acquire the sufficient amount of proof about the opponents trustworthiness.[38][39] Moreover, negotiation has also been investigated in scenario of a straightforward bargaining for resources and privacy protection practises, supported by rich semantics.[40][18][19][41] This type of negotiation is referred to as the *privacy policy negotiation*. Privacy policy negotiation as well as trust negotiation can be valuable tools for a fine-grained restitution of data to be disclosed and privacy protection rules before the actual data are released, thus implementing in reality Data minimisation principle and Finality principle. For a successful negotiation, taking it abstractly, there should always be a formal result in sense of an agreement, where the statements both sides agreed are evident.

Identity management. Numerous initiatives and technological standards have been created up to date for different kinds of frameworks for introducing identities into electronic transactions. YADIS [42] is an open initiative to build an interoperable lightweight discovery protocol for decentralized, user-centric digital identity and related purposes. With YADIS the capabilities of identities can be composed from an open-ended set of services, defined and/or implemented by many different parties. OpenID [43] is a distributed, decentralized network where identity is represented as a URL and can be verified by any server running the protocol. It is a part of the YADIS family of protocols. Light-Weight Identity (LID) [44] is a set of protocols and software implementations created by NetMesh Inc. for representing and using digital identities on the Internet without relying on any central authority. LID supports digital identities for humans, human organizations and non-humans (e.g. software agents, things, websites, etc.). XRI/XDI [45][46] is an international non-profit organization governing public services based on the XRI abstract identifier and XDI data interchange protocols. This new layer of infrastructure enables individuals and organizations to establish persistent Internet identities and form long-term, trusted peer-to-peer data sharing relationships. iNames are one form of an XRI, an OASIS open standard for abstract identifiers designed for sharing resources and data across domains and applications. One problem XRIs are designed to solve is persistent addressing – how to maintain an address that does not need to change no matter how often the contact data for a person or organization changes. XRIs accomplish this by adding a new layer of abstract addressing over the existing IP numbering and DNS naming layers used on the Internet today. Privacy is protected because the identity owner controls this resolution. Simple eXtensible Identity Protocol (SXIP) [47] is a protocol for automating the exchange of identity data on the Internet. It supports Single Sign-On access to different websites. User-Centric Verified Identity allows users to acquire and release verified “assertions” around their identity, enabling them to create richer profiles of their online identity. User Choice supplies added privacy by enabling users to be actively involved in the release of the data they store in their identity profile. Many of the technologies mentioned here are embraced under the Liberty Alliance initiative [48] for open standards, guidelines and best practices for federated identity management with its own identity architecture specification [49]. Another initiative for federated identity is Shibboleth [50] with architecture and open-source implementation for federated identity-based authentication and authorization infrastructure based on SAML.

CardSpace is a software which securely stores digital identities of a person, and provides a unified interface for choosing the identity for a particular transaction, such as logging in to a website. CardSpace [51] is a central part of the Microsoft effort to create an identity meta-system, or a unified, secure and interoperable identity layer for the Internet. The CardSpace software allows the users to create self-signed identities for themselves. CardSpace is built on top of Web Services Protocol Stack. Higgins trust framework [52] is a set of open source protocols and software applications that allow people to store their digital identities on their personal computers and share the stored information with commercial companies and other parties in a controlled fashion. Higgins is sponsored by IBM and Novell and is promoted as an alternative to Microsoft’s CardSpace.

Besides these efforts much research has been carried out for advanced identity management schemes in scope of some European projects. In projects PRIME and DAIDALOS the concept of multiple (virtual) identities is explored, where every person can have several identities, some of them with blurred or obfuscated data, to enhance user’s privacy.[53][41] This is one of the most powerful techniques for enforcing Data minimisation principle. The concept also makes use of pseudonymity and anonymity approaches.[54] The distinction between addressable and non-addressable pseudonyms is used to allow for the integration of identity management into the application logic. For example, non-addressable pseudonyms are used in the task assignment scenario (see [55], Section 4.5) to allow application internal processes to be supported by the identity management. In order to make a pseudonymous identity untraceable on the network level there exist models such as Onion Routing [56] and Crowds [57] which make possible routing, and thus internet communication, in an anonymous way. There are also software projects that enable this kind of protection.[58]

Data conservation. By data conservation different kinds of techniques are meant. First of all, an appropriate type of access control should be in place before access to private information is allowed to an arbitrary agent. The three most commonly refered models are Discretionary Access Control

(DAC), Mandatory Access Control (MAC), and Role Based Access Control (RBAC).[59] In DAC the owner of the resource decides who is allowed the access to the resource and what privileges they have. This may be achieved using access control lists or credentials, i.e. keys admitting access to the resource. In MAC agents have sensitivities assigned specifying their level of trust and resources have sensitivities assigned specifying the level of trust required for access; in order to access a resource, the subject must have a sensitivity equal to or higher than the resource. MAC can be achieved by a rule-based access control defining specific conditions for access to a resource, e.g. simple rules applying matching sensitivity labels of agents and resources; or it can be achieved applying lattice model, a mathematical structure, to infer complex decisions on relations between agents and resources. In RBAC collections of permissions that may include complex operations are the key to access a resource; agents are admitted to a resource if their roles assignments satisfy permissions. Recently, a new type of access control was introduced called Purpose Based Access Control (PBAC) with a scheme where access to a resource is allowed if the agent has a justified purpose for that.[60]

Finally, in order to protect privacy in the most general sense, data should be protected using cryptographic methods when transferring them over communication channels. More advanced protection is achieved by *steganography* [61] or other kinds of data obfuscation [62]. When privacy agreements are in question, in case they are represented in a digital form, their integrity and time of creation should be preserved by long term trusted archives [63]; failing to do so might disable a curtailed person to assert a privacy breach, since what should constitute a privacy breach in such a case is primarily measured by what the person was guaranteed by the opposite side and that should be saved for later reference in an agreement. As certificate material deteriorates over short periods of time (e.g. five years for a PKI certificate validity is already a long period), integrity and time origin of agreements or other digital documents have to be preserved using methods for long term trusted archiving.

Control of processing. When private identifying information is disclosed to data controllers and is under processing by data processors, the techniques mentioned until now can only have limited or no protection power. However, there has been a rich tradition in research of techniques for protecting privacy after disclosure, the so called *a-posteriori* privacy protection, albeit some of them may not be recognized as such in the past. One of the first attempts to protect privacy of data during processing (i.e. after their disclosure) was proposed by Rivest, Adleman, and Dertouzous [64] who introduced the idea of performing simple computations on encrypted data, the technique referred to as *privacy homomorphism*. The idea was studied in context of various cryptosystems and for different problems such as summation, multiplication, derivation, and integral of encrypted polynomials or union and intersection of encrypted sets [65]. The approach generally allows for the joint computation of a wide variety of functions, however its uses are limited by severe message expansion.

Related techniques were devised for the so called *Hippocratic databases* and privacy preserving data mining. The name “Hippocratic databases” was inspired by the concept of Hippocratic oath that has guided the conduct of physicians for centuries. The concept is a research initiative started in 2002 in IBM Almaden Research Center.[66] In the original paper the authors argue that privacy is the right of individuals to determine how and to what extend information about them will be communicated to others and suggest that the database community has opportunity to play central role in the privacy debate by re-architecting the database systems to include responsibility for the privacy of data as the fundamental tenet. The idea inspired the field of research on privacy preserving data mining.[67][68][69]

Further techniques for *a-posteriori* privacy protection were sought in direction of privacy auditing and introduced by European projects such as PRIAM [70]. The idea is that records of all the actions performed on data are stored in a trusted hardware module and authorities can perform auditing of those logs. A necessary prerequisite for this are the so called *sticky policies* [71]: to each unit of data a formal statement is attached where all the constraints about which actions are allowed on the data are defined by the owner of that data and described in a machine readable format. The sticky policy follows the data unit after it has been released to processing; the data along with the sticky policy should be cryptographically signed by the owner so that the data and the sticky policy can be proven to belong together. Data without a sticky policy should be deemed invalid. Every operation data processor commits on data is compared by the trusted hardware module to the sticky policy and if data processor has done operations to data that are not permitted by the sticky policy, then this can be signalled to the supervising agencies. This is referred to as *privacy audit trail*.[20][21]

4. Privacy Protection Patterns

We have seen that a variety of concrete techniques are available for privacy protection, which is in obvious contrast to the level of vagueness of legislation. The social situation calls for concrete solutions, but the gap between the legislation and the technologies keeps data controllers and data processors far away from using the concrete solutions in order to comply with legislation. What is missing is a collection of artifacts, which will be referred to as *privacy protection patterns*, each describing a particular technical solution or social model for privacy protection. Then legislation can be defined in terms of those artifacts.

4.1. The Structure of Patterns

A privacy protection pattern is an abstract scheme of how a particular approach to privacy protection is possible. It should include a unique name, a list of the actors included in the pattern, a list of properties the pattern provides, the context in which the pattern satisfies the properties, a succinct description of the solution and a diagram displaying the exact working of the solution. Additionally to this each pattern should include information on how it can be used in terms of controls which activate the pattern. Patterns will be laid down in the following form:

The form for specifying privacy protection patterns

Actors

A list of legal or natural persons involved in the pattern and their interests.

Properties

A list of abstract properties the pattern exhibits in relation to interests of the actors and in relation to the requirements implied by the context.

Context

A list of resources, instruments, or social and technical constraints and other factors which define the means and the possibilities how actors can achieve their ends.

Description

A succinct description of the way how the pattern operates on the context to moderate the actors' interests and achieve some aspect of privacy protection.

Controls

A list of features of that pattern allowing regulation or maintenance of the operation of the pattern described in terms of actions actors can take upon the context and the constraints for the admission of actors to commit those actions.

Definition

The activity diagram as specified in UML[72] displaying how the pattern operates on the context and how the controls influence this.

4.2. The Privacy Protection Patterns

The first pattern is about policies enabling exact specification of data protection rules and supporting machine reasoning. Principle of data subject consent, Finality principle, and Proportionality principle are impossible to enforce in an exact way without such a pattern; also what is a disproportional request in Problem Situation 7 is easier to resolve. Technical feasibility of this pattern is largely supported by policy specification languages and related research as was presented in Section 3.

Formally Provable Privacy Policies

Actors

Person – a legal or natural person.

Properties

- Deductibility: the structure of the policy supports computer aided inference of logical properties, semantic attributes and other constraints;
- Completeness: the policy is unambiguous;
- Soundness: the policy is not contradictory.

Context

- Privacy Policy: the certifying information and privacy protection rules of Person whose structure captures the first order predicate logic with extensions for knowledge representation;
- Theorem Prover: a system capable of deriving logical truths about Privacy Policy;
- Domain Model: a digital representation of the important notions from privacy protection with relations among them such as taxonomical relationships of inheritance of type or other ontological properties;
- Challenge: a request to disclose a certain personal identifying information of Person formated in the same structure as that of Privacy Policy.

Description

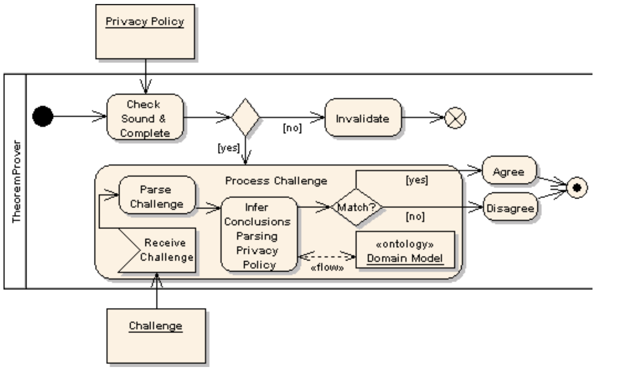
Privacy Policy captures the rules how the Person's data should be protected by data controllers and processors regarding privacy protection. The Domain Model provides the background knowledge for the Theorem Prover. The Theorem Prover makes possible a computer aided inference of facts and conditions regarding Privacy Policy against an arbitrary Challenge. Theorem Prover also makes possible checking completeness and soundness of the Privacy Policy.

Controls

For each data collected Person should be able to set in Privacy Policy at least the following:
- ENTITY identifies the Person;
- DATA describes the data which are subject to the following controls;
- PURPOSE specifies the allowed purposes for the collection or processing of DATA;[13]
- RECIPIENT identifies recipients of DATA;
- RETENTION indicates the kind of retention admissible for DATA;
- OPERATIONS defines the operations allowed on DATA;
- OBLIGATIONS states how data controllers and processors should protect DATA when handling them.

Each Challenge should define at least the following:
- REQUESTER identifies the data controller which requests a particular data;
- DATA describes the data requested;
- PURPOSE describes the purposes for data collection;
- OPERATIONS describe the operations REQUESTER wants to perform on DATA.

Definition



This pattern should be used in conjunction with

Privacy Preferences Helper Tool

Actors

Person – a legal or natural person;
Requester – a legal or natural person requesting a resource or data of Person or otherwise challenging Person's privacy.

Properties

- Maintainability: the Person is able to maintain the own set of privacy preferences;
- Friendliness: the Person is able to do so in a user friendly way.

Context

- Privacy Policy: Person's privacy preferences and other rules in a machine readable form;
- User Interface: a collection of commands and displays enabling insight into and management of Privacy Policy and related privacy preferences;
- Translator: a process which resolves user input from User Interface and relates this to the appropriate sections of Privacy Policy;
- External Application: an application through which Requester is able to challenge Person;
- Challenge: a request for a resource or data of Person or any other kind of challenge for Person regarding privacy protection.

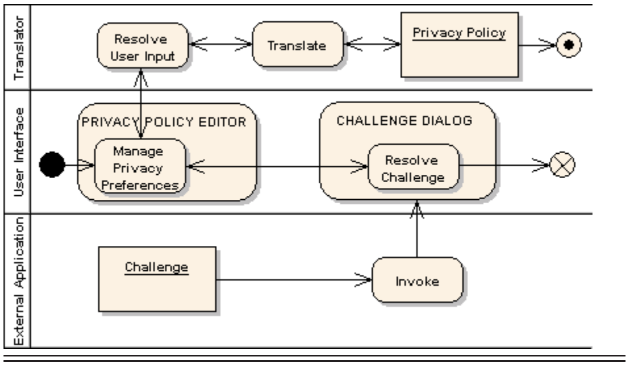
Description

User Interface enables Person to manage privacy preferences from Privacy Policy. Additionally, when Person's personal data are requested or Privacy Policy is challenged in another way by Requester, User Interface automatically informs Person of that challenge and Person is able to interactively resolve this.

Controls

User Interface should have at least two components:
- PRIVACY POLICY EDITOR is a display of Privacy Policy in a human readable way with controls which enable Person to add, update, delete, or otherwise modify privacy preferences and other rules from of Privacy Policy in a user friendly fashion;
- CHALLENGE DIALOG is a message display with appropriate controls and can be invoked by External Application when a Challenge occurs, enabling Person to interactively resolve the Challenge.

Definition



This pattern tells about a tool which makes possible a definition of privacy policy in a human friendly way so that user does not need to code the privacy policy. This is very important as policies generally are of the same complexity as programming languages and most of the people do not know how to do that. This pattern supports Principle of data subject consent and Finality principle, because the data subject can explicitly decide privacy protection preferences, it further supports Proportionality principle because it makes possible specification of which data and for what purposes and when should be disclosed. The pattern is technically possible as shown in Section 3. This pattern should be used in a combination with previous one and in combination with Privacy Policy Negotiation pattern.

The next pattern reflects upon the fact that whenever an individual or a company is about to disclose a particular information to another party it is all about trust and reputation, how much the other party will respect and protect privacy. What *respect* in the later case means can be defined as whether and how much the ways that other party handles the information are in accord with the data subject's privacy policy. Clearly, the case here is of a very specific trust domain, namely trust in how good or bad the information will be handled; accordingly, any trust representation, be it by number or category, will reflect on this. This pattern, especially in a combination with Privacy Policy Negotiation and Identity Management patterns privacy protection patterns, can be an important instrument for enforcing Principle of fair and lawful processing or Finality principle and to ward against Problem Situations 3, 4, 7 or 9; namely, market actors that indulge in such activities will get low reputation which will affect demand and consequently their market positions.

[13] This holds in either case: if Person is a data controller or processor then PURPOSE defines why DATA of data subjects need to be collected (or processed), or if Person is a data subject then PURPOSE defines for which purposes the data subject allows collection or processing of DATA.

Trust & Reputation Evaluation System

Actors

Person – a legal or natural person;
Peer – a legal or natural person requesting a resource or data of Person or otherwise challenging Person's privacy in a way so that Person has to disclose some data to Peer.

Properties

- Confidence: Person is able to deduce the degree of confidence in Peer based on rigorous trust / reputation evaluation model;
- Retribution: Person is able to reflect on the past experience with Peer in terms of trust and reputation and give Peer credit or accusation.

Context

- Trust Web: a community of persons with mutual trust relationships which also define the level of how high the whole of the community ranks each of the persons in terms of reputation;
- Trust & Reputation Manager: a tool on each person's node which makes possible assessment of trust and reputation by a rigorous trust / reputation evaluation model;
- Authority: a public body or agency moderating how persons can influence Trust Web and trust / reputation values.

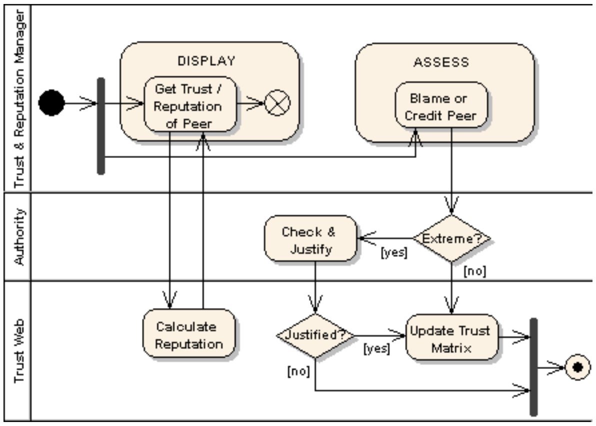
Description

(Every) Person has his own trust assessment of (Every) Peer, or at least of the close neighbours, and can adjust that using Trust & Reputation Manager. When Person wants to know reputation of Peer Trust & Reputation Manager can be used to obtain the grand total of trusts the whole Trust Web has in Peer. Trust & Reputation Manager should not allow changing Peer's reputation directly; instead, reputation should always be a cumulative of Trust Web based on trust evaluation model. In order to avoid abuse even Person's assessment of trust should not be unconditionally released into Trust Web so that possibility of unduly destroying reputation of a person is as low as possible: at least every request to gravely reduce reputation by a majority should be mediated by a special Authority.

Controls

Trust & Reputation Manager should have at least the following controls:
- DISPLAY is where trust or reputation values of peers can be read;
- ASSESS is where good or bad trust assessments can be injected into Trust Web, subject to Authority.

Definition



The next pattern identifies the need for a way to resolve privacy protection issues between two privacy policies of two parties in detail. This is required by Principle of data subject consent, Principle of right to object processing of personal data, Finality principle, and Data minimisation principle that cannot be enforced in practise without touching every issue in the two privacy policies and trying to converge them into an agreement that will display resolutions about explicit consent, opting in or out, the agreed purposes for processing, and the exact data to be disclosed. This pattern can greatly alleviate Problem Situation 7 since the agreement exactly defines the extent of data required for the transaction, subject to privacy policies and the purposes for collection. The pattern is feasible, as we have discussed in Section 3.

Privacy Policy Negotiation

Actors

Initiator – a legal or natural person which applies to Responder for some reason;[14]
Responder – a legal or natural person claiming data of Initiator or otherwise challenging Initiator's privacy with assumption to assist Initiator.

Properties

- Proportionality: the extent of data to be disclosed is proportional to the legitimate purposes for collecting the data;
- Confidentiality: the data and the privacy policy parts not directly required for the final resolution are not disclosed.

Context

- Initiator's Privacy Policy: the privacy policy of Initiator as defined in the pattern Formally Provable Privacy Policies;
- Responder's Privacy Policy: the privacy policy of Responder as defined in the pattern Formally Provable Privacy Policies;
- Agreement: an intersection of both the Privacy Policies with excerpts relating to the issues relevant for the actual situation;
- Initiator's Negotiation Agent: an agent system capable of negotiating issues of Initiator's Privacy Policy;
- Responder's Negotiation Agent: an agent system capable of negotiating issues of Responder's Privacy Policy;
- Theorem Prover: an abstract notation for a system capable of deriving logical truths about Privacy Policy;
- Domain Model: a digital representation of the important notions with ontological relations from privacy protection;
- External Application: an abstract notation for any instance of an application which allows a Negotiation Agent to invoke a dialog with the owner of the privacy policy.

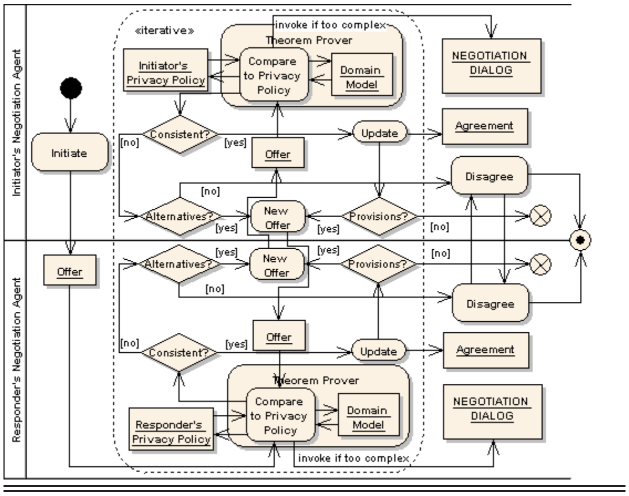
Description

The Initiator's Negotiation Agent and Responder's Negotiation Agent can automatically negotiate privacy protection issues between Initiator's and Responder's Privacy Policy. The complex reasoning for this is carried out by a Theorem Prover and associated Domain Model helps resolve the inherent semantic. The exchange of information between agents is done incrementally in packets called offers carrying only the data required to reach the final resolution and leaving the rest of the information undisclosed. In case both agree the final resolution is Agreement.

Controls

The External Application should support at least one control:
- NEGOTIATION DIALOG is a message display with appropriate controls which enable the owner of the privacy policy to interactively resolve the specific issues in negotiation that were not possible to resolve automatically.

Definition



As a follow up to the last pattern, it should be noted that Agreement should be regarded as a formal document and approved as such by legislation; it should be valid as a piece of evidence in a court case as any other formal contract is. To this end, Agreement should be archived using methods for long term trusted archiving of digital content [63].

The next pattern is about protecting identity of people in electronic transactions. It is deliberately that the very idea of digital identity on its own is not identified as a special privacy protection pattern; this is a security pattern used to identify and control what people do, whereas in privacy protection the aim is to protect the personal identifying information. Hence the following pattern proposes a special type of identity which still satisfies the security function but hides the true person behind. The Problem Situations 3, 4, 5 and 9 clearly show that the objection that such a pattern *only* supports people in illegal or illegitimate activities is far from correct. On the contrary, Problem Situation 5 shows that our personal identifying information is out there and since Problem Situations 3, 4 and 9 do happen Principle of fair and lawful processing, Finality principle and Data minimisation principle are impossible to enforce without hiding the real person behind the data disclosed in electronic transactions. The technologies required to implement this pattern range from identity federation frameworks which make possible referencing distributed personal identifying information, through protocols for data sharing and network addressing, and to anonymization models which assure unlinkability on network layer; these technologies will be referred to as the *identity infrastructure*.

This pattern isolates the personal identifiable information possible to infer out of the identity strictly to the provided information, depending on the strength of the pseudonymization. However, despite the pseudonymization such an identity can still be used for authentication and holders of such identities can still be accounted for their actions. This is possible in two ways. Either there is an authority that knows who is behind the pseudonym and can do legal interception; such a pseudonym can hide the holder and prevent Problem Situations 3 and 5 in the majority of cases with small actors in cyber crime; however, it cannot prevent Problem Situation 4 as authorities are typically involved.

Clearly, we need a way to provide people with a pseudonymous identity without possibility to trace the holder's real identity, but at the same time to prevent people to abuse them. This can be achieved through concept of *post*: a digital identity is pseudonymous to the degree whereby linking it to the true person is not possible, but in return the holder of such an identity should leave something of value (from now on referred to as *token*) at the post, but this is the only thing the holder needs to present of him/her/itself.[15] The digital identity should expire after some short amount of time; in order to get the token back or to renew the identity the holder should return to the post to do so. Meanwhile, if the digital identity in question was known to be used for illegitimate purposes, the holder can be held liable when coming to the post; if the holder does not come to the post then the token can be used to reimburse the people affected by the illegitimate uses of the identity. The token can be deposited as money, proportional to holder's income, and depending on how much money the holder has left at the post the pseudonymization can be weaker (i.e. subject to legal interception) or stronger (i.e. untraceable); for the period of time of validity of digital identity the agency running the post can make business using the money and at the revocation of the digital identity the money is returned to the holder with interest. Different agencies can compete providing better interest or better identity infrastructure.

It should also be clear that issuing a digital identity, which enables inspectors insight into parts of personal identifying information of the holder, has to be supported by appropriate access control to that information in order for the identity to be efficient: a party that was not given the identity explicitly by the holder should not have insight into the associated information. Mind also that standard measures

[14] In negotiation process also Initiator may claim data of Responder in order to aggregate enough proof for trustworthiness of Responder or to get access to Responder's resources – the process is symmetric in terms of the need for data exchange. Clearly, everything in privacy policy negotiation is about exchange of data: say one is negotiating for a service, then the service will be represented by a URI (i.e. a string of characters, thus data) so that the one can access it. Each resource in privacy policy negotiation is represented by appropriate data.

[15] That should say, one does not need to present physical appearance, real identity, bank account or any other identifying information.

for protecting integrity and confidentiality of that information are assumed to be in place using cryptography or other obfuscation techniques. These techniques do not constitute privacy protection patterns, at least they will not be identified as such in this paper; however they are important support for privacy protection and should not be neglected in realizations of privacy protection patterns. For example, before a digital identity has been issued all the data that are to be disclosed through this identity should already have been securely stored, the communications which are to be used to access these data should be protected by cryptographic protocols, and at the time the identity is defined access control should also be configured.

Virtual Identity

Actors

Person – a legal or natural person;
Inspector – a legal or natural person interested in Person;
Agency – a legal person providing Identity Infrastructure;
Authority – a legal person or other official body authorized for investigation and prosecution of cyber crime activities.

Properties

- Authenticity: demonstrates Person's credibility;
- Accountability: makes possible to hold Person liable;
- Pseudonimization: disturbs Inspector from knowing the true identity of Person;
- Unlinkability: makes hard to link the identifying information of Person outside of the directly provided data.

Context

- Person's Information: all the identifying information of Person, generally distributed on many places;
- Virtual Identity: a pseudonym of Person with references to pieces of Person's Information in function of a digital identity;
- Identity Infrastructure: a collection of technologies for federation of identity, single-sign-on, data sharing and addressing, and anonymization assuring unlinkability on network layer;
- Post: a place run by Agency where Person leaves a valuable when applying for Virtual Identity.

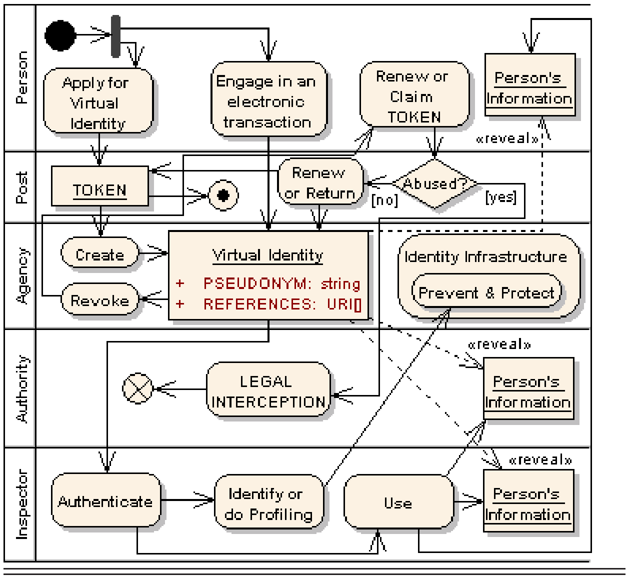
Description

Person's Information is distributed and known to public. Still, privacy protection can be achieved by using pseudonyms instead of real identities. Person can select arbitrary pseudonym and define the parts of Person's Information to be accrued to that pseudonym. This is made into a Virtual Identity and is issued by Agency after Person has left something of value at Post. Agency assures the properties of this pattern. After certain amount of time Virtual Identity expires and should be renewed by Person coming to Post. In case Person has abused Virtual Identity, Agency can withhold Person and call Authority for legal interception. When Virtual Identity is revoked, Agency has to return the valuable to Person with interest.

Controls

- PSEUDONYM is a string of characters used as the recognizing name of Virtual Identity;
- REFERENCES is a list or URIs to parts of Person's Information, packed into Virtual Identity, which represents its attributes;
- TOKEN is money deposited as valuable at Post;
- LEGAL INTERCEPTION is a procedure performed by Authority to find and prosecute the holder of Virtual Identity.

Definition



However, one of the supporting technologies for the previous pattern has for a long time been a synonym for privacy preserving internet browsing and definitelly deserves recognition as a privacy protection pattern.

Anonymizer

Actors

Person – a legal or natural person;
Inspector – a person looking for the true identity of Person.

Properties

- Anonymity: not being possible to trace the exact originator of a transaction.

Context

- Environment: a specific social or / and technical setting with infrastructure and corresponding controllers where communication between Person and Inspector is enabled;
- Anonymity Set: a collection of nodes which exhibit exactly the same traits in terms of capabilities that Environment offers Inspector for distinguishing among them;
- Anonymizer: a special technology or system which makes possible the appearance of Anonymity Set within Environment;
- Transaction: a communication between Person and Inspector carried out though use of facilities in Environment.

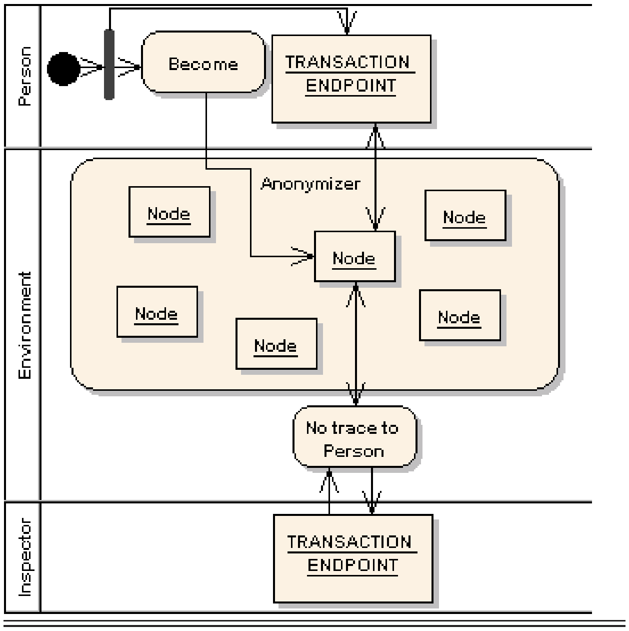
Description

In Environment Person appears as a node of Anonymity Set. Anonymizer makes impossible for Inspector to distinguish between Person and other nodes. This also implies that it is impossible for Inspector to trace, profile, or otherwise categorize objects that would make possible any kind of identification of Person inside Environment.

Controls

- TRANSACTION ENDPOINT is a system which makes possible for Person or Inspector to participate in Transaction and control it to certain extent.

Definition



Without this pattern Virtual Identity would be inefficient on the network layer since every attempt to conceal the true identity with a pseudonym could fail due to resolution of network identifiers frequently used by holder of virtual identity if they were (and normally they are) possible to trace back to the holder.

The next pattern speaks of a social model for data protection with using insurance for privacy breaches. Problematic situations, especially Problem Situation 3, have already given inspiration to some of the insurance companies[16] to think of insurance products that would “compensate for the liability arising from failure of network security protections, failure to protect or wrongful disclosure of private or confidential information, failure to protect personal identifying information from misuse or theft, or violation of any federal, state or local privacy statute alleged in connection with the failure to protect personal identifying information.”[73] The products cover

- expense of third party damages and legal claims;
- fines and penalties imposed by federal, state, and local governments;
- the expense incurred in notifying customers of a breach and the cost of mitigating reputational damage done;
- expense of defense costs within policy limits;
- expense incurred repairing or cleaning up the breach; and
- expense of fines levied by banks and credit card companies due to a privacy breach.

Privacy Breach Insurance

Actors

Person – a legal or natural person;
Insurance – a legal person involved in insurance business;
Perpetrator – a person responsible for the privacy breach.

Properties

- Safety: absence of catastrophic consequences.

Context

A specific socio-technical setting which is different for every case in separate with social and technical factors such as network infrastructure, market players, technical facilities and other installations and their deficiencies which make possible the privacy breach. In all that particular Person’s Data play the central role as their exploit is in the interest of Perpetrator and their protection is the aim of Insurance.

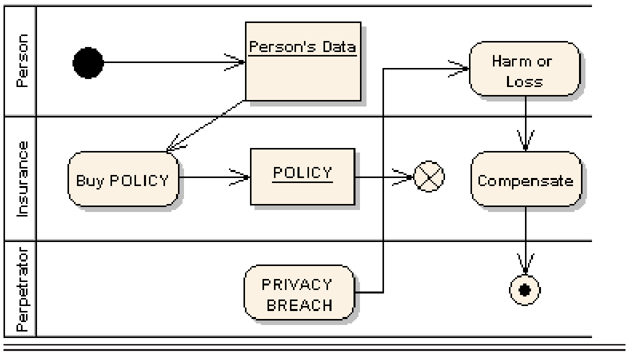
Description

At Insurance Person has bought an insurance policy for protecting Person’s Data. By some device Perpetrator is able to exploit the context and gain illegitimate access or control over Person’s Data. Perpetrator moreover causes harm or loss to Person by abusing Person’s Data. Person can claim compensation at Insurance to cover expenses of recovering from that harm or loss.

Controls

- POLICY is a contract between Person and Insurance on the extent of protection;
- PRIVACY BREACH is an exploit of context done by Perpetrator whereby Person’s Data are abused causing harm or loss to Person.

Definition



At that point it should be pointed out that all the privacy protection patterns which have been presente until now protect personal identifying information before it has been disclosed. This is referred to as *a-priori privacy protection*. As opposed to this a set of techniques is known to be able to (at least partially) protect privacy after the privacy identifying information was disclosed and are referred to as *a-posteriori privacy protection*. The remaining of the privacy

[16] cf. http://www.aig.com/Network-Security-and-Privacy-Insurance-(AIG-netAdvantage)_20_2141.html

protection patterns in this paper will focus on that last category and will close the whole cycle of privacy protection, as indicated in the introduction.

The first and probably the most important *a-posteriori* privacy protection pattern is that of a sticky policy. In this paper *sticky policy* is referred to as a fragment of Agreement as defined in pattern Privacy Policy Negotiation which pertains to an exact piece of private identifying information; such a fragment should hold only the certifying information and privacy protection rules that are somehow important for the protection of different parts of data of that piece of private identifying information. Any data without a sticky policy should be invalid and handling data without a sticky policy attached should be illegitimate.

This pattern is a prerequisite for the efficient performance of most of the following patterns of *a-posteriori* privacy protection and by this way importantly contributes to protection against some of the most problematic social situations, such as Problem Situations 4, 5 and 6. It should be clear that without such a pattern Finality principle and Time minimisation principle are impossible to enforce.

Sticky Policy

Actors

Data Subject – the person Data and Sticky Policy refer to as the subject;
Data Controller – the person which collects Data and defines purposes and ways how Data will be processed;
Data Processor – the person which actually processes Data.

Properties

- Intendment: the sense in which Data Controller should interpret Data Subject's volition about the way Data should be handled.

Context

- Data: a part of personal identifying information of Data Subject;
- Sticky Policy: a part of Agreement as defined in pattern Privacy Policy Negotiation adhering to Data;
- Data Token: the combined information made of Data and Sticky Policy and cryptographically signed by Data Subject.

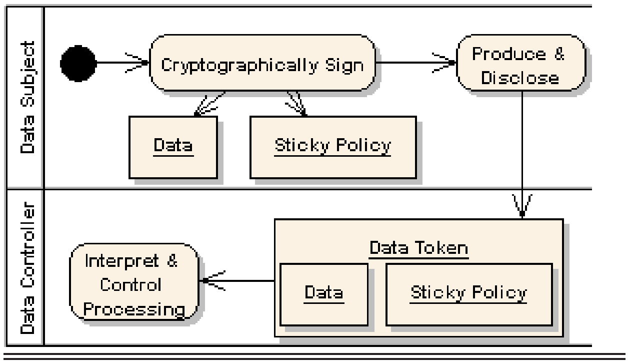
Description

Data and Sticky Policy are cryptographically signed by Data Subject and integrity of Sticky Policy adhering to Data is thus preserved in Data Token. Sticky Policy contains all the constraints and rules of how Data Processor should handle Data. Handling Data in disaccord with Sticky Policy is illegitimate.

Controls

- ENTITY identifies Data Subject;
- DATA describes Data;
- PURPOSE specifies the allowed purposes for the processing of DATA;
- RECIPIENT identifies allowed recipients of DATA;
- RETENTION indicates the kind of retention admissible for DATA;
- OPERATIONS defines the operations allowed on DATA;
- OBLIGATIONS states how Data Controller and Data Processor should protect DATA when handling them.

Definition



The next pattern can be of a great value when preserving Problem Situations 3, 4 and 5. Also Time minimisation principle or other forms of limitation of insight into data have a straightforward enforcement in this pattern. Technologies that support it have been outlined at the end of Section 3.

Privacy Preserving Data Processing

Actors

Data Subject – the person Data refer to;
Data Controller – the person which collects Data and stores them into Data Base;
Data Processor – the person which processes Data on behalf of Data Controller.

Properties

- Confidentiality: complete or selective hiding of data required for producing the final result of processing.

Context

- Data: a part of personal identifying information of Data Subject or other data owned by Data Processor;
- Data Base: the place where Data are stored at Data Controller, res. Data Processor, but under authority of Data Controller.

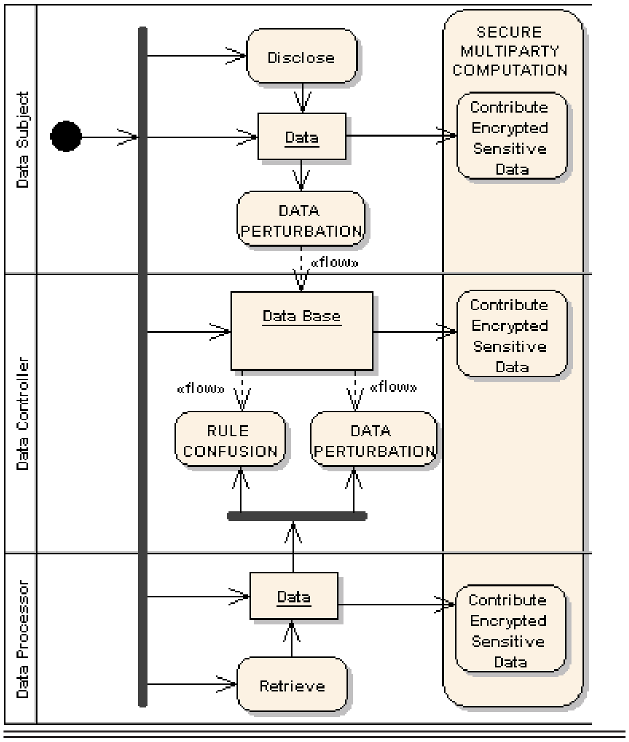
Description

Privacy of Data can be preserved at different stages before or at the actual time of processing. The first stage is at the disclosure of Data by Data Subject or Data Controller when DATA PERTURBATION can be used. The second stage is when Data are released to Data Processor from Data Base and at that point Data Controller can use RULE CONFUSION. The last stage is at the point of processing when SECURE MULTIPARTY COMPUTATION can be used.

Controls

- DATA PERTURBATION enables perturbation of Data so that their actual values are obfuscated, yet techniques exist which enable Data Processor to produce results of processing based on perturbed data of comparable quality as those obtained from the original Data;
- RULE CONFUSION makes possible truncation of data which gives Data Processor less cues to categorize raw Data and infer rules about their implicit associations, thus disabling Data profiling and deduction of sensitive personal identifying information, however this proportionally degrades the quality of results of processing;
- SECURE MULTIPARTY COMPUTATION enables parties to contribute encrypted or otherwise obfuscated Data to a processing which is capable to produce the same results out of these Data as though they were not obfuscated.

Definition



The following pattern describes one of the most typical data protection techniques, namely that of access control as described at the end of Section 3. The pattern, though, will try to introduce the idea of sticky policy into the decision process for controlling access. In such a setting this pattern is a strict precondition for almost any other privacy protection pattern, since without a close control which data and under what conditions and purposes are disclosed to whom also patterns such as Virtual Identity, Privacy Policy Negotiation or Sticky Policy have no actual effect.

Many problem situations need direct protection against intrusion in private data which this pattern can offer. Protection against Problem Situations 3, 4, 5, 6 and 7 can only be achieved with such a pattern in place and legal principles such as Finality principle, Data minimisation principle and Principle of right to object processing of personal data without this pattern have no definite power.

Access Control

Actors

Resource Controller – the person which controls access to Resource;
Requester – the person which requests Resource.

Properties

- Authorization: access to Resource is authorized according to predefined rules.

Context

- Resource: a particular resource requested by Requester;
- Access Control List: a list of requesters and their privileges to access resources;
- Credential: a piece of data encoding a voucher which somehow entails that Requester is entitled to access Resource;
- Sticky Policy: sticky policy adhering to Resource as defined in pattern Sticky Policy;
- Enforcement Point: the part of access control at the side where Resource is available;
- Decision Point: the part of access control where the decision about access to Resource is taken.

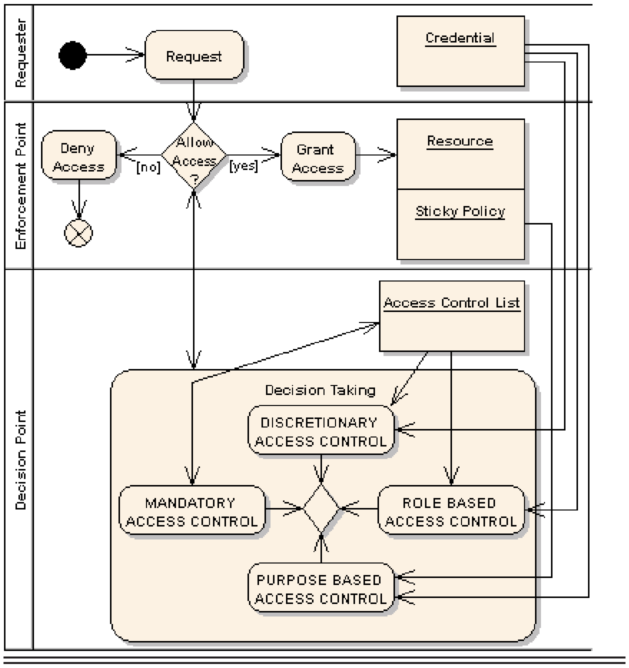
Description

Requester requests Resource at Enforcement Point which is in position to give access to Resource. However, for each type of resource a different Enforcement Point has to be implemented. Decision Point on the other hand can be centralized because decision can be taken based on formal methods and categorization of different types of resources, without need to know the special handling of each type of resource. Upon request Enforcement Point asks Decision Point whetehr access to Resource for Requester is allowed and if yes, then Enforcement Point enables Requester to use Resource. Decision is produced based on four different possible decision methods as described in **Controls** section.

Controls

Controlling that pattern is achieved through decision methods used by Decision Point, the four of them that have been used so far are:
- DISCRETIONARY ACCESS CONTROL enables granting access to Resource for Requester by defining this in Access Control List or by checking appropriate Credential of Requester;
- MANDATORY ACCESS CONTROL enables exact evaluation of sensitivities assigned to Requester and Resource as defined in Access Control List for the decision on whether Requester's sensitivity justifies access to Resource;
- ROLE BASED ACCESS CONTROL enables comparison of role of Requester, explicated by a suitable Credential, to permissions for accessing Resource as defined in Access Control List;
- PURPOSE BASED ACCESS CONTROL enables comparing purposes Requester shows by Credential stating Requester's legitimate business interests (or other kind of activities) to the actions allowed on Resource by Sticky Policy.

Definition



When data are processed, a record of all the actions and their committers should be preserved for later reference. On suspicion of abuse authorities should be able to access the logs and perform auditing. This can be achieved in two advancing levels of sophistication. On the basic level a record of the requester, the purpose or other inquiry related meta information, the time and the actual data requested should be taken at the point of access to the data by access control. It should be clear, however, that after this the data processor can freely ignore the sticky policy and handle data in disaccord with the data subject's will. This can be prevented by a notably more rigorous approach where data are held entirely in encrypted form and for any operation special trusted hardware modules (e.g. special hardware provided by trusted producers) have to be used that are able to perform computation based on techniques of secure multiparty computation or privacy homomorphisms as discussed at the end of Section 3. Using appropriate cryptosystems when encrypting data at disclosure data processors would be forced to design their software to make use of the application programming interface of trusted hardware module, because they would be unable to perform these operations themselves. This way data could only be processed using such trusted hardware modules which could then actually take records of details about who requested the operation, what was the operation, at which time and on what data it was performed, and whether this was in accord with sticky policy.

As it was pointed out at the end of Section 3 the techniques used for computing privacy homomorphisms may lead to severe message expension and this opens questions about feasibility of the second sophistication level. Nevertheless, there exist several efficient privacy homomorphism operations and other techniques from secure multiparty computation and, moreover, research could be done for optimization of complex parts in order to make possible entirely encrypted processing. Thus the second complexity level will be regarded as an interesting prospect and as a necessary conceptual part of the following pattern.

This pattern has potential to greatly mitigate Problem Situations 4 and 7, actually prosecution in case of any problematic situation can only be efficient with such an evidence in place. This evidence can be used in courts to prove violations of data protection principles in general.

Privacy Audit Trail

Actors

Data Subject – the person Data refer to;

Data Controller – the person which collects Data and defines rules for their processing;

Data Processor – the person which processes Data on behalf of Data Controller.

Properties

- Accounting: the ability to hold Data Processor liable for their actions;
- Auditing: the possibility to have insight into the actions of Data Processor.

Context

- Data: a particular data processed by Data Processor;
- Sticky Policy: the sticky policy, as defined in the pattern Sticky Policy, adhering to Data;
- Access Control: a component enabling control of access to data as defined in pattern Access Control;
- Trusted Hardware Module: a piece of hardware capable of secure encryption and decryption and running protocols for various operations based on techniques of secure multiparty computation and privacy homomorphisms;
- Log: a part of Trusted Hardware Module or Access Control where details of every operation performed on data are recorded, such as the requester, the action, and the time.

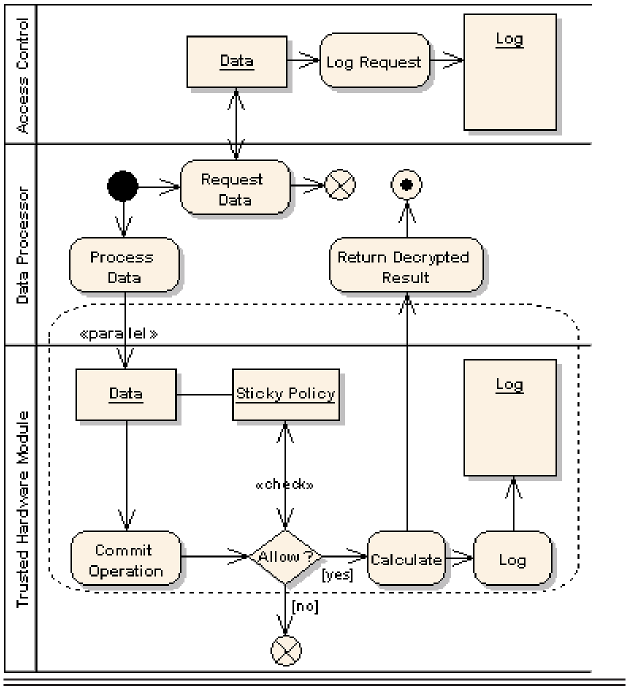
Description

At every request for Data Access Control logs Data Processor, the purpose, role, credentials, or other meta information, the time and the actual Data requested. Every processing of Data must be carried out by use of Trusted Hardware Module which logs the kind of operation, whether it was in accord with Sticky Policy, the Data Processor, the time and other important information. If the operation is in accord with Sticky Policy then the operation is performed. When Data are processed inside Trusted Hardware Module, they are kept in an encrypted form and special protocols enable processing despite their encrypted form; result is returned to Data Processor in a decrypted form.

Controls

Pattern changes behaviour depending on Sticky Policy.

Definition



The last pattern suggests a very social model for protecting privacy, namely that of a cyber police and prosecution through civil court. It should be clear that, ultimately, privacy can not be protected unless if legal frameworks and prosecution are not an integral part of the whole privacy protection paradigm.

Privacy Breach Prosecution

Actors

Perpetrator – the person who treats Data in disaccord with Sticky Policy;

Authorities – the agency or state department authorized to prosecute Perpetrator.

Properties

- Prosecutability: the ability to conduct criminal proceedings in court against Perpetrator.

Context

- Data: a particular data processed by Data Processor;
- Sticky Policy: the sticky policy as defined in pattern Sticky Policy adhering to Data;
- Privacy Audit Trail: logs obtained from privacy auditing as described in pattern Privacy Audit Trail.

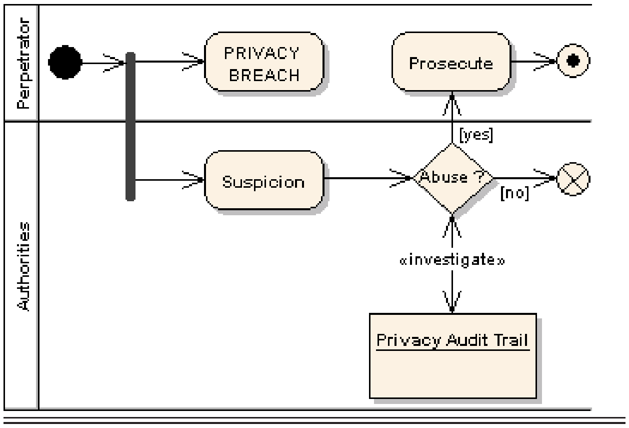
Description

When there exists a justified suspicion that Perpetrator has abused Data Authorities should collect relevant Privacy Audit Trail and check Perpetrator's actions regarding Data. In case Perpetrator really has abused data Authorities will prosecute Perpetrator.

Controls

- PRIVACY BREACH is an exploit of context done by Perpetrator whereby Data are handled in disaccord with Sticky Policy.

Definition



5. Privacy Protection Cycle

The proposed patterns can be combined into a higher order integration scheme showing how the patterns should be deployed in a real situation to make possible a systemic privacy protection. This integration scheme is referred to as the *privacy protection cycle* and is represented by the diagram on Figure 1.

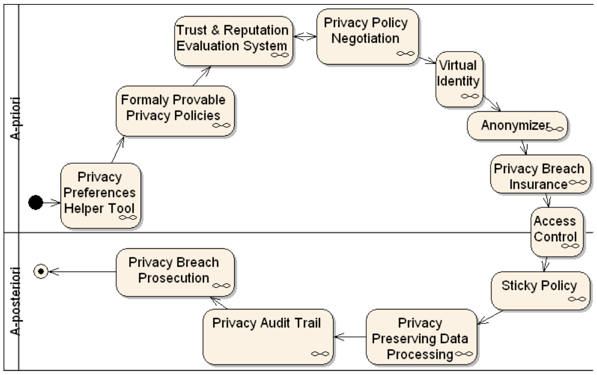

A-priori
Privacy Preferences Helper Tool
Formaly Provable Privacy Policies
Trust & Reputation Evaluation System
Privacy Policy Negotiation
Virtual Identity
Anonymizer
Privacy Breach Insurance
Access Control
A-posteriori
Sticky Policy
Privacy Preserving Data Processing
Privacy Audit Trail
Privacy Breach Prosecution

Figure 1. Integration scheme for privacy protection cycle

It should be noted that External Application of Privacy Preferences Helper Tool pattern was actually meant to be the Negotiation Agent of Privacy Policy Negotiation pattern; on the other hand, External Application of Privacy Policy Negotiation is Privacy Preferences Helper Tool. There are many more important correlations between patterns implied in privacy protection cycle. First of all, Privacy Policy Negotiation, although this was not explicitly mentioned in that pattern, should reflect on the level of reputation of peer when processing offers against privacy policy; privacy protection rules inside the privacy policy (e.g. subject to OBLIGATIONS) should define the level of reputation peer should meet in order to be allowed access to the resource; this way, if the reputation of data controller or processor would be too low, they would be refused at privacy policy negotiation time and automatically demand for their services would fall. This is the indispensable correlation between Trust & Reputation Evaluation System and Privacy Policy Negotiation if it should be possible to penalize misbehaving data processors or controllers automatically based on lower trust in electronic transactions. Close to this is another correlation which is that Authority of Trust & Reputation Evaluation System pattern could actually be Insurance of Privacy Breach Insurance pattern, because for Insurance it will be mandatory to be involved in investigations of privacy breaches; moreover, not only that Perpetrator of Privacy Breach Insurance should be blamed and their reputation lowered, but also if they were insured against abuse of private identifying information, then they should pay more for POLICY.

A more intuitive and illustrative presentation of privacy protection cycle is given on Figure 2. In this paper authors maintain that no weaker or partial scheme can be sufficient for actually protecting privacy of people given the facts pointed out in Subsection 2.2. Only a very systemic scheme where technologies and social models cooperate to make a total cover up of potential privacy threats can be regarded as a prospect towards a future where our private identifying information will be respected and the tremendous potential and capabilities of information technology for its exploitation regulated to the extent where the abuse in all its expressions will become an unlikely experience.

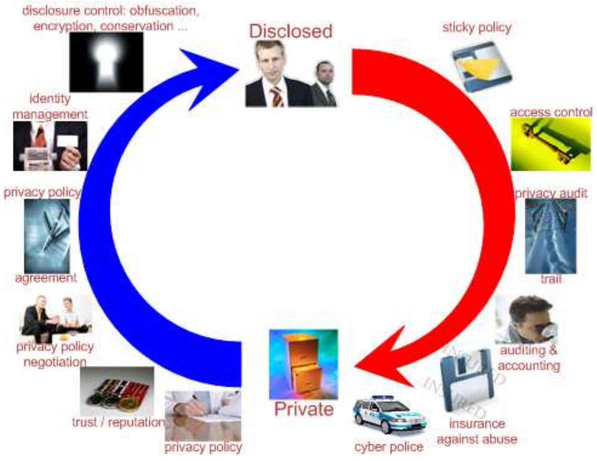


Figure 2. Intuitive representation of privacy protection cycle

6. Conclusions

A close look at the situation regarding protection of privacy shows that privacy related problems already are – and will in future become even more – a real and serious social problem. A thorough review of the state-of-the-art and existing research on the field of privacy protection reveals quite a rich assortment of potentially very powerful techniques for protecting privacy, but unfortunately legislation does not acknowledge them and brings no central organization into the field of privacy protection. Public initiatives and technologies which could potentially answer them are kept separated and social, legal and administrative frameworks that could promote and back up the technologies have not been established.

Authors of this paper believe that technologies have reached the point where most of the privacy protection patterns described in this paper can be implemented in their full potential and that the obstruction lies in the fact that there is no commercial interest in producing the kind of technology whatsoever on the current information technology market. This in turn owes much to the deficiency of privacy protection related legislation, which does not require of data controllers and processors to implement concrete schemes for privacy protection. For an illustration, if one takes Finality principle, it would make a lot of difference if instead of saying that

> *"Personal data must be collected for specified, explicit and legitimate purposes and may not be further processed in a way incompatible with those purposes,"*

regulation would additionally demand that

> *"Prior to processing of personal data Privacy Policy Negotiation pattern should be implemented on the part of data collection and any subsequent disclosure of data should be done with respect to Sticky Policy pattern and Privacy Audit Trail pattern should be implemented on the part of data processing."*

This way data controllers and data processors would be obliged to install in their information systems the software implementing the required privacy protection patterns which would create a real demand for such kind of software on information technology market. In this same light concepts such as virtual identity, privacy breach insurance or privacy preserving data processing would be given their exact and compelling legal formulations and regulation would actually fulfill its part in making privacy protection cycle a reality for the society.

References

[1] T. Otter, "Data protection law: The Cinderella of the software industry?" Computer law & security report 23, 2007, pp. 67-72.

[2] DIRECTIVE 95/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal L No.281, 23.11.1995.

[3] Y. Punie, S. Delaitre, I. Maghiros, D. Wright, (eds.) "Dark scenarios in ambient intelligence: Highlighting risks and vulnerabilities," SWAMI Deliverable D2. A report of the SWAMI consortium to the European Commission under contract 006507, November 2005. http://swami.jrc.es

[4] Convention for the Protection of Human Rights and Fundamental Freedoms, Rome, 4.11.1950.

[5] Convention on Cybercrime, Budapest, 23.11.2001. http://conventions.coe.int/Treaty/EN/Treaties/Html/185.htm

[6] DIRECTIVE 2000/31/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce), Official Journal of the European Communities L 178/1, 17.7.2000.

[7] DIRECTIVE 2002/58/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), Official Journal of the European Communities L 201/37, 31.7.2002.

[8] Privacy Bird. http://www.privacybird.org

[9] L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle, "The Platform for Privacy Preferences 1.0 (P3P1.0) Specification," World Wide Web Consortium Recommendation, April 2002. http://www.w3.org/TR/P3P/

[10] L. Cranor, M. Langheinrich, and M. Marchiori, "A P3P Preference Exchange Language 1.0 (APPEL1.0) W3C Working Draft," 2002.

[11] P. Ashley, S. Hada, G. Karjoth, C. Powers, and M. Schunter, "Enterprise Privacy Authorization Language (EPAL 1.2) specification submitted to W3C," 2003. http://www.w3.org/Submission/2003/SUBM-EPAL-20031110/

[12] T. Moses, "eXtensible Access Control Markup Language 3 (XACML), Version 2.0," OASIS eXtensible Access Control Markup Language Committee Specification, 1 Feb 2005. http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf

[13] SAML. http://saml.xml.org/saml-specifications

[14] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider: The Description Logic Handbook: Theory, Implementation, Applications. Cambridge University Press, Cambridge, UK, 2003. ISBN 0-521-78176-0

[15] OWL Web Ontology Language Overview, W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-features/

[16] http://kaon.semanticweb.org/

[17] http://kaon2.semanticweb.org/

[18] Travis Leithead, Wolfgang Nejdl, Daniel Olmedilla, Kent E. eamons, Marianne Winslett, Ting Yu, and Charles C. Zhang: "How to Exploit Ontologies in Trust Negotiation," ISWC – Workshop on Trust, Security, and Reputation on the Semantic Web, Hiroshima, Japan, November 7, 2004.

[19] Valentina Tamma, Michael Wooldridge, Ian Dickinson: "An Ontology Based Approach to Automated Negotiation," proceedings of the Fourth International Workshop on Agent-Mediated Elctronic Commerce (AMEC-2002), Bologna, Italy, July 2002.

[20] J. G. Cederquist, R. Corin, M. A. C. Dekker, S. Etalle, J. I. den Hartog, G. Lenzini, "Audit-based compliance control," International Journal of Information Security archive, Volume 6, Issue 2, Pages: 133 – 151, March 2007.

[21] S. Etalle, W. H. Winsborough, "A Posteriori Compliance Control," 12^{th} ACM Symposium on Access Control Models and Technologies (SACMAT), 20-22 June 2007, Nice, France. pp. 11 – 20.

[22] M. D'Agostino, D. Gabbay, R. Haehnle, J. Posegga (Eds), "Handbook of Tableau Methods," Kluwer,1999.

[23] J. Alan Robinson, "A Machine-Oriented Logic Based on the Resolution Principle," Journal of the ACM (JACM), Volume 12, Issue 1, pp. 23-41.

[24] http://www.cs.unm.edu/m̃ccune/otter/

[25] http://coq.inria.fr/

[26] Term Rewriting and All That. (1999) Baader, F. and Nipkow, T. Cambridge University Press.

[27] First Order Logic and Automated Theorem Proving. (1996) Fitting, M. Springer Verlag.

[28] http://owl.man.ac.uk/factplusplus

[29] http://www.mindswap.org/2003/pellet/index.shtml

[30] http://dl.kr.org/dig/

[31] http://jena.sourceforge.net/

[32] M. Nielsen, K. Krukow, "Towards a Formal Notion of Trust," PPDP'03 August 27-29, 2003, Uppsala, Sweden.

[33] A. Jøsang, "A logic for uncertain probabilities," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 9, No. 3 (June 2001).

[34] A. Jøsang, R. Ismail, C. Boyd, "A Survey of Trust and Reputation Systems for Online Service Provision," Decision Support Systems, 43(2), pages 618-644, March 2007.

[35] A. Jøsang, R. Hayward, S. Pope, "Trust Network Analysis with Subjective Logic," (ACSC2006), Hobart, Tasmania, Australia, January 2006.

[36] M. Carbone, M. Nielsen, V. Sassone, "A Formal Model for Trust in Dynamic Networks," In Proc. of International Conference on Software Engineering and Formal Methods (SEFM 2003), p. 54 – 63.

[37] M. Richardson, R. Agrawal, P. Domingos, "Trust Management for the Semantic Web," in Proceedings of the Second International Semantic Web Conference 2003, p. 351 – 368.

[38] Todd Barlow, Adam Hess, Kent E. Seamons: "Trust Negotiation in Electronic Markets," proceedings of the eighth research symposium on emerging electronic markets (RSEEM 01), Maastricht, The Netherlands, September 16-18, 2001.

[39] Wolfgang Nejdl, Daniel Olmedilla, and Marianne Winslett: "PeerTrust: Automated Trust Negotiation for Peers on the Semantic Web," proceedings of Secure Data Management, VLDB 2004 Workshop, Toronto, Canada, August 30, 2004.

[40] C.A. Ardagna, E. Damiani, S. De Capitani di Vimercati, and P. Samarati: "Towards Privacy-Enhanced Authorization Policies and Languages," in Proc. of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security (IFIP), Nathan Hale Inn, University of Connecticut, Storrs, USA, August 2005.

[41] K. Dolinar, "Daidalos Deliverable DII-A451, Architecture and Design: Security and Privacy for Pervasive Systems," DAIDALOS FP6 project consortium, 2006.

[42] J. Miller, "Yadis Specification, Version 1.0." The Identity and Accountability Foundation for Web 2.0, 18 March 2006. http://yadis.org/papers/yadis-v1.0.pdf

[43] OpenID. http://openid.net/

[44] Light-Weight Identity. http://lid.netmesh.org/wiki/Main_Page

[45] D. Reed, D. McAlpin, "Extensible Resource Identifier (XRI) Syntax, V2.0," OASIS Extensible Resource Identifier (XRI) Committee Specification, 14 November 2005. http://www.oasis-open.org/committees/download.php/15377/xri-syntax-V2.0-cs.pdf

[46] D. Reed, M. Sabadello, P. Trevithick, "The XDI RDF Model V11," OASIS XRI Data Interchange Comitee Specifications, 21. 10. 2008. http://www.oasis-open.org/committees/download.php/29748/xdi-rdf-model-v11.pdf

[47] Simple eXtensible Identity Protocol. http://www.sxip.com/background

[48] Project Liberty. http://www.projectliberty.org/

[49] T. Wason, "Liberty ID-FF Architecture Overview," Version: 1.2-errata-v1.0.

[50] Shibboleth Initiative. https://spaces.internet2.edu/display/SHIB/WebHome

[51] Michael B. Jones, "A One-Page Introduction to Windows CardSpace," MSDN, Microsoft Corporation, January 2007. http://msdn.microsoft.com/sl-si/netframework/cc196951(en-us).aspx

[52] Higgins trust framework. http://www.eclipse.org/higgins

[53] J. Camenisch, A. Shelat, D. Sommer, S. Fischer-Hbner, M. Hansen, H. Krasemann, G. Lacoste, R. Leenes, J. Tseng, "Privacy and Identity Management for Everyone," in ACM DIM 2005. http://www.zurich.ibm.com/%7Ejca/papers/cssf05.pdf

[54] A. Pfitzmann, M. Hansen, “Anonymity, Unlinkability, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for Terminology,” version 0.31, Feb 15 2008. http://dud.inf.tu-dresden.de/Anon_Terminology.shtml

[55] S. Fischer-Hübner, C. Andersson, “PRIME Deliverable: D14.0a – Framework V0,” version 6, June 2005.

[56] D. Goldschlag, M. Reedy, P. Syversony, “Onion Routing for Anonymous and Private Internet Connections,” Communications of the ACM, vol. 42, num. 2, February 1999. http://www.onion-router.net/Publications/CACM-1999.pdf

[57] M. Reiter, A. Rubin, “Crowds: Anonymity for Web Transactions,” ACM Transactions on Information and System Security 1 (1), 23 November 2005. http://avirubin.com/crowds.pdf

[58] Tor: anonymity online. https://www.torproject.org/

[59] V. C. Hu, D. F. Ferraiolo, D. R. Kuhn, “Assessment of Access Control Systems,” USA National Institute of Standards and Technology, Interagency Report 7316, September 2006. http://csrc.nist.gov/publications/nistir/7316/NISTIR-7316.pdf

[60] Ji-Won Byun, E. Bertino, N. Li, “Purpose Based Access Control of Complex Data for Privacy Protection,” SACMAT05, June 1-3, 2005, Stockholm, Sweden.

[61] N. F. Johnson, S. Jajodia, “Exploring steganography: Seeing the unseen,” Computer (1998A) 31(2):26-34. http://www.jjtc.com/pub/r2026.pdf

[62] R. Wishart, K. Henricksen, J. Indulska, “Context obfuscation for privacy via ontological descriptions,” 1^{st} International Workshop on Location- and Context-Awareness (LoCA), volume 1678 of Lecture Notes in Computer Science, pages 276-288, Springer, 2005. http://henricksen.id.au/publications/LoCa05.pdf

[63] A. J. Blazic, “Long Term Trusted Archive Services,” First International Conference on the Digital Society (ICDS’07), pp.29, 2007.

[64] Ronald L. Rivest, Len Adleman, Michael L. Dertouzous, “On data banks and privacy homomorphisms.” In Richard. A. Demillo, David P. Dobkin, Anita K. Jones, and Richard J. Lipton, editors, “Foundations of Secure Computations,” pages 169-177. Academic Press, New York, 1978.

[65] Kevin Henry, “The Theory and Applications of Homomorphic Cryptography.” A thesis presented to the University of Waterloo in fulfillment of the thesis requirement for the degree of Master of Mathematics in Computer Science Waterloo, Ontario, Canada, 2008.

[66] R. Agrawal, J. Kiernan, R. Srikant, Y. Xu, “Hippocratic Databases,” in VLDB, 2002, pp. 143 – 154.

[67] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, Y. Theodoridis, “State-of-the-art in Privacy Preserving Data Mining,” SIGMOD Record, Volume 33, 2004.

[68] R. Agrawal, R. Srikant, “Privacy-Preserving Data Mining,” ACM SIGMOD 2000 5/00 Dallas, TX, USA.

[69] D. Agrawal, C. C. Aggarwal, “On the Design and Quantification of Privacy Preserving Data Mining Algorithms,” PODS 2001, Santa Barbara, California, USA.

[70] C. Bryce, M.A.C. Dekker, S. Etalle, D. Le Métayer, F. Le Mouël, M. Minier, J. Moret-Bailly, S. Ubéda, “Ubiquitous Privacy Protection,” PRIAM Position Paper.

[71] G. Karjoth, M. Schunter, M. Waidner, “Platform for Enterprise Privacy Practices: Privacy-enabled Management of Customer Data,” 2^{nd} Workshop on Privacy Enhancing Technologies, Lecture Notes in Computer Science. Springer Verlag, 2002.

[72] Activity Diagram, “UML 2.0 Superstructure Specification,” Object Management Group, August 2003, p. 280.

[73] Jamey Heary, “Privacy Breach Insurance; new solution for mitigating the risk of credit card and identity breaches,” Cisco Security Expert on Tue, 03/18/08. http://www.networkworld.com/community/node/26132

A. Corroborations for Problem Situations

The following is a list of references to real affairs supporting the problem situations from subsection 2.2.

[4] There are known court cases for that problem: In Halford v. United Kingdom (27 May 1997), the Court introduced the criterion of the “reasonable expectations of privacy”. Miss Halford, a senior officer whose telephone calls were intercepted without warning, was granted privacy protection in her office space, although not absolute.

[5] cf. Zeller, Tom Jr., “Black Market in Stolen Credit Card Data Thrives on Internet”, The New York Times, 21 June 2005: “A ‘dump’, in the blunt vernacular of a relentlessly flourishing online black market, is a credit card number. And what Zo0mer is peddling is stolen account information - name, billing address, phone - for Gold Visa cards and MasterCards at $100 apiece.”
Vijayan, Jaikumar, “ID Theft Continues to Increase. More than 13 million Americans have been victimized, new study reveals,” Computerworld, 30 July 2003. http://www.pcworld.com/news/article/0,aid,111832,00.asp
Zetter, Kim, “TSA Data Dump Leads to Lawsuit”, Wired News, 14 July 2005. http://www.wired.com/news/privacy/0,1848,68560,00.html

[6] See Zeller, Tom Jr, “For Victims, Repairing ID Theft Can Be Grueling,” The New York Times, 1 Oct 2005. The story reports cases where victims have been trying to overcome the consequences of identity theft for more than two years: “Victims are still left with the unsettling realization that the keys to their inner lives as consumers, as taxpayers, as patients, as drivers and as homeowners have been picked from their pockets and distributed among thieves.”
cf. Solove, p. 110: “Identity theft can be a harrowing experience. According to estimates, a victim typically spends over two years and close to 200 hours to repair the damage that identity theft causes.” And p. 110: “Most identity thefts remain unsolved. Research firm Gartner Inc estimates that less than 1 in 700 instances of identity theft result in a conviction.”

[7] cf. OHarrow, Robert, No Place to Hide, p. 124: “LexisNexis, a subsidiary of the UK-based Reed Elsevier Group, maintains billions of records, including media reports, legal documents, and public records collected from thousands of sources around the world.”
cf. OHarrow, p. 34: “Acxiom is not a household name. But as a billion-dollar player in the data industry, with details about nearly every adult in the United States, it has as much reach into American life as Pepsi or Goodyear. You may not know about Acxiom, but it knows a lot about you.”
cf. OHarrow, p. 49: “‘InfoBase Enhancement’ enables Acxiom to take a single detail about a persons and append, on behalf of its customers, a massive dossier. This generally happened without the individual every knowing about it.”
cf. Perez, E.: “Identity theft puts pressure on data sellers,” The Wall Street Journal, 21 February 2005. http://www.post-gazette.com/pg/05052/460233.stm. A recent breakout: “Company ChoicePoint actually sold 145.000 peoples personal information to Nigerian scammers.”

[8] cf. Safire, William, “Goodbye To Privacy”, The New York Times, 10 April 2005: “Of all the companies in the security-industrial complex, none is more dominant or acquisitive than ChoicePoint of Alpharetta, Ga. This data giant collects, stores, analyzes and sells literally billions of demographic, marketing and criminal records to police departments and government agencies that might otherwise be criticized (or defunded) for building a national identity base to make American citizens prove they are who they say they are.”

[9] See, for example, OHarrow, p. 222: “HNC monitors 90 per cent of all credit cards in the United States and half of those in the rest of the world using artificial intelligence to seek out indications of fraud and deceit.”
Solove, Daniel J, The Digital Person, p. 20: “Wiland Services has constructed a database containing over 1,000 elements, from demographic information to behavioural data, on over 215 million people.”
Tuohey, Jasey, “Government Uses Color Laser Printer Technology to Track Documents. Practice embeds hidden, traceable data in every page printed”, 22 November 2004. http://www.pcworld.com/news/article/0,aid,118664,00.asp. See also Jardin, Xeni, “Your Identity, Open to All”, Wired News, 6 May 2005. http://www.wired.com/news/privacy/0,1848,67407,00.html

[10] Singel, Ryan, “Nun Terrorized by Terror Watch”, Wired News, 26 September 2005. http://www.wired.com/news/privacy/0,1848,68973,00.html

[11] cf. OHarrow, p. 48: “For years, the credit bureaus had been dogged by complaints. Information in their reports was chronically incorrect. They routinely failed to correct mistakes, and seemed arrogant when individuals called.”

[12] Krebs, Brian, ”Hacked Home PCs Fueling Rapid Growth in Online Fraud”, Washington Post, 19 September 19 2005. http://www.washingtonpost.com/wp-dyn/content/article/2005/09/19/AR2005091900026_pf.html

Threats to the Swarm: Security Considerations for Swarm Robotics

Fiona Higgins, Allan Tomlinson and Keith M. Martin

***Abstract*—Swarm robotics is a relatively new technology that is being explored for its potential use in a variety of different applications and environments. Previous emerging technologies have often overlooked security until later developmental stages, when security has had to be undesirably (and sometimes expensively) retrofitted. This paper compares swarm robotics with related technologies to identify their unique features where existing security mechanisms can not be applied. We then review some of the emerging applications where ineffective security could have significant impact. We conclude by discussing a number of security challenges for swarm robotics and argue that now is the right time to address these issues and seek solutions. We also identify several idiosyncrasies of swarm robotics that present some unique security challenges. In particular, swarms of robots potentially (i) employ different types of communication channels (ii) have special concepts of identity, and (iii) exhibit adaptive emergent behaviour which could be modified by an intruder. Addressing these issues now will prevent undesirable consequences for many applications of this type of technology.**

***Index Terms*—swarm robotics, security, autonomy, adaption, emergent behaviour**

I. Introduction

Swarm robotics is a relatively new area of research, and one which is growing rapidly. As with many emerging technologies, there is no formal definition of the field that engenders universal agreement, however comprehensive reviews of the state-of-the-art identify some characteristics that have been generally accepted [1]–[3]. These characteristics include robot autonomy; decentralised control; large numbers of member robots; collective emergent behaviour and local sensing and communication capabilities. Thus, from a security perspective, it is reasonable to consider swarm robotics as a special type of computer network with the aforementioned characteristics.

It has often been the case that the security of a new technology is an afterthought rather than an explicit design objective. This is not entirely surprising given the creative nature of research and the diversity of disciplines investigating the technology. Typically it is only as the technology matures, and begins to be deployed, that the security implications then become apparent. This was the case with, for example, mobile phone technology. The first generation of mobile phones were analogue, and easy to clone since they broadcast their identity clearly over the airwaves. It was also easy to eavesdrop on them by simply tuning a radio receiver to pick up conversations. Subsequently the underlying technology has been continuously modified in order to address threats that became apparent after deployment. The development of the Internet is another example of security being retrofitted to the technology.

In the case of swarm robotics, the particular security requirements of swarm robotic networks do not appear to have been investigated in any detail so far. Thus, for the above reasons, we believe that this is an opportune time to consider these issues, before any wide-scale deployment. Deferring security research until later in the technology's evolution could, depending on the application, be a risky strategy and may lead to undesirable consequences.

As far as we are aware, this is the first attempt to categorise security challenges for swarm robotics. Very little prior work appears to have been openly published. A notable exception to this is the work of Winfield and Nembrini [4] who identify several threats to a swarm of robots, which they classify as hazards. In identifying the main security challenges to swarm robotic networks it is our hope that this paper, and the work we presented at ICAS09 [5], will result in the development of robot swarm technology that is reliable and safe to deploy even in potentially hostile environments.

In Section II, we briefly review technologies that are similar to swarm robotics, highlighting the key differences and defining what we mean by a robotic swarm. In Section III we provide examples of applications that potentially will make use of this technology and show how vulnerabilities may exploited. In Section IV we discuss security, commencing with a short high level overview of security, and then cataloguing aspects of the swarm robotic environment which present challenges to security. Finally, in Section V we draw some conclusions.

II. Swarm Robotics and Related Technology

Before considering the security of swarm robotic networks it is necessary to establish the scope of the type of system we wish to secure. In other words, it is necessary to define what we mean by a swarm robotic network. There are many technologies which are similar to swarm robotic networks in some respects but differ in particular aspects. It is useful therefore to review how similar technologies, some of which have been subjected to a degree of security analysis, relate to robotic swarms. This will allow us to identify the unique features of robotic swarms that may benefit from closer scrutiny in terms of security. It is these unique features that we wish to focus on in order to identify vulnerabilities that are

The authors are with the Information Security Group, Royal Holloway, University of London, UK
f.l.higgins, allan.tomlinson, keith.martin@rhul.ac.uk

particularly pertinent to swarm robotic networks and perhaps to identify aspects of these systems that may be exploited to enhance security.

In this section we consider four technologies closely related to swarm robotic networks and then describe the unique distinguishing features of the latter.

A. Multi-Robot Systems

Like robotic swarms, multi-robot systems are a collection of robots, working together to achieve a common goal. To accomplish this, multi-robot systems are typically managed by a well–defined command and control structure. Swarm robotic systems differ from more traditional multi-robot systems in that their command and control structures are not hierarchical or centralised, but are fully distributed, self-organised and 'inspired by the collective behaviour of social insect colonies and other animal societies' [6].

Self-organisation means that sometimes the collective behaviour, even if unpredictable, may well result in solutions to problems that are superior to ones that could have been devised in advance. The parallel drawn with social societies in the animal world extends to communication – interactions between the robots can be indirect as well as direct. Fault tolerance, which is related to security, has already been extensively explored within the context of multi-robot systems with hierarchical command and control, notably in the work of Parker's ALLIANCE control architecture [7].

B. Mobile Sensor Networks

Sensor networks consist of collections of devices, or nodes, with sensors that typically communicate over a wireless network. A *mobile* sensor network is a sensor network where the nodes are either placed on objects which move [8] or where the nodes may move themselves [9]. In the latter case they are sometimes known as robotic sensor networks [10]. Hybrid systems also exist [11], where mobile robots work in conjunction with static sensors.

Although mobile sensor networks exhibit many similarities to swarm robotic networks, there are distinct differences. For example, robotic swarms may utilise a wider range of communications technologies, which extend to indirect communication such as stigmergy, as described in Section IV-D. Moreover, individual identity may be more significant in a sensor network if it is important to determine exactly where the sensed data originated.

Perhaps the most important difference is that a sensor network is not designed to have the collective emergent behaviour of a robotic swarm.

C. MANETs

Mobile Ad-hoc Networks (MANETs) consist of wireless mobile nodes that relay each others' traffic, with the nodes spontaneously forming the wireless network themselves. The special properties of MANETs, such as the lack of infrastructure, absence of trusted third parties, as well as possible resource constraints, make implementing security a very challenging task. MANETs can consists of many types of mobile devices and there is considerable existing work on their security [12], [13]. Although MANETs do not exhibit the emergent behaviour of swarms, some MANET security techniques could have relevance to swarm robotics depending on the communication method used by the swarm.

D. Software Agents

There is no universally agreed definition of a software agent, but we take one proposed by Wooldridge [14]: 'An *agent* is a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives'. A *multi-agent system* (MAS) [14], [15] is a system composed of multiple autonomous agents, where each agent cannot solve a problem unaided; there is no global system control; data is decentralised; and computation is asynchronous. A *mobile* agent is a particular class of agent with the ability during execution to migrate from one host to another where it can resume its execution [15]. Thus *mobile multi-agent systems* may share many features with swarm robotic systems, but in a virtual world.

Corresponding to the active interest in mobile software agents and their rapid adoption, there has been much interest in their security [15]. However this does not always translate easily to robotic swarms because of the particular characteristics of robotic swarms which differentiate them, such as their physical nature, diverse communication mechanisms and control structure.

E. Swarm Robotics

From the brief discussion above it is clear that producing a well–defined taxonomy of mobile robotic networks will require careful consideration. Instead we now attempt to define what is meant by a swarm robotic network and show how this differs from the above related technologies.

The term 'swarm robots' generally refers to a large collection of mobile robots working on a single task [16], and the development of this technology is growing very rapidly [6]. Some of the reasons for this can be found in the perceived benefits in the characteristic properties of problem solving by social insects. The properties exhibited by social insects result in flexible, robust, decentralised and self-organised systems [6] and it is the desire to imitate these natural systems that is influencing research in swarm robotics.

Motivation

Social insects are regarded as highly effective and, some would argue, the most successful life-form on the planet. There are many reasons why they are so successful, and the properties which make them so are highly desirable in a swarm of robots. Some of these desirable properties are:

1) Redundancy, Reliability, and Scalability: Each entity within a swarm is highly redundant. This redundancy means that the loss of individual entities has little impact on the success of the task at hand, unless all or the vast majority of them are lost. As early as 1989, Rodney Brooks at MIT proposed to NASA that teams of hundreds of inexpensive ant-like reactive robots be sent to Mars in an article entitled ‘Fast, Cheap and Out of Control’. The rationale for this was, in part, to provide a degree of fault-tolerance. Having a large number robots meant that any robots damaged in transit or during landing would not have a real impact on the overall mission [17].
2) Decentralised Coordination: Coordination is completely distributed, and the task in hand will be carried out regardless of whether one or more of the individuals is lost - there is no central point of control in the swarm.
3) Multiplicity of Sensing: In a swarm, many individuals sense the same data. This means that the signal-noise ratio is greatly increased.
4) Dynamic Adaptability to the Working Environment: A swarm will adapt itself to the environment to meet the needs of the swarm.

In addition to imitating the above characteristics of social insects, there may also be practical reasons that a swarm of robots working together may be desirable. For example:

- Some tasks may be too difficult for a single robot, and may require robots working in a team to complete them.
- Using several robots may increase the speed of performing tasks.
- Designing, building, and using several simple robots may be easier, cheaper and more fault-tolerant than using a single robot.
- Theories of self-organisation show that, sometimes, the collective behaviour of a swarm results in patterns which are qualitatively different from those that could be obtained by a single entity. Randomness or fluctuations in individual behaviour, far from being harmful, may in fact greatly enhance the system’s ability to explore new behaviours and find new solutions [6].

The foregoing has described some of the benefits that may be expected in systems that imitate swarms. However, as well as the advantages outlined above, there are also some potential challenges to be overcome. For example:

- Lack of global knowledge may mean that a swarm of robots does not have the information required to perform a task, and stagnates, unable to make any progress.
- Also, there is a ‘too many cooks spoil the broth’ effect. Having more robots working on a task or in a team increases the possibility that individual robots will unintentionally interfere with each other, lowering the overall productivity [16].
- Programming: The concept of a swarm robotic network is that individual entities are autonomous. Nevertheless their deployment implies that there is a task that they are required to do. Often the solution to the task is *emergent*, and it may be extremely complicated to program the robots to perform the task [6]. Swarm engineering is a new discipline proposed by Winfield [18] which aims to help solve this problem.
- Control and Mediation: Complex systems with swarm intelligence might be very difficult to control or mediate if they started to exhibit undesirable behaviour. Such systems would therefore need to be designed and validated for a high level of assurance that they exhibit intended behaviours, and equally importantly do not exhibit unintended behaviours [18].

Definition of Swarm Robotics

The preceding has identified the desirable properties of naturally occuring swarms thus motivating research into artificial swarms. The term ‘swarm’ as applied to robotics was coined by Gerardo Beni and Jing Wang in 1988 at a NATO robotics workshop in Italy [19]. At the time, the discussion was about ‘cellular robots’. This term ‘cellular robots’ was applied to a group of robots that could work like cells of an organism to assemble more complex parts. Beni was discussing a class of cellular robots that could behave in an unpredictable way and move and interact dynamically. For this application it was generally agreed that the adjective ‘cellular’ was not particularly descriptive, and that ‘swarm’ was much better. Moreover, it more accurately portrayed the characteristics of the robots that were under discussion, which were seen to behave in a similar way to swarms occurring in nature..

When Beni and Wang introduced the term, the concept of swarm robotics was largely theoretical, but now it is a fast-evolving reality, with many research projects taking place worldwide – examples being the EU ‘Guardians’ project [20], ‘Ultraswarm’ at the University of Essex [21], Maxelbot at the University of Wyoming [22], Idaho National Laboratory projects [23] and SYMBRION [24] which is a project funded by the 7th Framework programme of the European Union.

At the 2004 Swarm Robotics workshop, Erol Şahin has proposed the following definition for swarm robotics, along with a set of distinguishing criteria to differentiate this technology from other multi-robot research:

> “Swarm robotics is the study of how large numbers of relatively simple physically embodied agents can be designed such that a desired collective behaviour emerges from the local interactions among agents and between the agents and the environment.” [25]

Based on Şahin’s definition a number of criteria may be described that identify a robotic swarm. These are:

1) Autonomous Robots: They should have a physical embodiment in the world, be situated and should be able to physically interact with the world.

2) Large Number of Robots: There should be a large numbers of robots (or the studies should be applicable to the control of large robotic swarms)
3) Few Homogeneous Groups of Robots: There should be relatively few groups containing large numbers of homogeneous robots.
4) Relatively Incapable or Inefficient: The robots should be relatively simple and incapable such that the tasks tackled require the co-operation of the individual robots.
5) Robots with Local Sensing and Communication Capabilities: The robots should only have localised and limited sensing and communication abilities. This constraint ensures that the coordination between the robots is distributed. However, it is acceptable to use global communication channels for a purpose such as to download a common program onto the swarm, but not for coordination among the robots.

This allows us to compare swarm robotic networks with the more established technologies described at the beginning of this section. Table I attempts to summarise and compare the characteristics of swarm robotic networks defined above with the aforementioned more mature technologies. Where there is no entry there is no clear answer to whether the technology fits the criteria or not Table II does the same for some of the implicit characteristics not explicitly included in Şahin's definition.

III. Use and Mis-use of Swarm Robotics

Section II described the scope of the systems we wish to study and showed how swarm robotic systems differ from similar, more mature technologies. Within this scope we may now begin to look at the threats that may be unique to swarm robotic systems, and perhaps identify unique features of swarm robotic systems that may help mitigate these threats.

In analysing the threats to swarm robotic systems it is useful to have an idea, first of all, of how the systems may be used; and then how the systems may be mis-used for malicious ends. In order to describe how systems may be mis-used some security terminology is introduced.

A. Basic Security Terminology

The International Organization for Standardization (ISO) has provided definitions for a number of high level security concepts and we will follow the nomenclature of ISO 13335-1 [26] in our discussion.

ISO 133335-1 defines a *threat* as 'a potential cause of an incident that may result in harm to a system or organization.' In our case, this can be interpreted as any potential incident that may adversely affect the intended objective of the swarm robotic network. The threat may be a threat to the swarm itself or to the information being processed by the swarm. Moreover, the threat can be the result of deliberate or accidental actions. The standard provides a number of examples of threats such as eavesdropping; information modification; malicious code; and physical accidents. In the following we will expand on these examples to illustrate the threats pertinent to swarm robotic networks.

Threats that are not mitigated leave *vulnerabilities* in the system. These threats may be then exploited, causing harm to the system. In other words, although all threats define a potential cause of harm, it is only the unmitigated threats that leave vulnerabilities. In the remainder of this section we consider the threats, rather than specific vulnerabilities.

We refer to the deliberate exploitation of a threat as an *attack* and those that initiate their execution as *attackers* or *adversaries*.

ISO 13335 also defines the notion of the *impact* of the exploitation of a threat, and the *risk* of a threat being exploited. While these are considerations that should be made in the deployment of any system, our objective is not focused on the specific details of a particular application and thus we will focus on the threats.

An example of a threat could be that an unauthorised person might see top secret information; a vulnerability could be that trust is misplaced in a courier delivering this information in a document. The courier may accidentally lose the document; or an attack could be that someone steals the document in transit and publishes it. The impact of a information loss will depend on the content of the document.

Security in any environment, including swarm robotics, is fundamentally about the provision of core *security services.* These services can be defined at a high level without binding the service provision to a technology specific *security mechanism*. The ISO standard for the security architecture of the OSI reference model [27] identifies a number of security services, of which the following are relevant to swarm robotic networks.

Confidentiality
: The confidentiality service protects data from unauthorised disclosure. It may protect all data in a message or selective fields. It may also be used to prevent traffic analysis.

Integrity
: An integrity service prevents prevents data from being altered in an unauthorised or unintended way; for example, by modification, insertion or deletion. As with confidentiality, it may be selective or apply to the entire message. An integrity service may also be used to detect data that has been replayed.

Authentication
: Authentication services may be classed as *peer entity authentication* services or *data origin authentication* services. The former provides assurance that the peer entity in the communication protocol is who they claim to be. The latter provides assurance that data came from its reputed source.

Availability
: Although, strictly speaking, not a security *service*, availability is defined in ISO 7498-2 as the *property*

Table I: Comparison of Explicit Characteristics

	Swarms	Multi-Robot	Mobile Sensor Networks	MANET	Multi Agent Systems
Autonomous	✓	✗			
Large Number	✓		✓		
Few Homogeneous Groups	✓		✓	✗	✓
Simple	✓		✓		✓
Local Sensing and Comms.	✓	✓	✓	✓	✓

Table II: Comparison of Implicit Characteristics

	Swarms	Multi-Robot	Mobile Sensor Networks	MANET	Multi Agent Systems
Self Organising	✓	✗	✗		
Emergent Behaviour	✓	✗	✗	✗	
Co-operate to accomplish task	✓	✓		✗	✓
Distributed Command/Control	✓	✗	✗		✓
Mobile	✓	✓	✓	✓	✓
No ID Necessary	✓		✗	✓	

of being accessible and useable upon demand by an authorised entity. The term *denial of service* is often used in reference to loss of availability.

ISO 7498-2 defines two more services: access control, and non–repudiation. Although these services are important in the context of the OSI reference model, they are of less relevance to swarm robotics. Entities in a robotic swarm are typically simple devices that do not provide access to a service, and which operate in a closed network where disclaiming a previous transaction is not a high risk.

Security mechanisms used to provide the above services include *encryption*, for confidentiality, and *digital signatures* and *message authentication codes* for integrity and data origin authentication. Entity authentication usually requires the completion of a security protocol. The Handbook of Applied Cryptography [28] provides a good introduction to these mechanisms. Mitigating denial of service attacks is more dependent on the particular application. In any system, the provision of security is a holistic process. This requires careful management processes that oversee the use of specific security technologies that can be applied to devices and networks. These include *firewalls, access control mechanisms* and *network security protocols*. At the heart of most security technologies is the deployment of specific *cryptographic primitives*, which are mathematical tools that can be applied to data to provide the core security services. These normally rely on the careful protection and maintenance of *cryptographic keys*, which are critical data items that must be stored securely.

With this background we may now review several scenarios where swarm robotic technology is being considered for use, and look at the potential threats to these systems.

B. Military Applications

Swarm robotic networks are of particular interest to the military, and in these applications the need for security is perhaps self-evident. There is currently a great deal of research taking place in the military use of robotic swarms. In the United Kingdom in August 2008, a challenge called ‘The Grand Challenge’ [29] took place, which was searching for the best ideas in defence technology to help solve some of the evolving threats facing front line troops. One prominent entrant to this was ‘Swarm Systems’ [30], which used swarms of micro air vehicles.

In the United States, US Army Research are funding and working with BAE Systems on the The Micro Autonomous Systems and Technology (MAST) project, [31] which will ‘research and develop advanced robotic equipment for use in urban environments and complex terrain, such as mountains and caves. The project will create an autonomous, multifunctional collection of miniature intelligence-gathering robots that can operate in places too inaccessible or dangerous for humans’.

Mine clearance is another example of where robotic swarms may be deployed. Individual entities that constitute a swarm robotic system are dispensable, making the system suitable for domains that involve dangerous tasks. For instance, clearing a corridor on a mining field. Swarm systems would be better than a single more complex and expensive mine clearing robot because they can afford to be suicidal, and may be able to cover the area more quickly.

The major threat to military systems is from deliberate attacks on the robotic swarm. Such attacks may range from passive eavesdropping on communications, or monitoring traffic; to more sophisticated attacks where malicious robots may be injected into the swarm, much as viruses and Trojans are deployed in computer systems. Such sophisticated attacks may go un-noticed while the attacker manipulates data being processed by the swarm and possibly affects the emergent behaviour.

C. Monitoring

Robotic swarms are well–suited to environmental monitoring and, since they have motor as well as sensor capabilities, they could potentially provide solutions in the case of undesired environmental events. For example, the Elimination Units for Marine Oil Pollution (EU-MOP) project demonstrated that a robotic swarm could be used to detect environmental pollutants such as oil spillages, and subsequently clean them up [32].

At the Ecole Polytechnique Federale de Lausanne (EPFL) an investigation has been completed into how swarm robotics could be used for the autonomous inspection of complex engineered structures [33].

Accidental malfunction of the entities that make up the swarm is always a threat. The impact of this threat could be significant these applications if the swarm was the sole means of monitoring.

In addition to malfunctions or accidents, active threats to such robot swarms could arise from malicious organisations such as terrorists or criminals. Such groups could target the availability of the swarms, or the confidentiality or integrity of any information that they hold. For example, by injecting malicious code or malicious robots, they could physically or electronically hijack the system resulting in the loss of availability. As with current denial of service attacks on internet–based services, the threat of such an attack is itself sufficient to meet the adversary's requirements. As with internet–based services a malicious organisation could potentially use such a threat to extort money from the legitimate owners of the swarm. Many such attacks have already been launched against business websites on the Internet. The threat may even be exploited, with the robots only being returned to use after a ransom has been paid.

The data that a monitoring swarm holds could also be useful to an unauthorised third party. For example the location and extent of an oil spillage could be of interest to an environmental group; the location of faults in an engineering structure could be of great value to a competitor. Therefore, it is of importance that such data is kept confidential. Also, if such data could be corrupted accidentally or deliberately, it could lead to the swarm performing incorrectly, which could mean that monitoring is not taking place properly or the swarm is not trying to fix something that it is meant to be. Thus integrity protection would be a useful service to have in these applications.

D. Disaster Relief

The deployment of robot swarms during disaster relief operations is another application area that is considered for swarm robotic networks.

At the University of Utah, research has taken place into using swarms of robots to aid first responders in disaster situations [34] and the European Union 6th Framework GUARDIANS project [20] is addressing a similar application.

The GUARDIANS are a swarm of autonomous robots applied to navigate and search an urban space in situations which are dangerous and time-consuming for humans. The project's central example is an industrial warehouse in smoke, as proposed by the Fire and Rescue Service. The job is time consuming and dangerous since toxins may be released and humans senses can be severely impaired. The robots warn of toxic chemicals, provide and maintain mobile communication links, infer localisation information and assist in searching. They enhance operational safety and speed and thus indirectly save lives

In situations such as these, availability becomes a primary security requirement, as well as confidentiality, integrity and authentication/identification. Availability is necessary so that the swarm can respond as quickly as possible to the emergency at hand. If robots are unavailable due to malfunction, accident or because they have been hijacked either physically or electronically by an external agency, then they will be unable to perform their critical task. The motivation for such an attack may be difficult to comprehend, however, as discussed above, the threat of a denial of service attack may be sufficient for malicious groups to extort their demands.

Unauthorised access to data could also be a threat in this application. Eavesdropping on the robots communications may provide information to an attacker, for example about the location and extent of the damage and about any entities that the robots discover during the rescue operation. Such information may be highly sensitive and would require the protection of a confidentiality service.

Perhaps more importantly, there exists a threat of data manipulation. Integrity protection is necessary to ensure that the data being passed around the swarm is accurate, so that the robots respond correctly. In addition to integrity protection, data origin authentication may be required to provide assurance that information has come from a reliable source.

Threats arising from entity authentication failure are more subtle. It may be necessary to ensure that sensitive information obtained by the swarm is communicated only to legitimate parties e.g. the rescue service. If robots are communicating locally, problems could arise if the peer entity cannot authenticate itself. This scenario could arise if multiple swarms, perhaps with different goals, are operating in the same physical area.

Many relevant current technologies already provide full support for strong security. Communications between human personnel in emergency situations often use Terrestrial Trunked Radio (TETRA) [35] which is an open digital standard defined by the European Telecommunications Standard Institute (ETSI). Whether this technology is applicable to swarms, however, will depend on the particular implementation.

E. Healthcare

The use of swarm robotics has been considered for a wide range of healthcare services: from surgery and intrabody diagnostics, to more routine tasks such as medication provision and patient monitoring. The European IWard project is proposing to use swarms of robots to provide assistance to healthcare workers [36].

Entity authentication will be very important for swarm robotics in healthcare situations. For example, it will be of vital importance that only legitimate robots are introduced into a human body, or sent to deliver patient medicines or read patient data from monitoring stations. Failure to authenticate could result in the introduction of swarm robots that would harm the patient surgically or whilst inside their body, by delivering incorrect medicines or by reporting medical data to unauthorised entities.

The confidentiality or privacy of patient data is paramount, and is protected by law in many countries. Apart from patients wanting to be able to choose who knows their personal medical history, it must be kept from organisations who may wish to have it for reasons such as pharmaceutical research, or to simply try and deny an individual access to insurance, employment or services.

Integrity of medical information must be ensured. Otherwise a swarm could damage a patient by responding to incorrect information such as wrong organ position, elevated blood pressure or blood sugar levels. Consequently, if a swarm were to respond incorrectly, this could seriously damage the patient's health, maybe fatally.

Availability of swarm robots in a healthcare situation is important, especially so where they are deployed in situations with critically ill patients. If they are not available and able to respond immediately then such patients could suffer greatly, and maybe die as a result.

F. Commercial Applications

As the technology develops, the hope is that robotic swarms will find commercial use. Commercial uses could include some of applications already discussed, for example monitoring or healthcare, as well as many other routine tasks that are 'dull, dirty, or dangerous' [16]. Unfortunately in any commercial environment, the motivation to gain competitive advantage will undoubtedly result in attempts to steal information, manipulate data, and disrupt services.

For example, if an organisation can interrupt their competitors service and make it unavailable or unstable, damaging its reputation, then they will become the organisation of choice. If they can steal information from their competitor then they may be able to find out their trade secrets for their own commercial gain. If they can amend their competitors data then they can make them operate unpredictably, again damaging their reputation, and making themselves appear preferable.

Thus consideration of confidentiality, integrity, authentication and availability will be required for commercial applications to be successfully adopted and deployed.

IV. Security Challenges in Swarm Robotic Environments

Section II described what we mean by swarm robotics and discussed how this differs from related technologies. In the preceding section we have described the threats that may arise in the deployment of swarm robotic systems and identified the security services that may be applied to mitigate these threats.

It is appropriate now to consider the challenges to providing these security services in swarm robotic networks. It is clear that some security problems are similar to those experienced by other related technologies, and that some solutions from these technologies may apply to swarm robotics. However, not all of these shared problems have been fully solved. Furthermore, the swarm robotic environment introduces particular security challenges that do not exist in other technologies.

A. Resource Constraints

According to our definition of swarm robotics, the robots should be relatively simple. The less complicated a device is, the greater the challenge in providing security becomes. This is due to resource constraints: storage for static and ephemeral data is restricted; communication bandwidth is constrained; and processing power is limited. Most importantly, where mobile devices are concerned, power consumption has to be minimised to preserve energy.

Resource constraints restrict the types of existing security technologies that can be deployed and special cryptographic mechanisms may be required to reduce the consumption of resources on such devices [37]–[39]. However, attacks on the provision of resources can still lead to the device becoming inoperable – permanently so if the resource is not renewable e.g. a battery. This would result in loss of availability of the device and potentially the swarm.

B. Physical Capture and Tampering

Robotic swarms are unique in their combination of physical entities with autonomous behaviour, mobility, and distributed control. Consequently, the owner of a swarm may not know the exact location of each device and what other entities may be in the vicinity. Thus individual swarm robots may be captured by an attacker.

Physical capture of a robot may lead to immediate loss of availability. The attacker may also use the device to manipulate data being reported, and may attack the device hardware to extract any secret data.

In the worst–case scenario an attacker could modify the device and re-introduce it to the swarm, enabling a number of other attacks to be carried out. Such a rogue device may continue to manipulate data as the swarm moves to new locations. It may eavesdrop on communications. It may even be able to introduce malicious code or commands to other devices. In the worst case it would be able to alter the behaviour of the swarm without the attack being detected. This capture and 're-introduction attack' is unique to swarm robotic technology.

C. Monitoring and Control

Systems employing swarm intelligence do not have a hierarchical structure with specific points of monitoring and control. Moreover, the individual entities within these systems take decisions autonomously, based on local sensing and communications. With such systems it is evident that there could be many risks if, for any reason, deliberate or accidental, they went 'out-of-control'. These risks include many security violations such as loss of confidentiality, integrity or availability. Monitoring and control presents an interesting challenge to security within swarm robotics.

D. Communication

Unlike the related technology discussed in Section II, robotic swarms may be designed to interact either explicitly, or implicitly [40].

Explicit communication can be achieved via broadcast or directed messages. Radio frequency (RF) and infra-red (IR) technologies have been widely used for explicit communications within swarms. Other technologies include coloured LED display, body-language or sign-language, colour patterns on a robot's body, coil induction, haptics, audible sounding, combination of LED display and audio signalling and acoustic signalling in an underwater environment.

Implicit communication is unique, amongst the technologies discussed in Section II, to swarm robotics. By considering implicit communication, we include interaction via sensing other robots and their behaviours, and interaction via the environment. The latter acts as a sort of shared memory and is known as *stigmergy* [6], [41], [42].

From a security perspective, any open implicit or explicit communication method can be jammed, intercepted or otherwise disturbed relatively easily by an attacker. The security of RF and IR has been well–researched, but the security of the remaining more 'exotic' interaction methods needs to be thoroughly investigated and presents a fascinating security challenge.

E. Swarm Mobility

Security is difficult to provide in any mobile environment, however the mobility of robot swarms, combined with the autonomous behaviour, is quite unusual. This has some interesting characteristics that might make some security services easier to implement than for related technologies.

One example is entity authentication discussed below. Swarm robots may be able to move towards the peer that they wish to authenticate. The authentication service could then be provided through visual sensing and physical data exchange. A similar example is key distribution: robots may move within the swarm to distribute shared keys.

However any constraint on the movement of swarm members, for example to remain in the 'bounds' of the swarm, could present additional security issues.

F. Entity Authentication and Identity

As discussed in Section III, in some applications it may be very important for a swarm robot to determine whether it is interacting with a legitimate entity or not. Data origin and entity authentication often require some notion of *identity*. This is a particular problem where *individual identity* within a swarm is undesirable [43]. However, other work on robotic swarms has used *group identity* [44], or individual identity which is broadcast regularly [45].

If identity can be assumed or changed, then attacks can be launched on entity authentication, confidentiality, integrity and availability. The notion of identity within a robotic swarm thus presents an interesting challenge from a security standpoint.

G. Key Management

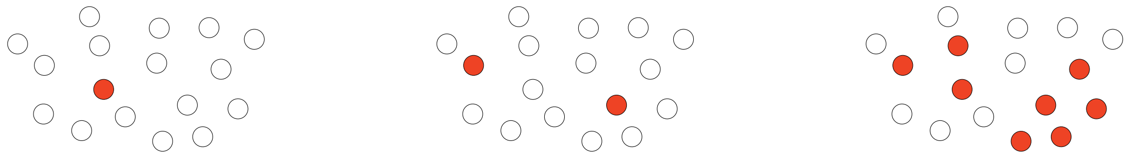
Security services deployed in a robot swarm inevitably require the need to manage cryptographic keys [46]. These keys define which pairs (or groups) of robots can apply security services. As robots join and leave a swarm, it may be necessary to update this keying material. Thus the dynamic and interactive nature of a swarm presents sophisticated key management challenges although the intelligent mobility of the swarm may provide some novel solutions to this.

H. Intrusion Detection

When an unauthorised entity joins a network it is sometimes called *intrusion,* and the field of network intrusion detection is well–established [47]–[49]. However, the physical nature of the entities in a swarm robotic network means that intrusion in a robotic swarm robotic network is not the same as intrusion in a traditional data network. Deliberate intrusion was alluded to in Section IV-B and accidental intrusion in Section III-D, where several swarms may operate in the same geographical location. In these cases the intrusion mechanism is the physical insertion of a rogue agent into the swarm, which is not addressed by the established network intrusion detection systems.

Intrusion detection systems typically attempt to detect anomalous behaviour in the network. In a robotic swarm this behaviour may extend to the physical behaviour of individual robots. New mechanisms will be required to detect anomalous physical behaviour. Moreover, the autonomous nature of robots and collective emergent nature of the behaviour of the swarm will make any anomalous behaviour difficult to detect.

If undetected, one or more foreign robots could infiltrate the swarm, either maliciously or accidentally, and ultimately affect the desired emergent behaviour. The situation is illustrated in figure 1 where the shaded nodes represent foreign robots infiltrating the swarm. In figure 1a a single intruder has entered the swarm. It is not unreasonable to expect the swarm to be able to detect this intruder. If several foreign robots can infiltrate the swarm, as shown in figure 1b, it may be easier for the intruder to affect the behaviour of the swarm. As



(a) Single Intruder

(b) Multiple Intruders

(c) Overlapping Swarms

Figure 1: Intrusion of rogue robots into a swarm

more foreign robots infiltrate the swarm, 1c, it will become more difficult to distinguish the intruders from the original swarm, particularly if the notion of identity within a swarm is forbidden.

Once an intruder is detected, an appropriate response will need to be formulated according to an *Intrusion Protection System* [49]. Depending on the application the response could be to simply ignore the rogue device, to monitor its behaviour, or to find a way to either disable it or remove it from the system. In some scenarios it may even be desirable to manipulate the behaviour of the intruder in a counter-attack.

Intrusion detection and protection looks to be particularly challenging in a swarm of robots, and will need a specifically tailored approach.

I. Managing Learning

Robotic swarms are designed to learn and react to environmental changes by means of adaption. A malicious entity might present changes to the *environment* which will cause a swarm to adapt in an undesired way. For example, if anomaly detection is used to detect intrusion based on learning and monitoring typical behaviour, then a malicious entity could manipulate the environment to change the pattern of 'typical' behaviour in order to gain entry to the network.

V. Conclusion

The development of swarm robotic technology has reached a point where many new applications beginning to emerge. Therefore, we believe that this is an opportune moment to take a closer look at the security of swarm robotic systems - before widespread deployment.

To that end we have identified several unique features of swarm robotic networks that distinguish them from related technology and consequently justify further study from a security perspective. Most notable of these characteristics are the autonomous behaviour of the swarm and the emergent behaviour. Although much has already been accomplished to provide security for related technologies, the characteristics of autonomy and emergent behaviour, combined with mobility and distributed control, make robotic swarms significantly different from these technologies to raise a number of new security problems. These new security problems are not only of theoretical interest but will have implications on many practical applications of swarm robotic technology.

Bearing this in mind, we described a number of challenges to robotic swarm security, many of which are unique to this technology. For example, the application of stigmergic communications may provide a new attack surface that will require the development of new security mechanisms. The notion of identity within a swarm may also necessitate research into the provision of entity authentication within a swarm. And finally the potential to modification of emergent behaviour if a malicious entity manages to infiltrate the swarm may require further investigation into intrusion detection, especially where the intruder is a physical mobile agent. We therefore believe that an investigation of the above areas is timely.

References

[1] E. Şahin and W. M. Spears, Eds., *Swarm Robotics: SAB 2004 International Workshop*, ser. LNCS. Santa Monica, CA, USA: Springer Berlin / Heidelberg, Jul 2005, vol. 3342.

[2] E. Şahin, W. Spears, and A. Winfield, Eds., *Swarm Robotics: Second International Workshop, SAB 2006*, ser. LNCS. Rome, Italy: Springer Berlin / Heidelberg, Sep 2007, vol. 4433.

[3] L. Bayindir and E. Şahin, "A review of studies in swarm robotics," *Turkish Journal of Electrical Engineering*, vol. 15, pp. 115–147, 2007.

[4] A. F. T. Winfield and J. Nembrini, "Safety in numbers: fault–tolerance in robot swarms," *International Journal of Modelling, Identification and Control*, vol. 1, no. 1, pp. 30–37, 2006. [Online]. Available: http://inderscience.metapress.com/link.asp?id=byu8g6fqm5g55v9x

[5] F. Higgins, A. Tomlinson, and K. Martin, "Survey on security challenges for swarm robotics," in *Fifth International Conference on Autonomic and Autonomous Systems (ICAS 2009)*, R. Calinescu, F. Liberal, M. Marin, L. P. Herrero, C. Turro, and M. Popescu, Eds. Valencia: IEEE Computer Society, April 2009, pp. 307– 312.

[6] E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm intelligence: from natural to artificial systems*. Oxford University Press US, 1999.

[7] L. E. Parker, "Alliance: an architecture for fault tolerant multirobot cooperation," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 2, pp. 220–240, Apr 1998.

[8] T. Wark, C. Crossman, W. Hu, Y. Guo, P. Valencia, P. Sikka, P. Corke, C. Lee, J. Henshall, K. Prayaga, J. O'Grady, M. Reed, and A. Fisher, "The design and evaluation of a mobile sensor/actuator network for autonomous animal control," in *IPSN '07: Proceedings of the 6th international conference on Information processing in sensor networks*. New York, NY, USA: ACM, 2007, pp. 206–215.

[9] K. Dantu, M. Rahimi, H. Shah, S. Babel, A. Dhariwal, and G. Sukhatme, "Robomote: Enabling mobility in sensor networks," in *Proceedings of Fourth International Symposium on Information Processing in Sensor Networks*, 2005, pp. 404–409.

[10] (2008) Robotic sensor networks. Minnesota University. [Online]. Available: http://rsn.cs.umn.edu

[11] J. Reich and E. Sklar, "Toward automatic reconfiguration of robot-sensor networks for urban search and rescue," in *Proceedings of the 1st International Workshop on Agent Technology for Disaster Management*, 2006, pp. 18–23. [Online]. Available: http://users.ecs.soton.ac.uk/sdr/atdm/ws34atdm.pdf

[12] L. Buttyán and J.-P. Hubaux, *Security and cooperation in wireless networks: thwarting malicious and selfish behavior in the age of ubiquitous computing*. Cambridge University Press, 2007.

[13] E. Hansson, A. Bengtsson, and A. Vidström, "Security solutions for mobile ad hoc networks," Swedish MOD, FOI Defence Research Agency Command and Control Systems, Tech. Rep. FOIR-1694-SE ISSN 1650-1942, Aug 2005.

[14] M. Wooldridge, *An Introduction to MultiAgent Systems*. Wiley, 2002.

[15] N. Borselius, "Multi-agent system security for mobile communication," Ph.D. dissertation, Department of Mathematics, Royal Holloway, University of London, 2003.

[16] R. R. Murphy, *Introduction to AI Robotics*. MIT Press, 2000.

[17] R. A. Brooks and A. Flynn, "Fast, cheap and out of control - a robot invasion of the solar system," *Journal of the British Interplanetary Society*, vol. 42, pp. 478–485, 1989.

[18] A. F. T. Winfield, C. Harper, and J. Nembrini, "Towards dependable swarms and a new discipline of swarm engineering," in *Swarm Robotics: SAB 2004 International Workshop*, ser. LNCS, E. Şahin and W. M. Spears, Eds., vol. 3342. Santa Monica, CA, USA: Springer Berlin / Heidelberg, Jul 2005, pp. 126–142.

[19] Beni and Wang, "Swarm intelligence in cellular robotic systems," in *NATO Advanced Workshop on Robotics and Biological Systems*, Il Ciocco, Tuscany, Italy, 1989.

[20] EU. (2007) 6th framework programme: Guardians project. European Union. [Online]. Available: http://www.guardians-project.eu/

[21] R. D. Nardi and O. Holland, "Ultraswarm: A further step towards a flock of miniature helicopters." in *Swarm Robotics: Second International Workshop, SAB 2006*, ser. LNCS, E. Şahin, W. Spears, and A. Winfield, Eds., vol. 4433. Rome, Italy: Springer Berlin / Heidelberg, 2006. [Online]. Available: http://gridswarms.essex.ac.uk/publications/DeNardi2006UltraswarmFurther.pdf

[22] W. M. Spears and D. F. Spears. Maxelbot project. [Online]. Available: http://www.cs.uwyo.edu/~wspears/ap.html

[23] Idaho National Laboratory. [Online]. Available: http://www.inl.gov/adaptiverobotics/robotswarm/index.shtml

[24] [Online]. Available: www.symbrion.eu

[25] E. Şahin, "Swarm robotics: From sources of inspiration to domains of application," in *Swarm Robotics: SAB 2004 International Workshop*, ser. LNCS, E. Şahin and W. M. Spears, Eds., vol. 3342. Santa Monica, CA, USA: Springer Berlin / Heidelberg, Jul 2005.

[26] *Information technology – Security techniques – Management of information and communications technology security – Part 1: Concepts and models for information and communications technology security management*, International Organization for Standardization (ISO) Std. 13 335, Rev. 1, 2004.

[27] *Information processing systems – Open Systems Interconnection – Basic Reference Model – Part 2: Security Architecture*, International Organization for Standardization (ISO) Std. 7498-2, Rev. 1, 1989.

[28] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *The Handbook of Applied Cryptography*, 5th ed. CRC Press, 1996. [Online]. Available: http://www.cacr.math.uwaterloo.ca/hac/

[29] (2008) The grand challenge. United kingdom ministry of defence. [Online]. Available: http://www.challenge.mod.uk

[30] (2008, Aug) Swarm systems: providing swarms of micro air vehicles. [Online]. Available: http://www.swarmsys.com

[31] BAE-Systems. (2008, April) Mast project. [Online]. Available: http://www.baesystems.com

[32] K. Gkonis, T. Pavlidis, N. Kakalis, and N. Ventikos, "Final project report for the elimination units for marine oil pollution (eu-mop) project," European Union, 6th Framework Programme, Tech. Rep., 2008.

[33] N. Correll and A. Martinoli, "A challenging application in swarm robotics: The autonomous inspection of complex engineered structures," *Bulletin of the Swiss Society for Automatic Control*, vol. 46, pp. 15–19, 2007.

[34] D. P. Stormont, "Autonomous rescue robot swarms for first responders," in *IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety (CIHSPS 2005)*, Orlando, FL, USA, Apr 2005, pp. 151–157.

[35] (2008) Terrestrial trunked radio. Tetra-Association. [Online]. Available: http://www.tetra-association.com

[36] EU. (2007) 6th framework programme: Iward: Intelligent robot swarm for attendance, recognition, cleaning and delivery. European Union. [Online]. Available: http://www.iward.eu/

[37] S. Kumar and C. Paar, "Reconfigurable instruction set extension for enabling ecc on an 8-bit processor," in *International Conference on Field-Programmable Logic and Applications (FPL) 2004,*, Antwerp, Belgium, 2004.

[38] L. Batina, J. Guajardo, T. Kerins, N. Mentens, P. Tuyls, and I. Verbauwhede, "Public-key cryptography for rfid-tags," in *IEEE International Workshop on Pervasive Computing and Communication Security*, New York, USA, 2007.

[39] A. Bogdanov, L. Knudsen, G. Leander, C. Paar, A. Poschmann, M. Robshaw, Y. Seurin, and C. Vikkelsoe, "Present: An ultra-lightweight block cipher," in *Cryptographic Hardware and Embedded Systems - CHES 2007*, 2007, pp. 450–466.

[40] L. E. Parker, "Current state of the art in distributed robot systems," in *Distributed Autonomous Robotic Systems 4*, L. E. Parker, G. Bekey, and J. Barhen, Eds. Springer, 2002, pp. 3–12. [Online]. Available: http://www.cs.utk.edu/~parker/publications/DARS_2000_overview.pdf

[41] P.-P. Grassé, "La reconstruction du nid et les coordinations inter-individuelles chez bellicositermes natalensis et cubitermes sp. la théorie de la stigmergie: Essai d'interprétation du comportement des termites constructeurs," *Insectes Sociaux*, vol. 6, no. 1, pp. 41–80, 1959.

[42] T. White, "Expert assessment of stigmergy: A report for the department of national defence," School of Computer Science, Carleton University, Ottawa, Ontario, Canada, Tech. Rep., 2005. [Online]. Available: http://www.scs.carleton.ca/~arpwhite/stigmergy-report.pdf

[43] P. Flocchini, G. Prencipe, N. Santoro, and P. Widmayer, "Gathering of asynchronous robots with limited visibility," *Theoretical Computer Science*, vol. 337, no. 1-3, pp. 147–168, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/B6V1G-4FC37VR-1/2/119f08be923e94ddbdb49819df053e63

[44] R. A. Russell, "Visual recognition of conspecifics by swarm robots," in *Australasian Conference on Robotics and Automation*, 2004. [Online]. Available: http://www.araa.asn.au/acra/acra2004/papers/russell.pdf

[45] J. Fredslund and M. Matarić, "A general algorithm for robot formations using local sensing and minimal communication," *IEEE Transactions on Robotics and Automation*, vol. 18, pp. 837–846, 2002.

[46] S. Dolev, L. Lahiani, and M. Yung, "Secret swarm unit reactive k – secret sharing," in *Progress in Cryptology – INDOCRYPT 2007*, ser. LNCS, vol. 4859. Chennai, India: Springer Berlin / Heidelberg, Dec 2007, pp. 123–137.

[47] J. P. Anderson, "Computer security threat monitoring and surveillance," James P. Anderson Co., Tech. Rep., 1980.

[48] D. E. Denning, "An intrusion detection model," in *Proceedings of the Seventh IEEE Symposium on Security and Privacy*, May, 1986, pp. 119–131.

[49] K. Scarfone and P. Mell, "Guide to intrusion detection and prevention systems (idps)," National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, Tech. Rep. SP 800–94, Feb 2007. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-94/SP800-94.pdf

www.iariajournals.org

International Journal On Advances in Intelligent Systems
➪ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS
➪issn: 1942-2679

International Journal On Advances in Internet Technology
➪ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING
➪issn: 1942-2652

International Journal On Advances in Life Sciences
➪eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO
➪issn: 1942-2660

International Journal On Advances in Networks and Services
➪ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION
➪issn: 1942-2644

International Journal On Advances in Security
➪ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS
➪issn: 1942-2636

International Journal On Advances in Software
➪ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS
➪issn: 1942-2628

International Journal On Advances in Systems and Measurements
➪ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL
➪issn: 1942-261x

International Journal On Advances in Telecommunications
➪AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA
➪issn: 1942-2601