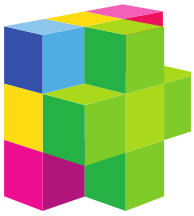
Sustainable
Society Network



WORKING PAPERS OF THE SUSTAINABLE SOCIETY NETWORK+
Vol. 3
February 2015

Conference Proceedings: First International Conference on Cyber Security for Sustainable Society 2015

Coventry University, 26-27 February 2015

ISSN 2052-8604

Introduction to the Working Papers of the SSN+

The *Working Papers of the Sustainable Society Network+* are a publication avenue for ongoing work by the members of the SSN+. Through these working papers, we aim to make academic research more accessible to the general public and other interested parties. Rather than follow strict academic style and referencing, therefore, we instead provide some bibliographies for readers interested in further information.

Disclaimer

Main Website

http://sustainablesocietynetwork.net/

Twitter

@SustainableSoci

Email

sustainable.society@imperial.ac.uk

Digital Economy
Transforming Business and Society

Sustainable
Society Network

ISSN 2052-8604

Table of Contents

ISSN 2052-8604

ISSN 2052-8604

ISSN 2052-8604

Sustainable Society

Dr Siraj Ahmed Shaikh, Conference Co-Chair, Reader in Cyber Security, Department of Computing, Coventry University

Dr Catherine Mulligan, Conference Co-Chair, Principal Investigator, Sustainable Society Network+, Imperial College London

Dr Stephen Lorimer, Network Coordinator, Sustainable Society Network+, Imperial College London

Welcome to our third volume of the Sustainable Society Network+ Working Papers. In this edition you will find the conference proceedings from the 1st International Conference on Cybersecurity for Sustainable Society 2015, covering a broad range of topics from network analysis and psychology to behavior change and trust.

As many of our readers are aware, the SSN+ was set up as one of four Digital Economy Program Networks designed to create communities of practice across the UK and globally. Below we outline why the use of digital technologies is required to help create a sustainable society.

On our website, you will find a number of upcoming events and funding opportunities. Please do visit and connect with us – we look forward to hearing from you and hopefully including your paper in future conferences that the Network sponsors!

Cathy, Siraj and Steve

Digital Technologies and Sustainable Societies

Why is this area important?

Digital technologies can be used to make services more sustainable and enhance current systems (economic, environmental and social), in a way that is accessible, affordable, bespoke and popular. The Digital Economy has potential to transform lifestyles and improve quality of life, having an impact on society as a whole.

Sustainable
Society Network

ISSN 2052-8604

Our ways of interacting and networking in research, business etc. is changing: using digital technologies to thread disconnected systems together, individuals can be well connected and informed on a personal level. By collating and using information to deliver timely and appropriate options, service providers can enable consumer choice and delivery of improved services at decreased cost.

On an increasingly instrumented planet, there is a need for creation of open standards and development of user accessible tools for life. In sustainable societies of the future, people will be able to make informed sustainable choices. Improved information delivery (economic, environmental, social and political) will foster changes in behavior to minimise the negative impact of our activities.

Why is the Digital Economy important to this area?

The Digital Economy (DE) will help provide the technologies and socio-economic understanding to deliver services sustainably.

Research in this area requires a holistic, multi-disciplinary approach to address not only the technical challenges, but also the human aspects (e.g. how to encourage personal motivation to engage and change attitudes and behaviours, how to build trust in information and services).

Information: people will need the right information at the right time in the right format for them. How do we get individuals well connected and informed? How can we ensure delivery of high quality, trusted information with a consistent, personal, bespoke user experience, but without breaking privacy? What are the challenges in creation of open standards and associated legal frameworks? How can we tackle the challenges of creating knowledge, from information, from data?

The DE has the potential to transform how we deliver services and thus have a transformational impact on society, for example:

- **Transport:** Which technologies and what cultural changes are needed to realise a society where there is no need to own a car and there are real alternatives to long-distance air travel?
- **Energy and the Environment:** Digital technologies can facilitate energy demand reduction at a number of levels of interaction between society and the energy

ISSN 2052-8604

system. How can the DE lower the negative environmental impact of societies and inform drastic decisions about scarce resources?

- **Healthcare:** Is the current model sustainable? How can we encourage a shift in focus of the current healthcare system to tackle long-term problems? The DE has a role in addressing the challenges of delivering high quality, personalised healthcare, managing wellness and driving down the cost of healthcare provision.

ISSN 2052-8604

First International Conference on Cybersecurity for Sustainable Society

The following papers were presented at Coventry University on 26-27 February 2015. Each of the papers was reviewed by the co-chairs and a team of reviewers before being included in the conference and these proceedings.

In the call for papers, the co-chairs emphasised that modern society aspires to be economically and socially sustainable. The digital infrastructure underpinning our society allows us to achieve these goals by various mechanisms including online social networks, trust and reputation, peer-to-peer and content sharing, secure transactions and online voting. Such mechanisms however are at increasing risk from cyber attacks, industrial espionage, online fraud and government surveillance; each serving to undermine trust and security at different levels. This conference offered a unique platform to address the issues of cyber security and trust in the context of sustainable societies. Authors were invited to submit original research papers addressing theoretical foundations, modelling and empirical approaches, along with more exploratory work, and position papers to raise open questions.

Topics of interest include but were not limited to:

- Cyber Attacks (Phishing, Reputation Attacks)
- Identity, Trust and Trustworthiness
- Privacy, Surveillance and Control
- Reputation Systems and Social Networks
- Security Governance and Public Good
- Human Factors and Behavioural Dynamics
- Ethics, Philosophy, Politics and Innovation
- Security Economics, Incentives and Liabilities
- Low Carbon and Resilient City and Transport

Program committee co-chairs

- Siraj Ahmed Shaikh (Coventry University, UK)
- Catherine Mulligan (Imperial College, UK)

Sustainable
Society Network

ISSN 2052-8604

Program committee

- Debi Ashenden (Cranfield University, UK)
- Anirban Basu (KDDI Laboratories, Japan)
- Eduardo Cerqueira (UFPA, Brazil)
- Tom Chen (City University, UK)
- Paul Curzon (Queen Mary Univ of London, UK)
- Zeynep Gurguc (Imperial College, UK)
- Chris Hankin (Imperial College, UK)
- Vassilis Kostakos (University of Oulu, Finland)
- Paddy Krishnan (Oracle, Australia)
- Santi Phi (Chiang Mai University, Thailand)
- Juan E. Tapiador (Carlos III de Madrid, Spain)
- Tim Watson (University of Warwick, UK)
- Chris Preist (Bristol University, UK)
- Stefanos Skalistis (EPFL, Switzerland)

Reviewers

- Dale Richards
- Alex Stedmon
- Tom Crick
- Madeline Cheah
- Adrian Venables
- Harsha Kumara Kalutarage

ISSN 2052-8604

Applying social network analysis to security

Elizabeth Phillips[1], Jason Nurse[2], Michael Goldsmith[2] and Sadie Creese[2]
[1]Oxford University Centre for Doctoral Training in Cyber Security
elizabeth.phillips@cybersecurity.ox.ac.uk

[2]Cyber Security Centre, Department of Computer Science,
University of Oxford, Oxford, UK
{*firstname.lastname*}@cs.ox.ac.uk

Abstract

In this paper, we set out to explore some of the many ways in which Social Network Analysis (SNA) can be applied to the field of security. In particular, we investigate what information someone (e.g., an attacker) could infer if they were able to gather data on a person's friend-groups or device communications (e.g., email interactions) and whether this could be used to predict the "hierarchical importance" of the individual. This research could be applied to various social networks to help with criminal investigations by identifying the users with high influence within the criminal gangs on DarkWeb Forums, in order to help identify the ring-leaders of the gangs. For this study we conducted an initial investigation on the Enron email dataset, and investigated the effectiveness of existing SNA metrics in establishing hierarchy from the social network created from the email communications metadata. We then tested the metrics on a fresh dataset to assess the practicality of our results to a new network.

Introduction

The Internet has transformed the way in which people communicate with each other within society. With the increase in communications, comes an added exposure associated with this additional traffic. This paper aims to focus on the specific test case of inferring hierarchy from such communications. The technique that we are specifically interested in is Social Network Analysis (SNA), i.e. a set of approaches that allow for the study of social links between elements (e.g. people, devices or things).

Social networks have been an attractive resource to analyse dating as far back as 1930[1]. Freeman in 1979 highlighted the initial works of Moreno, Jennings, Warner and others in investigating the social networks within schools, prisons and workplaces. However, the Real World Experiment of Travers et al. in 1969[2] was the first to highlight

ISSN 2052-8604

how connected our own social networks are with the “small world phenomenon”. The directed nature of communication allows SNA to be used to help create comprehensive network graphs that can be assessed visually and mathematically (through a range of SNA metrics) to help identify influential nodes and/or clusters within the network.

Email is widely accepted by the business community as the first broad electronic communication medium and was the first ‘e-revolution’ in business communication. Typically, email is used for alerting, archiving, task management, collaboration, and interoperability. According to Radicati’s 2014 Surveys[3], 108.7 billion business emails are sent and received daily (up from 89 billion in 2012 [4]). This accounts for 55.4% of the total email communication globally (196.3 billion). By 2018, this is expected to increase by 28.2% to 139.4 billion. Within an organisation, emails may be used to send messages regarding the latest football score or to discuss the latest draft of a report[5]. The diverse interactions that email mediates allow researchers a unique insight into the everyday workings of an organisation and may help reveal informal hierarchies that may not be evident to an individual outside of the organisation[6], [7].

Since the revelations of metadata collection exposed by Edward Snowden in June 2013 [8], the importance of metadata from emails is gaining awareness. In the light of these revelations, organisations are investigating the current risk exposure of their own data[9] and the extent to which the US surveillance schemes may affect their organisation. In order to collect a sufficiently large dataset along with the associated ground truth, we decided to focus on email communication networks. As these techniques are improved, it may be possible to apply these techniques in order to identify influential players within DarkNet forums or other criminal networks in order to help with criminal convictions.

Research Question and Approach

In this paper we set out to investigate the effectiveness of existing SNA techniques when applied to hierarchical analysis based upon the metadata from email communications. As there has been research on this topic in the literature (e.g., the specific objective here will be towards enhancing the accuracy of inferring these relationships and using fewer metadata elements to complete the inference. In particular, we aim to answer “***To what extent can SNA techniques be used to assess email communications metadata to identify known, but also hidden social groups***”.

We will split the research into four main tasks, namely:-

ISSN 2052-8604

- **Initial investigation**: This task focuses on implementing several of the existing SNA methods and metrics, and applying them to a communication dataset to see how well they perform in identifying groups and their structures (i.e. hierarchies of individuals). We put special emphasis to the number of data elements required to define structures and the accuracy with which these structures can be identified. For this experiment we use the Enron email communications dataset given the availability of ground truths to evaluate the methods and support our findings, and also its large size.
- **Enhancing the discovery of groups and social structures:** Having investigated the effectiveness of existing SNA techniques, we will aim to enhance the accuracy of these techniques in predicting the "hierarchical importance" of an individual. We will also introduce new methods through which groups and social structures can be identified. For an initial evaluation of these new approaches, we again use the Enron dataset.
- **Collecting a new email communications corpus**: To test our enhanced inference techniques, we collect a new communications corpus from willing volunteers and use our techniques established above to compare our predicted hierarchy with the true hierarchy in the dataset. We use the metrics identified as useful from the first two experiments.
- **Evaluating the enhanced inference methods**: At this stage, we evaluate our SNA proposals and the level of accuracy with which they can identify the known social groups (as documented in the sample's ground truth). As we are using an organisational dataset for our analysis, we are also interested in discovering whether our approaches can discover the organisational hierarchies.

Methodology

In order to address the research question aims, we began by collecting the emails from the dataset of interest. From the email collection we were able to extract the metadata from each email from which we can build our network. Once we have extracted the data from the email communication network, we then created a graph of the new social network where each node will represent an employee and each directed edge ***a*** → ***b*** represents an email sent from **a** to **b**. The weight of each edge corresponds to the number of emails sent from **a** to **b**.

Once we have created our graph, we then set out to identify metrics on our network that may be useful in helping to determine the relative "***importance***" of an individual within it. Once these metrics have been calculated for each node of our network, our next task is to apply Supervised Machine Learning (SML) to identify the metrics that are useful when

ISSN 2052-8604

determining hierarchy within the organisation. SML allows us to create a model which links the metrics to a corresponding hierarchical job "***category***" within the organisation as well as allowing us to exclude particular metrics from future experiments due to their lack of contribution. After performing SML we identified a number of useful metrics that can be used to determine the relative importance of an individual. Once our model was created, we tested the validity of our results on a real dataset. We apply our trained model to this new dataset in order to determine how accurate it is at identifying the senior management in the group.

Link Analysis and SNA

Complex interactions between entities can be modelled as networks. These networks include the Internet [10], food webs [24] and biochemical networks [15]. Each of these networks consists of a set of nodes or vertices (e.g. computers or routers on the Internet or people in a social network), connected together by links or edges, representing data connections between computers, friendships between people etc.

Link Analysis (LA) is the analysis of relationships and information flow between a network of individuals, groups, organizations, servers and other connected entities, and has been a topic of study for several decades[10], [11]. A Social Network (SN) is defined as the representation of networks with people as nodes and relationships between them as links in a graph. Social Network Analysis (SNA) is defined as the application of Link Analysis to a social network. We can perform SNA on our newly created Enron social network in order to determine the hierarchical structure of the organisation. Within a group's social network, we define the "hierarchical importance" of an individual as the seniority of the individual within the group.

SNA Metrics

Within the field of SNA, there are a range of metrics that can be used to assess a network and the nodes (individuals) within it. In this experiment we aim to assess whether these (or enhanced variations of them) could be used to determine the importance of an individual simply through a broad set of Email-Communications data.

ISSN 2052-8604

Attribute Name	Description
Sent Messages (SM)	The number of emails sent by an employee.
Received Messages (RM)	The number of emails received by an employee.
Degree Centrality (DCS)	The number of distinct employees within the network that an employee has sent emails to.
Betweenness Centrality Score (BCS)	The betweenness centrality measure for an employee.[11]
Pagerank Score (PRS)	The PageRank score an employee.[27]
Markov Ranking (MR)	The markov ranking of an employee. [20]
HITS Authority Score (HAS)	The authority score for an employee (if several users with high hub weights send an email t the user then they will have a higher authority score). [18]
HITS Hub Score (HHS)	The hub score for an employee (if the user sends emails to users with high authority scores then they will have a higher hub score). [18]
Clique Score (CS)	The number of cliques (maximal subgraphs) an employee is in using the Bron and Kerbosch algorithm.[6]
Weighted Clique Score (WCS)	The weighted clique score for each user, weighted by the number of users within each clique.
Average Distance Score (ADS)	The average distance between the user and all other users in the graph.
Clustering Coefficient (CC)	The extent to which vertices in a graph tend to cluster together. [35]

TABLE 1:- DESCRIPTION OF OUR CHOSEN SNA METRICS

Our assumption that p_2 in Figure 1 plays a central role is due to the proportion of the network that they connect with. This is formally known as the Degree Centrality of the node and is one of many SNA metrics that may be of use in our analysis. Table 1

ISSN 2052-8604

contains the metrics that we decided to investigate as part of our analysis. The metrics were chosen based on a literature review of previous research and their ability to identify nodes of influence within a SN[12]. We present these in terms of their use with our Enron dataset where the nodes represent employees and the graph edges represent email communications between employees.

Initial Investigation

For our first investigation, we used the Park et al.[13] dataset for our analysis. This was based on the original dataset of Adibi and Shetty in ISI[14], but has been modified to delete extraneous duplicate emails and fix some anomalies in the data. Our final dataset consisted of 184 email addresses corresponding to 147 employees and a total of 517,431 emails. The ground truth was obtained by investigating information available from the original dataset[14], previous papers [15], articles available online[16], [17] and the request for immediate managers issued by FERC1 which contains the job role and the immediate supervisor of 480 Enron employees[18].

In total, we chose 7 categories which reflect the hierarchical level of each employee from their organisational role based upon the generalisation of the key roles described in the official FERC report [18]. These categories are similar in nature to previous research articles [14]. Below we present the 7 categories.

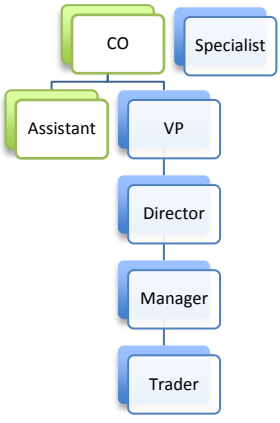


FIGURE 1:- HIERARCHY OF THE ENRON CORPORATION

- **Chief Officer (CO):-** The 11 senior C-Suite Officers in their divisions e.g. CEO etc.
- **Vice President (VP)**:- 24 employees with divisional control of 100+ employees.
- **Director**:- 24 employees who control larger teams (>60)
- **Manager**:- 29 employees with control of up to 10 employees.
- **Trader**:- 37 low-level employees who perform the day-to-day trading.
- **Specialist**:- 17 employees with specialist roles (such as IT administrator).
- **Assistant**:- 5 personal assistants to senior VPs and CO's.

Figure 1 shows the visual representation of the categories. We leave the "Specialist" category separate from the main chain of

ISSN 2052-8604

hierarchy as these individuals interact with all members of the organisation at the different levels of hierarchy and move between groups within the organisation.

Tool Support

Over the last few years several SNA tools have been developed for different purposes such as Gephi[19], GraphViz[20], VisOne[21], Netlytic[22], UCINet[23] and Socilyzer[24]. Whilst these are all ideal for their own purposes, none provided us with all the analysis that would be needed in order to calculate the selected metrics. As such, we decided to create our own tool that would allow us to calculate all the metrics identified in the previous section in the same software. Figure 2 shows a representation of our social network with our new tool.

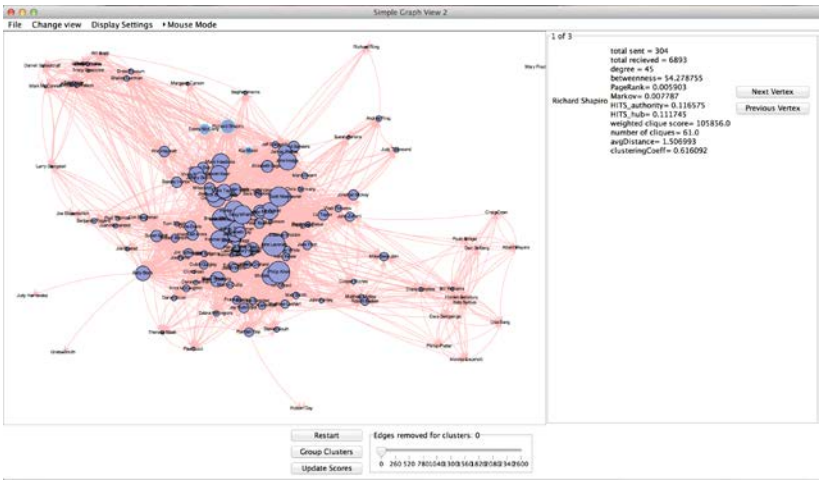


FIGURE 2:- IMAGE OF OUR NEWLY CREATED TOOL SUPPORT

Results from our initial experiment

In our first experiment, we evaluated the effectiveness of our metrics by their ability to distinguish between the 7 categories defined previously. **Error! Reference source not found.** shows the breakdown of the metrics on a category-by-category basis. A full breakdown of our results can be seen in

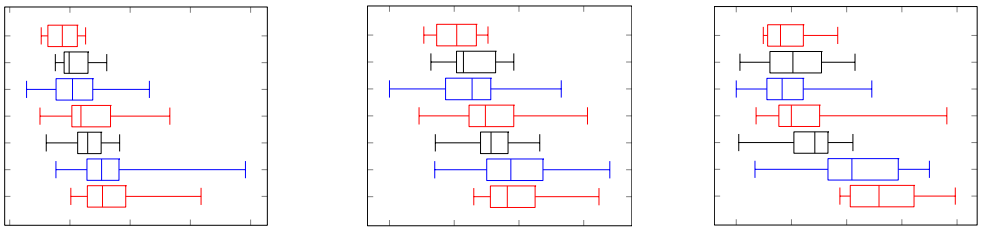


FIGURE 3:- DISTRIBUTION OF OUR TOP THREE SNA METRICS

ISSN 2052-8604

our paper[1].

The Markov Centrality scores were capable of separating off the VPs and COs and the remaining categories, but provided little distinction between other categories. Similarly, the PageRank Scores were able to provide some level of distinction between senior employees and the other categories.

As the HAS and HHS are closely related, we would have expected similar performance from both. Our initial results confirm this and showed that the HAS was significantly greater on average for VPs and Cos when compared to the other categories, leading us to believe that HAS may be a useful indicator of seniority within the network, however there are still some VPs with a low HAS. Our results also showed that the WCS outperformed CS. If an individual has a Weighted Clique Score greater than 200,000, then they have a high likelihood of being in one of the more senior categories. Conversely, all of our traders had a score less than 200,000. This leads us to believe that there may be a stronger correlation between the WCS and the employee category than between the CS and employee category.

Both the NSM and NRM were useful in identifying assistants, managers and directors as they sent comparatively fewer messages, but was not able to help distinguish further. The DCS and the BCS were useful in distinguishing some of the categories. The DCS metric proves effective at distinguishing between COs/VPs and other categories, and was useful in identifying senior employees. The BCS was able to help highlight COs and other senior members within the organisation, but several mid- seniority Managers also had high scores and these outliers may restrict the metric's utility.

The ADS were noticeably good at distinguishing between the COs and the other categories (with the exception of Assistants) as COs tended to have an ADS of 1.5 or greater whereas those that were not in a position of authority had a lower ADS. It was less good, however, at distinguishing between the employees of lower seniority. The CCS proved ineffective when attempting to find a correlation with the employee category. Alone, it gave little insight into the difference in employee categories.

[1] http://www.cs.ox.ac.uk/people/elizabeth.phillips/

ISSN 2052-8604

Summary

Many of the conclusions from our initial analysis coincide with some real world assumptions. The ADS, for example, was expected to provide a good distinction between COs and other categories as most employees would not contact the CO directly but would communicate through their line manager.

Similarly, due to the nature of the HHS and HAS metrics, a higher HAS for senior management is expected as lower hubs (i.e. employees of lower seniority) would send several messages to them and they would also send numerous messages to lower-seniority employees. The WCS was expected to be useful as many COs would be the critical nodes in the graph and as such, would be part of many more complete sub-graphs (and in turn, gain a higher WCS). From the initial investigation, it emerged that there are a number of potentially useful metrics that can aid in identifying individuals of hierarchical importance within an organisation or group. We therefore decided to test these metrics in order to assess their effectiveness in a more rigorous manner.

Enhancing discovery of social groups and hierarchies

In order to calculate the social structure, we applied a Machine Learning approach to associate the metrics with the role Category. This would allow us to use the metrics obtained above and the ground truths to train a model that would predict the employee's category based only on the SNA metrics of the employee.

ISSN 2052-8604

Actual Category	Classified as						
	CO	VP	Director	Manager	Trader	Specialist	Assistant
CO	9	1	0	0	0	0	1
VP	10	4	6	1	3	0	0
Director	1	4	6	0	13	0	0
Manager	2	3	4	0	20	0	0
Trader	1	2	7	1	26	0	0
Specialist	0	3	2	0	11	1	0
Assistant	1	0	1	0	1	1	0

TABLE 2:- CONFUSION MATRIX FOR INITIAL CATEGORIES

To test the ability of the supervised learning algorithm to predict the employee category, we began by testing the dataset using a Bayesian Network Classifier. In order to validate the created models, we used 10-fold cross validation. Table 2 shows the classification results of the Bayesian Network model in a confusion matrix. The table revealed that 20 of the 29 Managers were incorrectly classified as Traders. This discrepancy could be due to the structure of the underlying network. Within the Enron corporation, many individuals were assigned the role of a manager but were only managers of small teams and were performing the role of a trader. This problem is exacerbated further due to the discrepancies between the ground truth sources.

In order to address this problem, we reduced the number of categories from seven to two, as we were primarily interested in identifying the senior employees. The new “Boss” category corresponded to the previous CO and VP categories whilst the “Not_Boss” corresponded to the remaining five categories. Despite the lower level of granularity of the employer’s category that we were now able to predict, it allowed us to focus on highlighting the employees of greatest interest within the organisation.

Breakdown of reclassified data

Table 3 shows the statistical breakdown of the network once they have been reclassified using the 2 new categories while Table 3 shows a breakdown of some of the most useful metrics. From the analysis of the figures, we were able to identify the metrics that have a

ISSN 2052-8604

different distribution of values for each category, which in turn makes them potentially useful contributors to the Machine Learning algorithm in order to distinguish between the two categories. In particular, ADS, DCS, HAS, WCS and MCS all showed a distinction between the two categories and hence they may be useful metrics.

Attribute	Category	REC	DEG	BC	PR	MAR	HAS	HHS	WCS	CS	ADS	CC
Max	Boss	6893.00	132.00	1889.20	0.0196	0.0170	0.20	0.28	369852.00	490.00	1.70	0.70
Max	Not Boss	2972.00	92.00	1507.12	0.0133	0.0153	0.19	0.27	338456.00	360.00	1.67	1.00
Min	Boss	216.00	22.00	26.81	0.0051	0.0065	0.05	0.02	692.00	14.00	1.39	0.24
Min	Not Boss	0.00	1.00	0.00	0.0014	0.0000	0.00	0.00	1.00	1.00	1.29	0.00
Mean	Boss	1414.20	56.94	335.43	0.0089	0.0101	0.12	0.11	106370.29	114.14	1.52	0.48
Mean	Not Boss	530.18	28.80	128.95	0.0057	0.0066	0.05	0.04	13218.01	37.08	1.43	0.56
StdDev	Boss	1505.63	24.30	383.09	0.0035	0.0029	0.04	0.06	111219.20	109.60	0.06	0.11
StdDev	Not Boss	583.48	16.91	208.03	0.0021	0.0029	0.03	0.04	45734.44	51.01	0.07	0.17

TABLE 3:- RESULTS FROM RECLASSIFIED DATA

Once we had created our two new categories, we tested the effectiveness of our new model using a variety of different Machine Learning Methods. In total we selected seven models, namely Naive Bayes (NB), Bayesian Network (BN), Multi-Layer Perceptron Model (MLP), IB1, K-Star and SMO, and compared them to random guessing. The overall best performing classifier is the MLP, with the NB and BN close behind by providing a greater True Positive (TP) rate for the Boss category and producing a greater Receiver Operating Characteristic (ROC) curve area and F-Score. A higher F-Score and ROC curve area is an indication of a good classifier.

Random Guessing						
TP	FP	Precision	Recall	F-Score	ROC Area	Class
0.522	0.455	0.934	0.522	0.67	0.5	Not_Boss
0.545	0.478	0.085	0.545	0.146	0.5	Boss

Multi-Layer Perceptron						
TP	FP	Precision	Recall	F-Score	ROC Area	Class
0.956	0.273	0.977	0.956	0.967	0.939	Not_Boss
0.727	0.044	0.571	0.727	0.64	0.939	Boss

TABLE 4:- ANALYSIS OF OUR MLP MODEL VS. RANDOM GUESSING

Table 4 shows a breakdown of the results for our MLP model vs. random guessing. Our results show us that by categorising the Enron dataset into two categories and by introducing the new metrics and categorisations, we have been able to predict whether an individual is a Boss with an F-Score of 0.64 and an ROC Area of 0.939 compared to random guessing which achieved 0.146. It also identified five critical attributes, namely

ISSN 2052-8604

WCS, ADS, HAS, HHS and DCS. This has enabled us to improve on existing metrics, which are accurate to only 82.37% [25]and 87.58%[26] respectively.

Summary

From our analysis using our new role categories, we were able to identify five metrics that have a different distribution for Bosses than ordinary employees, which in turn can make them useful contributors to our model to predict the employee's role category. In order to quantify how effective each metric was, we decided to use machine learning metric evaluators. In particular we used the Relief-F evaluator[27] which was chosen for its consistency and its ability to cope with the dependence between our attributes.

Experiment 2

For our second experiment, our new dataset was considerably smaller than the Enron dataset and represented the communications amongst a single group. For this group, we collected a total of 6,936 emails sent amongst the ten members of the group over a twelve-month period from 20 June 2013 to 20 June 2014. Each email was sent to an average of 1.97 recipients. As our data-collection scripts hide the identity of email recipients of emails sent outside of the group, the actual number of recipients in an email may well have been much higher than this.

After establishing our initial network, we then proceeded to collect the ground truth for the actual hierarchical structure of the network. Within this network, there was one official Boss for the research group (Employee #0) who acted as the main supervisor for many (but not all) of the projects. Employee #4 was also in a unique position as they had worked on a variety of different projects with various members of the group in the past. They are considered a senior member in the group because of the various interactions across projects (often simultaneously) and we therefore categorised employee #4 as a Boss as well.

ISSN 2052-8604

From the initial network, we discovered that the graph was almost fully connected, with 84 out of the 90 possible edges between the ten employees established based on their email communication which led to some of our SNA metrics being ineffective as they were unable to differentiate important connections from insignificant ones. For each email sent, we add 1 to the thickness of each graph edge. The distribution of weights was expected given the small size of the group and the interaction between members for non work-related purposes associated with a close research group.

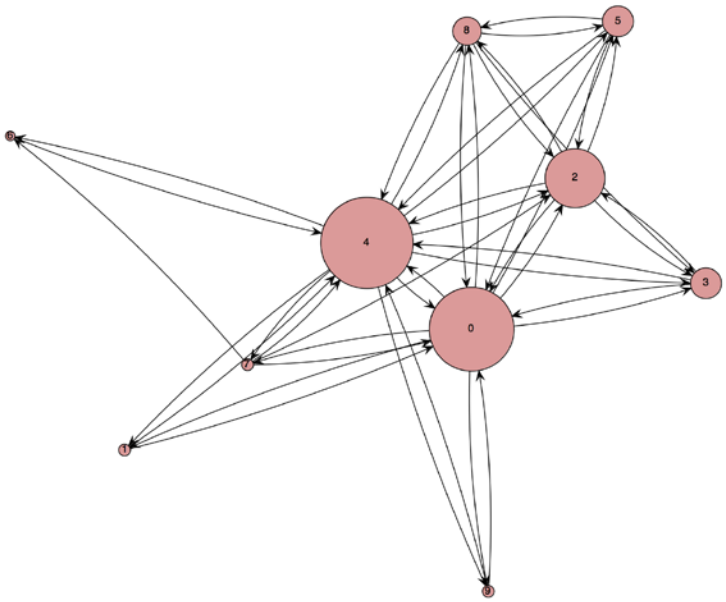


FIGURE 4:- GRAPH REPRESENTATION OF OUR NEW

In order to overcome this, we decided to only consider edges of weight 30 or more in order to only identify strong ties between members. Whilst this pruning might lead to us potentially missing some important connections, it is more important to prune the edges that may not have been central to the work-focused network. 17 shows the structure of the new network with nodes sized according to their Authority score and edges of weight 29 or less removed from the network and is laid out using the force-directed layout of Fruchterman[28] and uses the notion of “force” and connectivity between nodes and their edges to determine where they should be placed.

The graph immediately identifies employee #4 and employee #0 as strongly connected nodes due to their close positioning in the graph (with 1095 emails sent between the 2 employees). It also identified employee #9 as an employee that is linked to only a few members in the group; this reflects the fact that employee #9 only worked on one project

ISSN 2052-8604

with the 2 senior members of the group and as such, had little collaboration with other members. Similarly, Employee #6's distance from the cluster reflects the fact that they had only recently joined the group (March 2014).

Results from experiment 2

Employee	WCS	HAS	DCS	ADS	MCS	Class
0	6.168	0.459	0.8	1.900	0.195	Boss
1	1.149	0.175	0.2	1.563	0.043	Not_Boss
2	3.870	0.409	0.55	1.692	0.128	Not_Boss
3	1.320	0.314	0.35	1.600	0.088	Not_Boss
4	7.316	0.471	0.9	2.000	0.249	Boss
5	1.320	0.314	0.4	1.643	0.087	Not_Boss
6	1.149	0.144	0.15	1.529	0.032	Not_Boss
7	2.380	0.175	0.3	1.643	0.044	Not_Boss
8	2.639	0.302	0.45	1.692	0.092	Not_Boss
9	1.149	0.175	0.2	1.563	0.043	Not_Boss
Average	2.846	0.294	0.43	1.682	0.100	
StdDev	2.252	0.124	0.254	0.153	0.072	

TABLE 5:- RESULTS FROM EXPERIMENT 2

Table 5 presents the distribution of employees and their metric scores. The results of our new analysis support our previous theory as both employee #4 and #0 are the true Bosses and have notably higher scores than the other employees. All other employees' scores are less then 1 standard deviation above the mean for each of the top 5 metrics. This finding strengthens our initial belief that these metrics are a good measure of "hierarchical importance" within an organisation.

The results of our second experiment demonstrated that the metrics identified in Experiment 1 performed as expected and were reasonably effective at distinguishing between the two employee categories. This confirmed the utility of using the 5 metrics (especially the Weighted Clique Score) in allowing the inference to be made from email-communication metadata to the hierarchical structure of a group.

The work assumes that supervisors and bosses are active users of email in order for the communication network to reflect the true communications within the network. Whilst some management styles prefer to use other tools (such as phone calls) to communicate, if we were able to collect this form of data, then our abstraction of the email

ISSN 2052-8604

communications to a social network would allow it to be incorporated into our network by increasing the edge weight based on the type of communication, so as to create a new network which better reflects the underlying hierarchy, on which we can perform the same SNA analysis.

Conclusions and Future Work

Our results have identified five SNA metrics which have proved effective in distinguishing between the employees that are assigned a Boss category and those who are assigned to a Not_Boss category based only on the email communications between them; namely Weighted Clique Score, HITS Authority Score, Average Distance, Markov Centrality Score and Degree Centrality Score.

The primary value of our research is the improvement in selecting and improving on existing metrics whilst using the minimum amount of data, so as to enable the methods to be applied to any generic communications network including Dark Net Forums, Social Networking Sites as well as phone records and other offline communication networks such as face-to-face meetings.

One direction of future research is to apply our metrics to a communications network established from other sources such as the 2012 dataset extracted from the ISI-KDD Challenge of the Dark Web forums 2[2]. This should allow us to identify the most influential contributors to the forum which may help identify the ring-leaders of criminal groups that use the forums. Another direction our research could take is within Insider Threat Detection within organisations. This in turn could be a feature of Machiavellianism, which as one of the Dark Triads personality traits [29]could be a potential predictor for a malicious insider. Further research would be required to investigate to what extent uncharacteristically high influence relates to Insider Threat Detection.

References

[1] L. C. Freeman, "***Centrality in social networks conceptual clarification***," *Soc. Netw.*, vol. 1, no. 3, pp. 215–239, 1979.

[2] J. Travers, S. Milgram, J. Travers, and S. Milgram, "***An Experimental Study of the Small World Problem***," *Sociometry*, vol. 32, pp. 425–443, 1969.

[2] Available at http://128.196.40.222:8080/CRI Indexed new/datasets/ansar1.txt

ISSN 2052-8604

[3] Sara Radicati, "***Email Statistics Report, 2014 - 2018***," Radicati Group, Apr. 2014.

[4] Sara Radicati, "***Email Statistics Report, 2012 - 2016***," Radicati Group, Apr. 2012.

[5] Chron, "T***he Use of Email in Business Communication***," *Small Business - Chron.com*. [Online]. Available:http://smallbusiness.chron.com/use-email-business-communication-118.html. [Accessed: 22-Jun-2014].

[6] Atul Kachare, "***Analysis and Visualization of E-mail Communication Using Graph Template Language***," *SAS Glob. Forum*, 2013.

[7] L. Sproull and S. Kiesler, "***Reducing Social Context Cues: Electronic Mail in Organizational Communication***," *Manag. Sci.*, vol. 32, no. 11, pp. 1492–1512, Nov. 1986.

[8] "***Edward Snowden***." [Online]. Available: http://www.theguardian.com/world/edward-snowden. [Accessed: 21-Jun-2014].

[9] D. Wright and R. Kreissl, "***European Responses to the Snowden Revelations: A Discussion Paper,***" IRISS, Dec. 2013.

[10] L. Getoor and C. P. Diehl, "***Link Mining: A Survey,***" *SIGKDD Explor Newsl*, vol. 7, no. 2, pp. 3–12, Dec. 2005.

[11] S. Wasserman, ***Social Network Analysis: Methods and Applications***. Cambridge University Press, 1994.

[12] T. Coffman, S. Greenblatt, and S. Marcus, "***Graph-based Technologies for Intelligence Analysis***," *Commun ACM*, vol. 47, no. 3, pp. 45–47, Mar. 2004.

[13] Park, "Enron employee status." [Online]. Available:http://cis.jhu.edu/~parky/Enron/employees. [Accessed: 24-Jun-2014].

[14] J. Shetty and J. Adibi, "***The Enron email dataset database schema and brief statistical report***," *Inf. Sci. Inst. Tech. Rep. Univ. South. Calif.*, vol. 4, 2004.

[15] G. Creamer, R. Rowe, S. Hershkop, and S. J. Stolfo, "***Segmentation and automated social hierarchy detection through email network analysis***," in *Advances in Web Mining and Web Usage Analysis*, Springer, 2009, pp. 40–58.

[16] "***John Arnold: Ex-Enron billionaire trader retires at 38 | Mail Online,***" *Daily Mail Online*. [Online]. Available: http://www.dailymail.co.uk/news/article-2138890/John-Arnold-Ex-Enron-billionaire-trader-retires-38.html. [Accessed: 15-Jun-2014].

[17] R. Partington, "***The Enron cast: Where are they now? - Financial News,***" *Financial News*. [Online]. Available: http://www.efinancialnews.com/story/2011-12-01/enron-ten-years-on-where-they-are-now. [Accessed: 15-Jun-2014].

[18] federal energy regulatory commission subpoena duces tecum, "Request no. 11" [Online]. Available: https://raw.githubusercontent.com/diehl/Enron-GraphML-Data-

ISSN 2052-8604

Documentation/master/EnronManagerSubordinateRelationships.pdf. [Accessed: 15-Jun-2014].

[19]"***Gephi, an open source graph visualization and manipulation software.***" .

[20]J. Ellson, E. Gansner, L. Koutsofios, S. North, and G. Woodhull, "***Graphviz— Open Source Graph Drawing Tools,***" in *Graph Drawing*, vol. 2265, P. Mutzel, M. Jünger, and S. Leipert, Eds. Springer Berlin Heidelberg, 2002, pp. 483–484.

[21]***"visone."*** [Online]. Available: http://visone.info/. [Accessed: 22-Jun-2014].

[22]"***Netlytic.org.***" [Online]. Available: https://netlytic.org/home/. [Accessed: 24-Jun-2014].

[23]S. P. Borgatti, M. G. Everett, and L. C. Freeman, ***Ucinet for Windows: Software for Social Network Analysis***. Analytic Technologies, 2002.

[24]"***An Easy-to-Use Social Network Analysis Tool - Socilyzer.***" [Online]. Available: https://socilyzer.com/. [Accessed: 24-Jun-2014].

[25]E. Gilbert, "***Phrases That Signal Workplace Hierarchy,"*** in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, New York, NY, USA, 2012, pp. 1037–1046.

[26]A. Agarwal, A. Omuya, A. Harnly, and O. Rambow, "***A Comprehensive Gold Standard for the Enron Organizational Hierarchy,***" in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, Stroudsburg, PA, USA, 2012, pp. 161–165.

[27]M. Robnik-Sikonja and I. Kononenko, "***An Adaptation of Relief for Attribute Estimation in Regression***," in *Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, CA, USA, 1997, pp. 296–304.

[28]T. M. J. Fruchterman and E. M. Reingold, "***Graph Drawing by Force-directed Placement***," *Softw Pr. Exper*, vol. 21, no. 11, pp. 1129–1164, Nov. 1991.

[29]J. McHOSKEY, "Narcissism and machiavellianism," *Psychol. Rep.*, vol. 77, no. 3, pp. 755–759, Dec. 1995.

ISSN 2052-8604

Information trustworthiness as a solution to the misinformation problems in social media

Jason R.C. Nurse[1], Ioannis Agrafiotis[1], Michael Goldsmith[1], Sadie Creese[1], Koen Lamberts[2], Darren Price[3], and Glyn Jones[3]
[1]Cyber Security Centre, Department of Computer Science,
University of Oxford, Oxford, UK
{*firstname.lastname*}@cs.ox.ac.uk

[2]University of York, UK
koen.lamberts@york.ac.uk

[3]Thales UK Limited, Research and Technology
{*firstname.lastname*}@uk.thalesgroup.com

Abstract

The advent of the Internet has reshaped the way we communicate and interact in our daily lives. It provides an ideal medium through which we can share information and ideas, form groups, and contribute to a variety of discussions. In this position paper, we focus specifically on the information now available online – especially content from social media – to consider in detail the challenges that such information poses to modern-day society. Typical examples of challenges include the prevalence of mistaken information and deliberate misinformation and rumours. With an understanding of these challenges, we then introduce the notion of information-trustworthiness measures as a potential solution to the problem of misinformation in social media. The idea here is to use quality and trust metrics to assess information, and then, based on values attained, advise users whether or not they should trust the content. This paper extends our previous research in the field by assessing the misinformation problem in much greater detail, and also presenting our current agenda for future work.

Introduction

The Internet has revolutionised the way that we, as humans, communicate and interact with each other. It provides a ubiquitous and, in many ways, ideal medium, through which we can share information and ideas, contribute to a range of discussions, and discover more about the world around us. Given its suitability for communication there should be

ISSN 2052-8604

little surprise at the extent to which it is currently used and the vast amount of data being shared every day, in particular via social media. To take Facebook as an example, each day 2.5 billion content items are shared (including information, photos, posts) [1].

Similar to their more traditional predecessors, social media have become a critical tool in influencing people's perceptions and decisions. Whilst this influence is often positive and well-intended (e.g., a tweet from a local council informing motorists of a blocked road), real-world cases continue to show instances of individuals poisoning information for their own malicious ends, and unfortunately, with serious consequences. The case of the London Riots in 2011, where deliberate rumours led to confusion over where emergency services should be deployed [2], is one example.

In this position paper, we reflect on the problem of misinformation on social media, and the use of information quality and trust metrics to help address it. This paper considers and also builds on previous research to outline an agenda for our future research aimed at addressing these outstanding issues. The main aim is developing a full system that is able to consume social content, assess the trustworthiness that should be associated with it, and generally help understand what might be happening in on-going scenarios such as emergencies or crises.

The misinformation problems with social media

Social media have provided us with many opportunities to discover, learn and interact. Unfortunately, however, there are several problems accompanying this capability, one of the largest being the misinformation or information poisoning problem (i.e., the posting of inaccurate or misleading information) and its use to negatively influence individuals. Take the examples below.

In the summer of 2011 several British cities experienced a significant period of unrest with spates of rioting and looting. The violence originated in London, but rapidly spread across the UK mostly affecting large cities including Manchester and Birmingham. In the aftermath, social media were put in the spotlight. Governmental authorities claimed that information poisoning facilitated the spread of rioting, either via circulating rumours presenting an overly chaotic situation or via sharing photos of police officers who remained indifferent while looting was taking place in their presence [3]. The role of social media in encouraging the riots and disrupting essential response was deemed so critical, that even the prospect of temporarily blocking access to Twitter and Blackberry

ISSN 2052-8604

Messenger was raised by authorities. This gives some insight into the significance of the problem faced and challenges to official responders.

One of the main avenues in which rumours were spread during the riots was the micro-blogging platform, Twitter. According to retrospective reports, thousands of individuals re-tweeted dubious content leading to a sea of misinformation as the incident unfolded [4]. What is of great interest, though, is the extent to which people appeared to question their knowledge and common sense to embrace the rumours. For instance, an image portraying the London Eye in flames was heavily re-tweeted initially, and only after being online for a while did someone expressed doubts about the trustworthiness of the tweet; they rightly noted that the London Eye is made of iron and thus, it was difficult to imagine it ablaze. Even after the tweet debunking the rumour, more than 700 people within the next three hours re-tweeted the image expressing their anger at the destruction of the London attraction [5].

An additional problem was the enormous amount of data generated as a response to such tweets. This had a direct negative impact on police efforts to analyse the situation in the places where riots were taking place and to respond accordingly. Chris Sims, chief constable of West Midlands police, said his “force was actively engaged in trying to dispel information it believed to be untrue”, thus wasting valuable police resources [2]. In addition, the gold commander of Greater Manchester police described the amount of data from social media as overwhelming, recognising that “police struggled to analyse it even in the most basic way”, and also calling for innovative systems to elicit actionable intelligence from social media in an effective and quick manner [2].

Another case where misinformation from social media affected people's decisions with dramatic consequences was the Boston Marathon bombing in 2013 [6]. Within seconds of the first explosion, speculation, rumours and reactions from the masses dominated social media discussions. While first responders were on route to the incident, there were posts reporting additional explosions, library buildings being targeted, increased casualties, and even accusations against the Muslim community as being responsible for the attack [7]. Although the motives behind these rumours may not all have been malign, certainly such misinformation hindered authorities in allocating their resources effectively.

An example of the potentially devastating impact of such misinformation emerged from the rush to identify the perpetrators of the bombing attack. Once the FBI released photos from the scene where it took place, several social-media users responded by reviewing

ISSN 2052-8604

the information and naming anyone that looked similar as a potential suspect [7]. This took a dramatic turn when a tweet claiming that the Boston Police department had declared Sunil Tripathi and Mike Mulugeta as suspects, went viral, with thousands of individuals re-tweeting the names. Possibly as a result, Sunil Tripathi, who had nothing to do with the case, disappeared the same day and was found dead one month later [7].

From the cases above it is evident that social media can be exploited to misinform and to circulate inaccurate information with sometimes devastating consequences, even the loss of innocent lives. The need to develop mechanisms to evaluate the quality and trustworthiness of social-media information is therefore more urgent now than ever before.

Measuring the trustworthiness of online content

Previous research

Information quality and trustworthiness have been of interest to researchers for some time. To assess the quality of information, a typical question is, how fit is the information for its intended use. Trustworthiness can be thought of as an extension of quality, as it looks at the perceived likelihood that a piece of information will preserve a user's trust and belief in it [8]; presuming the information is of high quality therefore, the likelihood might arguably be high as well.

There have been numerous proposals that aim to utilise the quality and trust factors identified above to measure the trustworthiness of social content automatically. Agichtein et al., for instance, focus on the problem of finding high-quality content in social media and propose a classification framework for combining evidence (especially related to the quality factors discussed prior) from different sources of information [9]. As it pertains to the trustworthiness and credibility of online content, Castillo et al. draw on similar general factors (regarding features of the message, the information's source, and the topic) and use a supervised classifier (machine learning) to produce automated measurements of a tweet's credibility [10]. These are just two of the many approaches that aim towards this problem; space limits how much we can cover here, but readers are free to read more in [11]. Through the use of these automated techniques there is hope for a more general approach to tackle the misinformation problems plaguing online content.

Our work in the TEASE project

ISSN 2052-8604

The TEASE research project was born out of the need to address the misinformation problems commonly faced with online social-media content. Our objective was to research and prototype a computer system that was able to measure the trustworthiness of information, and feed this back to users to assist them in making decisions. There were several significant contributions made by TEASE. The first was a novel methodology and framework for assigning trustworthiness measures to openly-sourced information, including tweets, Facebook posts, and news reports [12-13]. This approach considered key trustworthiness aspects, including provenance, intrinsic quality, and infrastructure integrity, and their related sub-factors such as the identity of a source, their reputation and competence, how timely the information was, and the vulnerabilities and threats to the infrastructure through which information traversed before reaching the user. Through an analysis of information (and its related metadata) in terms of these factors, we were able to produce trust scores (one per item) that could then be displayed along with the related content. These would therefore help to identify misinformation early on and hopefully prevent its spread.

With regard to the user interface and ensuring that it was highly usable, we engaged in numerous user experiments, both with the general public, and for specific use cases, with experts (e.g., in crisis management). There were several notable findings from our experimentation. For instance, traffic lights are much more effective communicators of trustworthiness than other visual means such as stars or transparency [13]; that is, lights were better able to direct individuals away from bad information and towards good information. Another crucial finding was that individuals are astoundingly capable of combining trustworthiness ratings and evaluative information to make efficient judgements [13]. The experiment in this case was based on the common assumption that individuals can easily combine sets of information (e.g., tweets describing what's happening in a scenario) and their respective trust scores (e.g., assignments of various trustworthiness levels to the tweets) to first, understand what might be happening in the scenario, and then to make decisions. Both these findings assisted in our interface design but also contributed to broader research in the field of communicating quality and trust.

Looking towards the future

This section looks towards the future and ways to extend current research to tackle the outstanding challenges of misinformation in social-media. We propose a research and development agenda which draws on our previous work, and is concentrated on the use of social information for official response purposes.

ISSN 2052-8604

Social media present society with a plethora of opportunities, especially with regards to information to make decisions. The only way that these can be realised, however, is if the users of online content are able to identify inaccurate and misleading information, and have the tools to isolate high quality content. TEASE tackled this problem with notable success, in the creation of a flexible framework for measuring trustworthiness and an interface that emphasised usability. Nonetheless, there were important areas unable to be completely addressed in the lifetime of the project. One of these areas was the creation of a fully automated system, capable of working with live Internet feeds. The real challenge here is the research and design of a scalable system able to consume content about a specified topic (e.g., a bombing in Boston), use the TEASE methodology to measure the trustworthiness of all the items, and present information and annotated trustworthiness levels back to users in a timely manner. This is all with the understanding that in crises, there are typically hundreds of social-media posts per minute, a myriad of new users joining to contribute (thus, persons with unknown reputation levels), and metadata about content often missing (e.g., the location of an information source is key to assessing an eyewitness attribute).

Another feature that would be extremely valuable in such a system is the notion of World Views introduced in [12]. A World View is a cluster of social-media information (e.g., tweets and posts) that is related to each other (i.e., about the same topic) and is somewhat consistent, i.e., there is little discrepancy between the information items. Our research pursuit with respect to World Views therefore, would be defining how to create the clusters. We envisage an approach involving Natural Language Processing (to better understand the information and facilitate comparison) and formal modelling (to build consistent clusters). Even then, considering that the range of text is so expansive, it will be crucial to scope the problem – this is another reason that we have chosen crisis response. In this field, there are several existing encoding formats for content that will be invaluable. Additionally, we will be able to blend social-media content with closed-source intelligence (e.g., reports from emergency-service personnel) within World Views to create a more complete picture for responders.

With a fully functional system, the next aim will be evaluating it, and particularly its use in supporting decision-making during crisis situations. We propose a set of experiments where experts use the system first within a controlled context, where we can carefully monitor for any usage issues, and then, once any feedback has been incorporated, in the field. To clarify, we do not envisage a fireman with a tablet PC searching through rubble, but rather, a control centre directing first responders based on information now marked

ISSN 2052-8604

with trustworthiness scores. The utility of the system could be judged based on interviews and questionnaires after response to events.

Acknowledgements: TEASE was a collaborative research project that involved the University of Oxford, University of Warwick, HW Communications Ltd and Thales UK Research and Technology. The project was supported by Innovate UK's Trusted Services Competition (www.innovateuk.org) and the Research Councils UK Digital Economy Programme (www.rcuk.ac.uk/digitaleconomy).

References

[1] CNET News: Facebook processes more than 500 TB of data daily (2012) http://news.cnet.com/8301-1023 3-57498531-93/facebook-processes-more-than-500-tb-of-data-daily.
[2] Guardian News: Riot rumours on social media left police on back foot (2012) http://www.theguardian.com/uk/2012/jul/01/riot-rumours-social-media-police.
[3] Guardian News: UK riots 'made worse' by rolling news, BBM, Twitter and Facebook (2012) http://www.theguardian.com/media/2012/mar/28/uk-riots-twitter-facebook.
[4] Guardian News: How twitter was used to spread and knock down rumours during the riots (2011) http://www.theguardian.com/uk/2011/dec/07/how-twitter-spread-rumours-riots.
[5] Guardian News: How riot rumours spread on twitter (2011) http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter.
[6] Los Angeles Times: Boston bombings: Social media spirals out of control (2013) http://articles.latimes.com/2013/apr/20/business/la-boston-bombings-media-20130420.
[7] The Atlantic: #bostonbombing: The anatomy of a misinformation disaster (2013) http://www.theatlantic.com/technology/archive/2013/04/-bostonbombing-the-anatomy-of-a-misinformation-disaster/275155/
[8] Kelton, K., Fleischmann, K.R., Wallace, W.A.: Trust in digital information. Journal of ASIST 59(3) (2008) 363-374
[9] Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: International Conference on Web Search and Data Mining, (2008) 183-194
[10] Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: International Conference on World Wide Web, ACM (2011) 675-684
[11] Moturu, S.T., Liu, H.: Quantifying the trustworthiness of social media content. Distributed and Parallel Databases 29(3) (2011) 239-260
[12] Rahman, S.S., Creese, S., Goldsmith, M.: Accepting information with a pinch of salt:

ISSN 2052-8604

handling untrusted information sources. In: Security and Trust Management Workshop. Springer (2012) 223-238
[13] Nurse, J.R.C., Agrafiotis, I., Goldsmith, M., Creese, S., Lamberts, K.: Two sides of the coin: measuring and communicating the trustworthiness of online information. Journal of Trust Management 1(1) (2014) 1-20

ISSN 2052-8604

Practical attacks on PXE based boot systems

Philipp Holler[1], Christian Roth[2], and Hartmut Richthammer[2]
[1]Department Business Information Systems IV - IT Security Management, University of Regensburg, Germany
Philipp.Holler93@gmail.com

[2]Department Business Information Systems IV - IT Security Management, University of Regensburg, Germany
{firstname.lastname}@ur.de

Abstract

Sustainable energies are an important resource for the twenty-first century and the generation of renewable energy raised during the last years. Decentralised energy production rises with households being able to feed wind and solar power into the grid. To handle this bidirectional flow of energy and hold the whole system in balance, an innovative management is necessary. This was the birth of the smart grid idea. The organisation and handling of the grid and all stakeholders is managed in a Supervisory Control And Data Acquisition (SCADA) system centre. To administrate all computer devices within a SCADA centre, one solution might be thin clients with Preboot Execution Environment (PXE) boot up because network boot systems are an elementary part of a modern system management solution. An important fact of such an approach is the security aspect. However, the PXE standard comes without any security functionality.

In our research, we analyse possible attack vectors and create a testbed to simulate different attacks on such an infrastructure successfully. The attack vectors are based on the protocols Dynamic Host Configuration Protocol (DHCP) and Trivial File Transfer Protocol (TFTP)

Introduction

The sustainable energy revolution calls for new concepts on the energy industry sector. Nowadays energy production and energy feeding is not only possible by industry companies. Also every household can participate as an actor on this market. For this purpose it is important to manage the production, transmission, flow and distribution of energy. Otherwise, the grid structure could be damaged if insufficient or too much energy is available. The Smart Grid (SG) [5] could be a solution for this challenge.

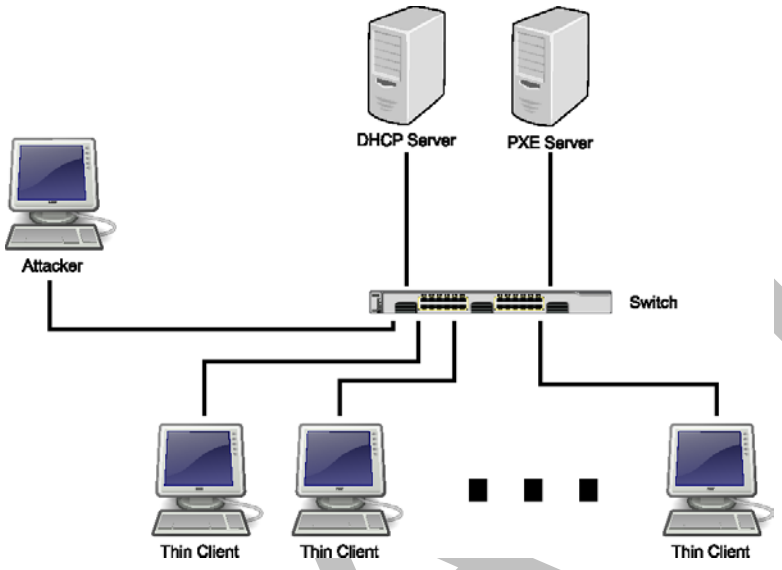


FIGURE 5: EXAMPLE OF A DISTRIBUTED SYSTEM INFRASTRUCTURE[3].

To administrate and handle the "smart" technology, a centralised management system is necessary. One subsystem in the approach from Metke and Ekl [9] is the SCADA centre, which has a core role in this infrastructure. A SCADA centre is a computer centre which administers distributed systems centralised. To manage this task, a lot of different computer systems and devices with installed Operating Systems (OSs) and software are necessary within this centre. However, software distribution is a challenge especially in larger IT environments found in industry. To reduce the administration and assistance cost for a SCADA carrier and to stay competitive, a common practice therefore in the industry is the usage of operating system images provided by a centralised instance [3]. Such images can be rolled out to thin clients over network using a widespread protocol named Preboot Execution Environment (PXE) [7]. For instance, Thinmanager[4] is a software vendor for thin client and software management in distributed environments whose products also use the contemplated protocol to distribute software. But defining the perimeter of a SCADA network and proper access control can be a challenge [12]. One common misconception regarding to the security of SCADA systems and its network infrastructure was that the nodes are electronically isolated [1,10].

[3] *Some picture elements were taken from http://www.openclipart.org (ujmoser, warszawianka, Rob Fenwitch) and stay under the GNU Public License.*

[4] *http://www.thinmanager.com*

ISSN 2052-8604

In figure 1, an example of a distributed system infrastructure is depicted. A central DHCP and PXE Server instance distributes Internet Protocol (IP) addresses and boot file images to Thin Clients. The detailed procedure and explanation how this works will be explained in section 2.

In SGs and SCADA systems, a lot of sensitive and privacy relevant information and datasets are processed [2]. As a result security is an important factor and should not be neglected in this topic. In our study, we analysed the PXE protocol with the focus on security flaws. The application of PXE without security concepts can be a high risk for the confidentiality, integrity and availability of the SCADA centre and the whole SG infrastructure. During our research we find out, that there are only a few publications which consider security in combination with the PXE protocol. References for the implementation of PXE into an existing network structure can be found very easily but most of them did not consider the security aspect.

The remainder of this paper is structured as follows. First we describe the basic principles of PXE in section 2. In section 3 we illustrate two attacks on the PXE protocol and touch some further attacks. After concluding the paper in section 4, we give an outlook of further work in the context of network software distribution in section 5.

PXE Basics

PXE was initially developed by Intel as a successor of the former Bootstrap Protocol (BOOTP) and basically combines the Dynamic Host Configuration Protocol (DHCP) [4] and the Trivial File Transfer Protocol (TFTP) [11] to provide information about the network infrastructure as well as transmitting bootable programs called Network Bootstrap Programs (NBPs). On the client's side PXE is usually implemented in the firmware of the computer's Network Interface Controller (NIC), but in rare cases it may also be loaded from another medium like a CD or the client's hard disk. The current version 2.1 was released in 1999 [7].

DHCP servers are used to dynamically distribute network information such as IP addresses in an IP based network reducing the management effort for network administrators. In PXE environments, DHCP servers are also responsible for offering information on how network images provided by (multiple) TFTP servers can be accessed. Therefore the standard DHCP ports are used and commands such as DHCPDISCOVER or DHCPOFFER are enriched with PXE related metadata, namely a

ISSN 2052-8604

client's architecture, its network identifier and its machine identifier in form of a GUID, for instance [8]. Since PXE is built on top of existing protocols, non-PXE-enabled DHCP servers and clients, respectively, simply ignore the—for them—unknown packet parts. It is also possible that one DHCP server provides the network information while another one called Proxy DHCP is responsible for the PXE part within a network. This enables the usage of PXE in networks which may not have the ability to reconfigure their DHCP servers. For instance, mid-sized enterprises often use specific appliances as their DHCP servers which may not be sophisticated enough to offer the configuration depth necessary for PXE. Such companies therefore have to use the fallback solution with Proxy DHCPs.

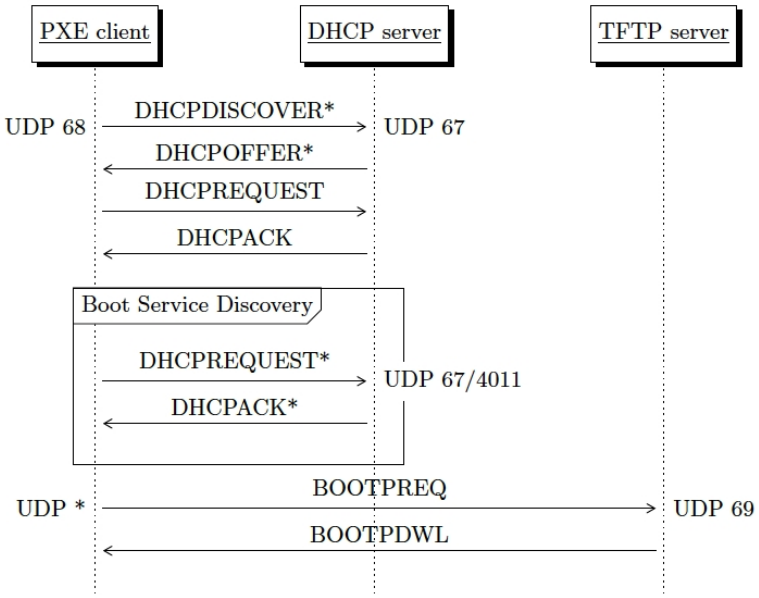


FIGURE 6: THE PXE PROCESS CONTAINS A DHCP HANDSHAKE AND A TFTP DOWNLOAD [7].

The protocol process is illustrated in figure 2. Once a client tries to perform a network boot he broadcasts an extended DHCPDISCOVER packet requesting network settings and signaling the PXE enabled DHCP server that he wants to perform a PXE boot. The server responds with an extended DHCPOFFER containing the PXE settings among other things. The clients stops the boot process if the required parameters are missing[5]. The client requests one of the provided IP addresses using a DHCPREQUEST. The server

[5] *For instance, DHCP servers which do not support the PXE protocol are lacking the needed parameters.*

ISSN 2052-8604

receipts the request either by agreeing (DHCPACK) or denying (DHCPNACK) the client's choice. After having received a valid IP address, further communication will be done using unicast instead of broadcast in most cases. If not already happened, a client has to send another extended DHCPREQUEST to receive the location of a boot server offering a NBP. This procedure is called *Boot Service Discovery* and can be done using multicast, broadcast or unicast (prioritised in this order), but requires that the client already has a valid IP address. However, the server has contingently restricted which send method is valid within the network. At this point, the client knows all needed parameters to load the file image from a TFTP server into its RAM and to finally execute the optionally validated bootable program.

Attack Vectors

The following chapter depicts the actual attack vectors against PXE systems. The attacks were tested against a working PXE environment using the ISC[6] DHCP server, PXELINUX and a Memtest86+ boot image. The three actors, PXE server, PXE client and the attacker are connected via an unmanaged switch. This software configuration comes up quite commonly when searching on the internet for instructions on how to setup a PXE environment. It also can be found slightly altered in enterprise networks but with the shown attacks working nonetheless.

The DHCP server recognises the PXE client on the basis of its MAC address. Thus, the DHCP server assigns a special IP to the client while it serves dynamic IPs for all other clients. The server also points to its own IP as the "next-server" and refers to the PXELINUX (pxelinux.0) file as Network Bootstrap Program. The contents of the TFTP directory for the exemplified configuration are as follows. A Memtest86+ boot image, a NBP and a configuration directory named pxelinux.cfg with the default configuration file provided by the Ubuntu Linux 14.04 distribution. Furthermore, it is possible to use client dependent configuration files by placing them in the named directory. A configuration is then applied if either a client's MAC address or its IP address[7] matches a filename within the pxelinux.cfg directory.

[6] *ISC or the Internet Systems Consortium is a non-profit organisation maintaining several core protocol implementations and applications, necessary for a self- organising internet.*

[7] *IP addresses have to be hex encoded. For instance,* `192.168.0.1` *corresponds to* `C0A80001`.

ISSN 2052-8604

The remainder of this chapter will go into further detail about the attack vectors found. It will also describe the course of action taken to exploit the vulnerabilities. The goal of most of the following attacks is to execute arbitrary code on a target system by manipulating either the boot image or the NBP. For the sake of simplicity, we simulate a successful exploit by replacing the mentioned Memtest86+ reference image with another image of the same software. Both images only differ in the version of Memtest86+: the reference file image provides the software in version 5.01 while the attacker file image uses the older version 4.20. By comparing the version numbers after the PXE boot one can easily see whether or not an attack worked in the intended way.

The actual implications of malicious code execution before the actual boot for the victim are dependent on the type of the machine in use. For standard clients without disk encryption, consequences can be severe as the attacker is able to read, manipulate and delete any data on the disk. This can be used to persistently implant viruses in the actual Operating System (OS) of the client or to simply steal data from it such as saved credentials or critical documents. Thin Clients that are started via PXE show similar weaknesses, although, due to lack of nonvolatile storage, potentially embedded malware is not persistently compromising such devices. While it is not possible to persistently implant malware into their operating systems, they can be repeatedly manipulated to load malicious code for the working OS.

Rogue DHCP Takeover

Due to the lack of writable memory on the PXE client, it is not possible for the client to discern between trustworthy and malicious PXE servers. Hence, an attacker can create his own PXE environment in the client's subnet and force the client to choose his installation over the real PXE server. This process is transparent for the client, i.e. the client cannot reenact which one he is intended to connect to.

Contrary to many other attacks where the attacker must be situated between server and victim, the only requirement for this attack is that broadcasted DHCP packets from the attacker must be able to reach the victim. As for most attacks, the goal is to execute arbitrary code on the victim's machine by forcing him to boot an image chosen by the attacker. For that purpose the attacker basically sets up a completely separate PXE environment on a system under his control and tries to redirect all boot requests to his server. Because the client doesn't know which server is the correct one, he simply connects to the one whose DHCPOFFER packet reaches him first (as long as it supplies

ISSN 2052-8604

the parameters necessary for PXE). This behaviour creates a race condition between all active DHCP servers which are able to serve the booting client within a network.

Table 1 shows the chronologically ordered sequence of DHCP packets sent when booting in PXE mode while a rogue DHCP server is present in the network. One can see that, in the exemplified case, the rogue DHCP server wins the race condition. Therefore the DHCPREQUEST packet of the client refers to the rogue server. Because the network's actual DHCP server is configured to be authoritative, it answers with a DHCPNAK to this request. The client, however, ignores this and finishes its DHCP communication after having received the DHCPACK from the rogue server. As intended by the attacker this leads to the client using the rogue DHCP's parameters, which in turn persuades him to boot the attacker image (Memtest86+ v4.20).

Source	Destination	Packet type
Client	Any DHCP server	DHCPDISCOVER
Rogue DHCP server	Client	DHCPOFFER
Network DHCP server	Client	DHCPOFFER
Client	Selected (rogue) DHCP server	DHCPREQUEST
Rogue DHCP server	Client	DHCPACK
Network DHCP server	Client	DHCPNAK

TABLE 6: PXE BOOT INITIATION WITH A ROGUE DHCP SERVER INVOLVED SHOWN IN CHRONOLOGICAL ORDER.

The approach outlined above relies on the victory over the *race condition* between attacker and network DHCP server. Thus, the success rate can be improved by ensuring that the malicious DHCPOFFER packet reaches the client first. There are different methods to achieve this. The most obvious way is to simply hinder the original DHCP by, for instance, employing a Denial-of-Service (DoS) attack against it. Another measure is setting up the physical network environment in a way that the DHCP server is either unreachable or connected by a longer or slower route.

Man-in-the-Middle TFTP Hijacking

PXE is lacking encryption and authentication mechanisms at all stages. Therefore the process is very susceptible to Man-in-the-Middle (MitM) attacks of any kind. This section covers their use to hijack the role of the TFTP server and thereby distribute malicious images.

ISSN 2052-8604

The general scenario and prerequisites for this attack are the same as in section 3.1. This attack can be utilised in settings where DHCP traffic is blocked so that the attack from the previous section won't work anymore. This can happen, for example, due to network management that restricts DHCPOFFER packets to one port on the network's switch or by generally restricting broadcast traffic in a network.

	Source	Destination	Protocol	Packet type
1	Client	DHCP server	DHCP	DHCPREQUEST
2	DHCP server	Client	DHCP	DHCPACK
3	Client	Broadcast	ARP	Who has TFTP-IP?
4a	Attacker	Client	ARP	TFTP-IP is at Attacker-MAC
4b	TFTP server	Client	ARP	TFTP-IP is at TFTP-MAC
5	Client	Attacker	TFTP	TFTP RRQ
6	Attacker	Client	TFTP	TFTP ACK
	...	...	...	...

TABLE 7. THE PACKET SEQUENCE IN A TFTP HIJACKED PXE BOOT PROCESS SHOWN IN CHRONOLOGICAL ORDER (EXCERPT).

Because the attacker is not able to manipulate the official TFTP server, he has to impersonate the real one. Because TFTP is not a broadcast protocol like DHCP, employing a MitM attack becomes a necessity to intercept, manipulate and inject traffic into the communication. Table 2 shows an excerpt of the packets sent in a PXE boot process in chronological order including the ones sent by an attacker. The relevant parts are the ARP packets between the end of the DHCP and the start of the TFTP communication: the clients broadcasts an ARP request (3) whereupon he receives two responses (4a and 4b) in a compromised setup. These packets are necessary because the client does not know the MAC address of the TFTP server yet. The success of the attack can be seen through the attacker's ARP reply coming before the TFTP server's reply as well as the client sending his first TFTP packet to the attacker.

ISSN 2052-8604

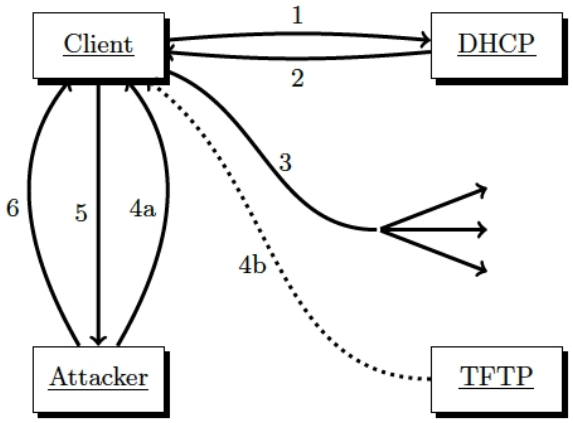


FIGURE 7: THE HIJACKED PXE BOOT PROCESS ONLY WORKS IF THE ATTACKER CAN DELIVER HIS PACKET PRIOR TO THE TFTP'S PACKET.

Usually, ARP cache poisoning is used for a MitM attack wherein the attacker relies on the target systems caching his spoofed ARP packets and therefore sending their traffic to the attacker. However, PXE clients do not implement an ARP cache, but rather use the very first ARP reply that reaches them. Similar to the “Rogue DHCP Takeover” attack, this creates a race condition between the attacker and the real TFTP server. However, this behavior is not represented by the common publicly available ARP spoofing tools. Therefore we prototyped a script specifically for this use case.

```
   Input  : IP adress (PXE client and server): string[]
   Input  : MAC adress (PXE client): string[]
   Input  : Spoofing duration and rate: int[]
 1 Turn on IP forwarding
 2 Set manual ARP entries for faster spoofing
 3 Set up iptables to redirect packets for the PXE server to itself
 4 Start tshark with filter on TFTP and DHCPDISCOVER packets
 5 while read line from tshark do
 6 |   if line is DHCPDISCOVER then
 7 |   |   Send forged ARP replies with own MAC to the client
 8 |   else if line is TFTP then
 9 |   |   Stop sending ARP replies; wait for keypress; reset network settings; exit
10 |   end
11 end
```

ALGORITHM 1: PXE ARP SPOOFING SCRIPT

ISSN 2052-8604

Algorithm 1 shows the concept of the script that is used for the spoofing attack. Necessary inputs are the IPs and the MAC addresses of PXE client and server, respectively, as well as the separate spoofing rate and duration values for them. After some setup and optimisation the script starts a tshark[8] session for packets sent from the PXE client. The tool reviews each packet that is read by tshark if it's either a DHCPDISCOVER packet with PXE parameters or a TFTP packet. While the former indicates a new PXE boot request, the latter shows that the spoofing worked since the attacker now receives the client's TFTP communication. On receipt of a DHCPDISCOVER packet the script starts to send forged ARP reply packets to persuade the client of the attacker having the original TFTP server's IP address. Rate and duration of this transmission is defined by the respective input variables. As soon as the first TFTP packet arrives, the script just waits for a keypress to reset network settings and exits.

Additionally to the MitM attack, the attacker has to set up his own TFTP server. The NBP that is served must correspond to the filename supplied by the DHCP server, otherwise the client is unable to successfully download it. When these requirements are met, the attacker has to start the spoofing script and it will automatically intercept the PXE boot request and redirect the TFTP communication to the machine the script is run on.

Other attack vectors

Aside from the two major attacks that have been covered in the previous sections, there are more security implications to consider with the most significant ones being shortly presented in this section.

PXE systems are vulnerable to non-overload-based DoS attacks due to the use of UDP as transport layer protocol. The attack from section 3.1, for instance, can be slightly altered so that the packets sent by the rogue DHCP contain invalid PXE parameters and therefore preventing the client from booting. Another possibility would be the injection of UDP packets with random data into the TFTP file transfer communication. This would corrupt the boot image and results in a crashing client.

Another consideration is the general lack of security features in TFTP. Any client in a network that would be theoretically able to boot via PXE is also able to access and

[8] *tshark is the command line version of the well-known packet analyser wireshark available at https://www.wireshark.org/.*

ISSN 2052-8604

download any NBP or boot image on the TFTP server. This can lead to security problems when critical data is included in such a boot image. For example, such a file image can contain certificates for authentication at a company server.

As a last note, it is quite easy to inject packets into the TFTP file transmission communication. An attacker could simply replace data in configuration files or boot images as long as he knows their structure (which he can easily achieve due to the lack of authentication in TFTP). Replacing whole images is usually harder this way, though, since every TFTP packet sent by the attacker has to win a race condition to be accepted by the client. Therefore the probability of getting a corrupt image from two different sources is very high.

Conclusion

This work has shown that PXE is vulnerable to a number of attacks resulting in potentially comprised security within SCADA like IT infrastructures. We exploited missing security features like the absence of authentication and encryption mechanisms to boot our own (potentially compromised) system image by only having physical access to a network. Depending on the type of client, these security problems can have differing consequences for the booting clients. While a thin client is only untrustworthy up to the next secure boot, a stateful client is potentially insecure until the point of reinstallation. To prevent these type of attacks on SCADA Systems, an approach could be to secure the network physically. Any unauthorized entities should not gain entry into the network [6]. Physical access could be limited by preventing access to network plug sockets, on the wall and client side, for example.

Unfortunately there are very little security safeguards built into the PXE specification and neither DHCP nor TFTP are more robust concerning security. The *Rogue-DHCP Takeover* attack is constructive if an attacker can ensure that his DHCP is used during the boot process because the victim has no possibility to verify a DHCP server's identity. *TFTP Hijacking* is based on the well know Man-in-the-Middle principle and is targeted on the lack of encryption during the PXE communication, thus an attacker can foist a different file image on a client. An offender can use the exemplified attacks to transparently introduce malware in a company. Hence, we conclude that the usage of the by definition unsafe PXE protocol within a company needs additional and individual provisions to circumvent the shown procedures. The protection of PXE systems is best done via general network management. Restricting DHCP traffic and broadcasts,

ISSN 2052-8604

preventing ARP spoofing attacks and controlling the physical network access are all valuable measures to ensure a secure PXE environment. As an additional security layer it is advisable to use disk encryption to prevent data theft as well as persistent alteration of data even in case of an insecure boot process.

Further Steps

From a security perspective it would be interesting if the different implementations of PXE such as Microsoft Windows Deployment Services (WDS) or iPXE possess the same weaknesses as the original PXE implementation. However, the usual way of booting up these systems is by loading a special bootloader over the network card's PXE stack. Because this first stack is the same one we examined in this paper – save for replacement of the NIC's firmware – all the attacks would work accordingly. Also, because these implementations are more feature rich, a new set of weaknesses could be potentially found, possibly with severe consequences for devices within a network. WDS could, for example, open attack vectors on products like Microsoft System Center Configuration Manager (SCCM). With such a system management solution being a critical part of many infrastructures this would be an aspect worth exploring.

A different direction, being slightly more application-focused, is the development of a system with the optimal combination of comfort of centralised software/operating system distribution and protection against offenders of all kinds as well. One possibility is piecing together such a system from existing protocols, systems and applications. However this may not necessarily work and it might be necessary to develop a new protocol for software distribution which meets requirements such as guaranteeing integrity, creating automated and dynamic workflows or authenticating clients against a central instance once they demand a PXE image. If applicable, existing and well-known protocols like SFTP or HTTPS can be a good starting point.

Acknowledgments

The research leading to these results was supported by "Bavarian State Ministry of Education, Science and the Arts" as part of the FORSEC research association (http://www.bayforsec.de/).

References

[1] Rolf Carlson. Sandia SCADA program high-security SCADA LDRD final

ISSN 2052-8604

report. April 2002.

[2] Thomas M. Chen. Survey of cyber security issues in smart grids. In *SPIE Defense, Security, and Sensing*, volume 7709. International Society for Optics and Photonics, April 2010.

[3] T. Cruz, P. Simoes, F. Bastos, and E. Monteiro. Integration of pxe-based desktop solutions into broadband access networks. In *Network and Service Management (CNSM), 2010 International Conference on*, pages 182–189, October 2010.

[4] R. Droms. Dynamic Host Configuration Protocol. RFC 2131, March 1997.

[5] Xi Fang, Satyajayant Misra, Guoliang Xue, and Dejun Yang. Smart grid – the new and improved power grid: A survey. *Communications Surveys Tutorials, IEEE*, 14(4):944–980, 2012.

[6] Vinay M. Igure, Sean A. Laughter, and Ronald D. Williams. Security issues in SCADA networks. *Computers & Security*, 25(7):498–506, 2006.

[7] Intel Corporation. Preboot Execution Environment (PXE) specification. Website, September 1999. Visited 2014-08-28.

[8] M. Johnston and S. Venaas. Dynamic host configuration protocol DHCP options for the intel preboot execution environment PXE. RFC 4578, November 2006.

[9] Anthony R. Metke and Randy L. Ekl. Security technology for smart grid networks. *Smart Grid, IEEE Transactions on*, 1(1):99–107, October 2010.

[10] Allen Risley, Jeff Roberts, and Peter Ladow. Electronic security of real-time protection and SCADA communications. In *Proc. of the 5th Annual Western Power Delivery Automation Conference*, pages 1–3, 2003.

[11] K. Sollins. The TFTP protocol (revision 2). RFC 1350, October 1992.

[12] Jason Stamp, Phil Campbell, Jennifer DePoy, John Dillinger, and William Young. Sustainable security for infrastructure SCADA. *Sandia National Laboratories, Albuquerque, New Mexico (www.sandia.gov/scada/documents/SustainableSecurity.pdf)*, 2003.

ISSN 2052-8604

Trustworthy Systems Design using Semantic Risk Modelling

Ajay Chakravarthy, Stefanie Wiegand, Xiaoyu Chen, Bassem Nasser, and Mike Surridge
University of Southampton IT Innovation Centre, Gamma House, Enterprise Road, Southampton, UK
{ajc,sw,wxc,bmn,ms}@it-innovation.soton.ac.uk

Abstract

In this paper, we present tools and methods for the design of trustworthy and secure ICT systems, and an intuitive mechanism through which risk management can be performed on the composed system. The work carried out is part of the OPTET project, which addresses the management of trust and trustworthiness in socio-technical ICT systems. In our approach, the system trustworthiness can be achieved and maintained by making it resilient to the risks that may compromise its functions. This should be addressed during design-time as well as runtime. This paper describes our semantic risk modelling approach which is suitable for dynamic and evolving systems. Further, we focus on the usefulness and usability of our modelling approach through the **System Composer**, which is an intuitive and easy to use graphical user interface used by system designers to compose trustworthy systems and perform risk analysis without having to deal with the complexity of the underlying semantic risk models.

Introduction

Humans, organisations, and their information systems interact and influence on each other as part of a Socio-Technical System (STS) [26]. These systems, nowadays, are distributed, connected, and communicating via the Internet in order to support and enable digital business processes, and provide benefits for the economy and society. An example is a tele-healthcare system involving wearable telemetry devices, which ultimately support interaction between care workers and patients. Trustworthiness of such Internet-based software systems, apps, services and platform is critical for their use and acceptance by organisations and end-users.

Untrustworthy systems are likely to behave in ways that users don't expect and find objectionable (e.g. intermittent error, leak of private information, unacceptable delay, etc.). Trustworthiness is a property of ICT systems that make them worthy of its users trust. We argue that this can be achieved and maintained by making the system resilient to risks that may compromise its functions. This requires a risk management approach to be integrated early in the design phase and consistently applied during runtime.

Our approach is defined in terms of the automated and systematic identification of risks to the assets within that system (human and technological) as well as their knock-on consequences, countermeasures to mitigate these risks, and a runtime system for monitoring and detection of these risks. The identified threats depend not only on what assets are involved but also on how they are related to each other. Addition or removal of an asset, or changing the composition of existing assets will result in different threats identified. This goes beyond the current risk management methodologies [13][6][4][31] in terms of usability and applicability to dynamic and adaptive multi-stakeholder ICT systems.

In this paper we focus on the design-time aspects by describing the models for system asset composition and threat derivation for dynamic ICT socio-technical systems and the mitigation strategies to deal with these threats using semantic ontologies. We follow a layered ontology modelling approach in OWL. We divide the models into a core ontology which provides a basic, high level structure for modeling risks in an STS. The generic ontology builds on this by modeling system independent classes of assets, threats, controls and behaviours. The design-time trustworthiness model specializes these classes to model the system specific classes of assets and threats. Finally a concrete model of OWL instances which captures the instantaneous system composition and status.

We describe in this paper how these semantic models are made useful to system designers (who are not expected to have semantic or deep security expertise) using an easy to use and intuitive graphical interface: the **System Composer.** The composer hides away the complexity of the underlying semantic models and presents only the most relevant information in a simple, point-and-click user interface. Abstract models of systems can be easily created by dragging and dropping assets on a canvas and defining their relationships. The System Composer then automatically creates all the system specific threats (risks are defined in terms of threats) for each asset using generic threat

ISSN 2052-8604

information encoded in the underlying generic semantic model which is domain independent.

The rest of the paper is organised as follows. Section 2 presents related work on existing state-of-the-art on semantic risk modelling and the graphical tools which currently exist for this purpose. Section 3 presents the semantic risk modelling approach we developed for dynamic and adaptive ICT socio-technical systems. Section 4 describes the System Composer tool in detail. Finally we conclude the paper in Section 5

Related Work

Security and Threat Modelling Tools

Threat modelling is a practice for identifying and predicting security threats of a given system, and then defining countermeasures to prevent or mitigate the effects of identified threats to the system. A threat model can then be handed over to the system design team as a guidance to optimally mitigate application risk. According to Shostack,A [25], the threat modelling process can be best presented using a four-stage framework (Figure 1) that was designed to align with software development lifecycle and operational deployment. The key steps are: modelling the target system; finding threats based on the system model; defining mitigation tactics and technologies to address threats found; validating the effectiveness of defined mitigations by integrating threat model to software and system test processes.

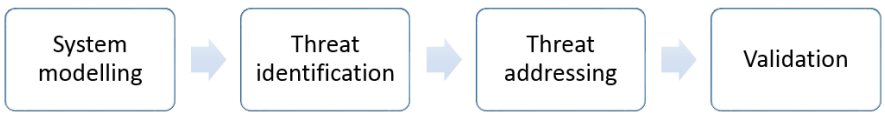


FIGURE 8: FOUR-STAGE THREAT MODELLING FRAMEWORK

There are three common structured approaches to model threats, which can be characterized as asset-centric, attacker-centric or software-centric. The asset-centric approach starts from identifying the assets or things of values in a STS. For each identified asset, an attack tree should be generated to show how the value of the asset might be compromised. Countermeasures can then be identified to reduce the risk of compromise from these identified threats. The main problem with this approach is that it

ISSN 2052-8604

involves a manual analysis, and the identification of threats depends on the expertise of the analyst, who may therefore overlook some important threats. The use of checklists to prompt the analyst provides only a partial solution to this.

The attacker-centric approach takes the reverse approach. Here one starts by identifying types of attackers and the ways they might seek to attack the system. Then one considers whether the system may be vulnerable to the attacks. Given its human-centered nature, the attacker-centric approach are more likely to find human threats, e.g. from social engineering attacks. However, the process still depends on a manual analysis by an expert. The analyst may also have to work harder to decide how threats relate to their particular system than in an asset-centric approach. Both asset-centric and attacker-centric approaches are widely used by security professionals to analyse threats of a given system.

In contrast, software-centric approach intended to integrate the threat modelling process into software development lifecycle. A good example is Microsoft's Secure Development Lifecycle (SDL) framework [9].

Software-centric approaches represent a threat model using software architecture diagrams, for example, Data-Flow Diagrams (DFD), use case diagrams or component diagrams. Such approaches are mainly used by a software development team to develop secure software by addressing identified security requirements and countering threats through the design and implementation of their software. This approach still depends on security expertise, but has the advantage that checklists (e.g. types of threats) are already related to the system implementation. However, it doesn't cover broader classes of threats, e.g. involving humans or hardware, etc.

In all these approaches, a key step is identification of threats to the system. There are many ways to do this, and numerous tools designed to help reduce human error in different situations, most notably when using a software-centric approach to threat modelling. For example, STRIDE [29] is a framework adopted by Microsoft SDL. It was designed to help developers by identifying types of attacks that tend to affect software. STRIDE stands for six categories of threats (spoofing, tempering, repudiation, denial of service, information disclosure, and elevation of privilege). These categories and the countermeasures defined in STRIDE provide guidance for developers who are new to security as well as acting as reference material for security professionals.

ISSN 2052-8604

Attack trees are another common method used by security professionals to analyse system threats in greater depth to find out other possibilities go beyond those common threats (e.g. those defined in the STRIDE). An attack tree [25] is a multi-levelled diagram consisting of one root representing the attacking target, and leaves/children representing the condition. From bottom up, child nodes are must be satisfied to make the direct parent node true, when the root is satisfied the attack is complete. The attack elicitation process requires iteration over each node in the tree and consider if that issue impacts the target system. When creating an attack tree, one need to decide on a presentation and select a root node. With that root node, STRIDE or other literature review (e.g. ISO 27001 [14]) can be used to find threats to add to nodes.

Although STRIDE has been adopted as a standard method of threat analysis and elicitation, it is arguably too high level and can be replaced with a more detailed commonly occurring attacks in different context. This motivated the emergence of attack libraries, such as (MITRE's Command Attack Pattern Enumeration and Classification) CAPEC [1] and (the Open Web Application Security Project) OWASP [20].

There are now many such threat modelling tools (Table 1) available in the market or offered as open source. These tools can be used to help threat modelling in many ways. A tool can help engagement in the threat modelling steps and provide assistance in performing those steps. These tools, such as TRIKE [22] SecurTree [11], and SeaMonster [18] are mainly used by security professionals who are familiar with the threat modelling process but requiring assistance in assurance of legibility and completeness. Both ThreatModeller [30] and Microsoft's SDL threat modelling tool adopted software-centric approach and provide an integrated solution for developer to generate threat model based on software architecture diagrams.

TABLE 8: LIST OF THREAT MODELLING TOOLS

Tool	Approach	Threat finding method	Target users
TRIKE	Asset-centric	Attack Tree	Security professionals

ISSN 2052-8604

SeaMonster	Software-centric	Attack Tree	Security professionals
Secur|Tree	Attacker-centric	Attack Libraries	Security professionals
ThreatModeller	Software-centric	Attack libraries	Developer
Microsoft's SDL threat modelling tool	Software-centric	STRIDE	Developer

However existing tools focus on either assisting threat modelling by security professionals, or integrating modelling of software-related threats into a software development process. Organisations still depend heavily on security consultants to deliver custom threat models, and deliver reports to those who implement and operate the analysed system. Organisations that use software-centric tools like Microsoft's SDL threat modelling tools may be more likely to overlook threats beyond the scope of STRIDE (e.g. human-centred attacks) unless they also involve security professionals in the loop.

Our approach aims to use automation to better integrate security expertise into a system design and engineering methodology, and to reduce human error. In order to do this, we chose to use machine understandable models, based on semantic reasoning technology. By using semantic technologies and well-defined semantic modelling stacks, our approach allows security experts to analyse case-by-case threats and to encode threat generation rules in a both human-readable and machine-understandable manner. Their expertise can then be used as a basis for automated analysis of a specific system, based on a semantic model of the system architecture created by its designers.

ISSN 2052-8604

Semantic Risk Modelling and Machine Reasoning

We are not the first to consider using machine understandable semantic models for security analysis, although few have attempted a similar level of automated reasoning. Work by Kim et al. [17] uses an extension of the Secure Tropos language to support the modelling of security risks. The domain model is mainly structured around three groups of concepts: asset-related, risk-related and risk-treatment related. Further security criteria for each of these assets are identified in terms of confidentiality, integrity and availability. This work is an extension of the work on Secure Tropos, and includes the development of syntactic, semantic and methodological extensions that would support security risks and their counter measures. Our threat representation is diagrammatic in nature and is used to present abstract syntax elements for risk modelling and the rules on how these can be combined together.

The domain model by Hogganvik et al.[8] is similar to the core ontology model we use (consisting of Threat, Asset and Control). They present work on a graphical approach to identify, explain and document security threats and risk scenarios. A graphical notation was developed to perform the five phases needed for security analysis 1. Context establishment, 2. Risk identification, 3. Risk estimation, 4. Risk evaluation and 5. Treatment identification. Diagrams are created during each of these steps (similar to UML models) under the guidance of a domain expert. We have followed a similar approach for the identification of threats and their mitigation strategies. However, the work described in [8] does not go beyond the modelling phase (diagrammatic modelling). The novelty of our system modelling approach is that we use an abstract modelling approach based on the Web Ontology Language (OWL) to address the challenge faced in adaptive systems (where the composition of the system is not known in advance). This allows our models to be much more expressive (due to expressive nature of OWL syntax) and encoded in a way such that existing rule based languages (e.g. SPIN) and semantic reasoners (e.g. TopSPIN) can be used with these OWL ontologies for the automatic identification and mitigation of threats.

An approach to address the problem of risk management when the composition of dynamic multi-stakeholder systems is unknown is to use machine reasoning to analyse risks, so this can be done rapidly whenever the system composition changes. There is an existing body of research into how one might create semantic models of a system to support such an automated analysis, though with the motive of capturing security standards and expertise so tools can be developed to support non-experts. A useful

ISSN 2052-8604

overview is provided by [1]. For example, the NRL Security Ontology [16] provides a way to describe the security properties of Web Services, which was later used as a starting point for a Web Service vulnerabilities ontology [32]. The Ontology of Information Security [10] by Herzog et al describes a system in terms of assets, vulnerabilities and threats, so making the link to system risks (via threat models). However, this ontology uses N-ary relationships making it hard to deduce security properties via machine reasoning.

The Security Ontology from Secure Business Austria (SBA) [5] uses only conventional RDF relationships, and captures security threats and controls from the German IT Grundschutz Manual [15], so providing a way to model systems with common threats and control strategies. The SBA approach goes a long way towards the goal of capturing security expertise in a form that can be reused (with supporting tools) by non-experts. However, this ontology describes only deployed systems and security controls, and cannot be used to create a design-time model for a dynamically composed system whose concrete composition is not known at design time. The SBA approach also makes extensive use of OWL instances, which makes it hard to cater for multi-stakeholder systems where it is often necessary to attach different properties to the same threat depending on which stakeholder or sub-domain is targeted. Our core and generic ontology models use many of the ideas from the SBA approach [5], but uses OWL classes (as opposed to instances) to model security concepts, enable design-time reasoning about dynamically composed systems. At run time, instances of these classes can be used to represent different parts of the system (e.g. different users, servers, etc.), allowing expertise captured at class level to be used in many different contexts in the real system. Our model also uses a simpler upper ontology, so that run-time reasoning can provide useful insights with a modest number of asserted facts which can be automatically derived from run-time system monitoring data. This is consistent with our goal of maximising the use of automation throughout the system lifecycle.

Semantic Risk Modelling Approach

Our approach is designed to address three main challenges:

1. System designers using a risk-based approach to capture security requirements often lack expertise in potential threats and countermeasures;
2. System designers may overlook threats they don't understand, or that they subjectively believe are not important for their particular system;
3. Information from design-time risk-based analysis is not yet utilised systematically at run-time.

ISSN 2052-8604

The last of these points is due partly to the fact that risk analysis still largely depends on humans with relevant expertise (even if assisted by a standardised procedure and check list). Run-time analysis is only useful if it can be done rapidly when systems change or their behaviour and status change. This often means risk-based analysis cannot be used at all when considering dynamically composed (e.g. Future Internet) systems. In the next sections, we present the details of the semantic model layers. Each layer is implemented using OWL for ontology construction and SPARQL + SPIN for rule encoding and semantic interference.

Core Model

The Core semantic model contains only the fundamental concepts, modelled using the following top level classes:

- Asset: anything of value in a socio-technical system;
- Threat: a situation or event that if active could undermine the value of an asset by altering its behaviour;
- Misbehaviour: a condition on asset behaviour that, if met, means the behaviour is unacceptable and the value of the asset is undermined;
- Control: a trustworthiness requirement that, if met by an asset, will block or mitigate a threat, enabling the asset to resist the threat and/or prevent any misbehaviour.

Relationships used in the core model are also of the most fundamental and stable variety, as shown in Figure 2.

ISSN 2052-8604

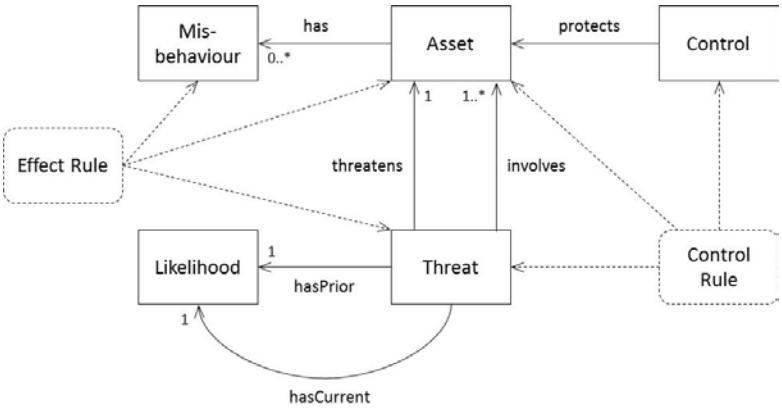


FIGURE 9: CORE TRUSTWORTHINESS MODEL

This model says that assets can misbehave, be threatened by or involved in threats, and protected by controls. Threats may become active and so cause asset misbehaviour. This is captured via two likelihood values (probabilities), one denoting the prior expectation (probability) that a threat will become active, and the other indicating the likelihood that it is currently active.

It is worth noting that the Control Rule and Effect Rule elements are not actually classes, but rules that refer to asset, threat, control and misbehaviour classes. We use OWL to encode the class structure and SPARQL + SPIN to encode all rule based information within our ontologies.

A Control Rule specifies which controls must be in place (i.e. which trustworthiness requirements must be met) for each involved asset in order to block or mitigate a given type of threat. The logic for a control rule is as follows: Suppose there is a threat T threatening a single asset A1 and involving other assets e.g. {A1, A2, A3}. If there are control measures (C1, …, Cn) such that C1 protects A1, C2 protects A2 and C3 protects A3 (in a pattern encoded in the rule) then the threat is considered blocked (if the controls are proactive and prevent the threat arising) or mitigated (if the controls are reactive and mitigate the effects if the threat does arise).

An Effect Rule describes the relationship between threats and asset misbehaviour. There are two types of Effect Rule:

ISSN 2052-8604

- Induced Effect rules specify what asset misbehaviour is caused by an active threat (usually in terms of a probability that threat activity would lead to a specific asset misbehaviour).
- Secondary Effect rules specify how a threat to one asset may be caused by misbehaviour in this or other involved assets, i.e. how misbehaviour can cause threat activity;

There is also a special 'null' induced effect rule specifying how likely it is that asset misbehaviour will be detected without being caused by any threat. This covers either spontaneous misbehaviour which is then detected, or false positive misbehaviour detection.

The Generic Model

The generic trustworthiness model specialises the core model so that it can easily be used to encode risk knowledge and apply it in a multi-stakeholder, agile service oriented systems possibly restricted to specific application domains. The main features of the generic model are:

- subclasses of core:Asset representing services offered by the primary stakeholder, customers who use those services, and resources composed in order to deliver those services;
- subclasses of core:Threat representing the different ways that the value of these assets could potentially be compromised;
- subclasses of core:Control representing security or system management mechanisms that could be used to block these threats or mitigate their effects;
- subclasses of core:Misbehaviour representing specific types of adverse behaviour that could be induced by active threats.

Asset Model

The main asset subclasses are those that will be sub-classed and related to each other explicitly by the system designer to express the structure of their system:

- LogicalAsset: used directly by the designer to describe system processes – at this stage we do not distinguish between software and other processes;

ISSN 2052-8604

- Human: an individual user of the system – a type of Stakeholder with independent volition acting (or trying to act) in their own interests;
- Organisation: a (commercial or non-commercial) business – a type of Stakeholder that is made up of many people acting according to some organisational objectives and control;
- Host: a collection of physical apparatus including one or more connected ICT systems that support logical processes;
- WiredNetwork: a network that uses physical (wired) connections to transmit data between Hosts;
- WirelessNetwork: a network that uses radio signals to transmit data between Hosts;
- CellularNetwork: a network in which radio signals are used to transmit data between Hosts and the network, but wired connections are used within the network.

Figure 3 shows these classes (shaded) and their relationships:

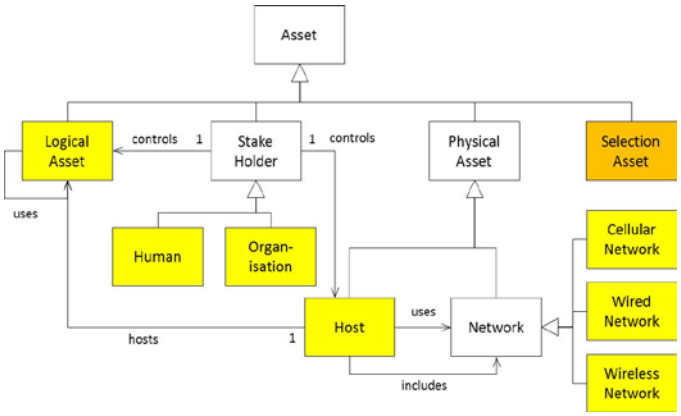


FIGURE 10: GENERIC ASSET CONFIGURATION.

In our earlier work [28] the asset subclasses were defined such that system designers had to classify their system processes into different architectural layers (client, service, consumer etc.). In the new approach, this classification will be inferred from the 'uses' relationships between two LogicalAssets, which indicate that one of the associated processes initiates some activity in the other.

ISSN 2052-8604

We do want the designer to distinguish between Human and Organisation stakeholders because Organisations are subject to some threats (e.g. from malicious insiders) which may make them act against their own interests. These don't normally affect Human users – a Human may or may not be malicious, but their actions usually reflect their true motivations. Note that stakeholders control both logical assets (i.e. system processes) and the Hosts that support these processes.

A Host may represent a single ICT system (e.g. a phone or a server), or an entire organisational infrastructure (including buildings and people as well as ICT systems and internal networks).

We distinguish between two types of relations between Host and Networks. When a Host uses a Network, it means that the Host communicates with other Hosts over that network. When a Host includes a Network, it means that the Host communicates internally over that network, and therefore the Host includes more than one connected ICT systems. It is worth modelling the internal communication networks separately because in some cases (e.g. if the network is wireless), it may be possible to compromise the Host by attacking the network from the outside (e.g. by jamming it).

The classification of physical assets into Hosts and Networks allows modelling of threats that use networks to attack hosts and vice versa. However, intrinsic threats against Networks differ depending on the type of network – e.g. a WirelessNetwork can be jammed by electromagnetic interference, but a WiredNetwork is largely immune to such an attack, and a CellularNetwork can only be jammed locally. For this reason, the system designer should subclass the appropriate type of network when modelling a network in their system.

The SelectionAsset class represents a special case, indicated by a different shading in Figure 3. Some types of SelectionAsset are used to model the options for communicating between hosts over networks. This is a 'fine grained' detail that needs not be asserted by the system designer as it can be inferred from the other 'main' asset classes and relationships (i.e. if two hosts use two networks to communicate with each other, then these two networks are considered to be part of a network pool which is a SelectionAsset). Other types of SelectionAsset are used to model the options for composing LogicalAssets (e.g. backup assets). These can't be deduced from the rest of

ISSN 2052-8604

the structure, so their subclasses and relationships have to be asserted to fully define a composition.

The model is designed to reduce the complexity of system design by automatically creating necessary asset types inferred from the relationships between system-specific subclasses of the main assets show in Figure 3. This includes new class assertions as well as asset classification.

Supplementary class inference

The above classes are sufficient to describe a system in terms of the components from which it is composed. However, this is not enough to properly represent the action of certain types of threats. For example, in the communication between Hosts over Networks, some types of attacks (e.g. packet flooding attacks) can be launched against a specific Host over a specific Network. The attack will effectively disable any processes running on that Host that need to communicate over that Network. However, it won't disrupt communication with other Hosts over the same Network or with the same Host over any other networks it may be connected to.

To fully represent this type of threat, it is necessary to add a further class to represent the real focus of such an attack, which is actually the interface between the Host and the Network. Figure 4 shows how this is related to the corresponding Host and Network:

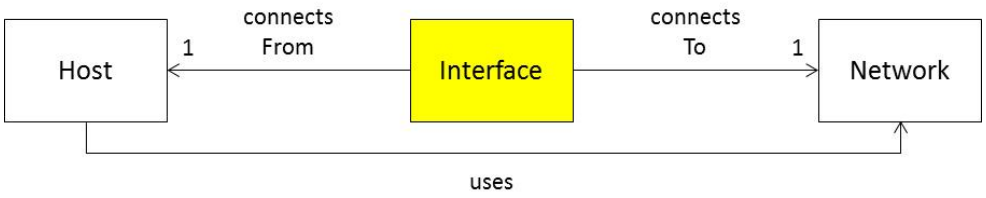


FIGURE 11: CLASS ASSERTION PATTERN FOR INTERFACE ASSET TYPES.

Figure 4 represents the relationships that should exist between the Interface, Host and Network. It says that whenever we have a Host that uses a Network, there should be exactly one Interface related to the Host and Network as shown. Note that the cardinality of these relationships is constrained – each Interface is related to exactly one Host and Network, which means we need a separate Interface for each Host and Network.

ISSN 2052-8604

The system model produced by a system designer need only specify one or more system-specific Host subclasses, with one or more system-specific Network subclasses for the possible networks over which they may be connected. The pattern from Figure 4 is then used to infer that wherever one of the system-specific Host subclasses uses one of the system-specific Network subclasses, there must be a system-specific Interface subclass representing interfaces between the relevant system-specific types of Host and Network. Table 2 shows the SPIN (pseudo) encoding of the above pattern within the design-time trustworthiness ontology. For space considerations, the exhaustive SPIN code has been summarized to highlight the main logic.

TABLE 9: INTERFACE AND HOST CLASSIFIER EXAMPLES

```
Find all hosts that use or include a network

FOREACH of the found pairs DO

        Only consider  host/network combination
        that don't share a common interface yet

        Create a new interface to connect host and
        network

        Create the necessary restrictions to
        connect the newly created interface to the
        related assets

DONE
```

Asset classification patterns

Another aid which this modelling approach provides is automatic classification of asset types based on their relationships without having to explicitly classify them manually into their architectural layers (client, server, consumer, delegate, agent etc.). We need to know this classification because while some threats could affect all LogicalAssets (e.g. a bug in the software), others will only arise because of the role of the asset (e.g. denial of service attacks can be made on services but not really on clients).

ISSN 2052-8604

The simplest relationship relates to the interactions between logical processes, using a rule represented by Figure 5.

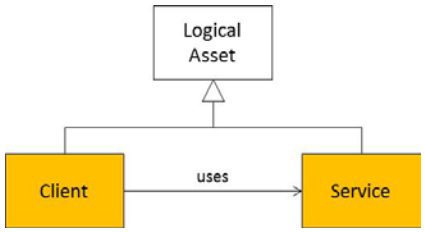


FIGURE 12: SIMPLE CLIENT-SERVICE CLASSIFICATION PATTERN.

The first issue to note is that Figure 5 defines a Client and a Service to be two LogicalAssets that have a mutual 'uses' relationship, i.e. the process implemented by the Client depends on some sub-process implemented by the Service. However, this doesn't by itself explain who controls the uses relationship. The simplest case is shown in Figure 6, which both defines a new classifier (the Consumer class), and introduces a new subclass of SelectionAsset (the ServicePool).

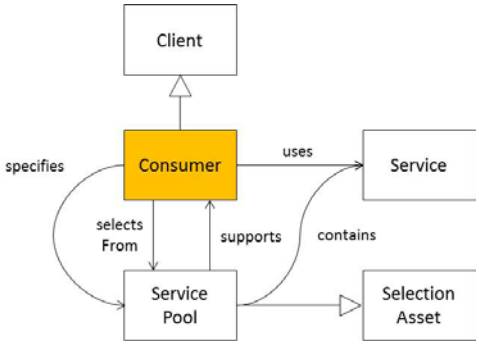


FIGURE 13: CONSUMER CLASSIFICATION PATTERN.

Like all SelectionAssets, the ServicePool represents the opportunity to choose how the system behaves, and it contains the candidate assets from which this choice is made. The existence of system-specific ServicePool subclasses is not inferred from the presence of other related system-specific asset classes. The system designer must explicitly define a system-specific subclass of ServicePool associated with each 'uses' relationships between system-specific Client and Service subclasses.

ISSN 2052-8604

The ServicePool always ‘supports’ the Client at one end of the ‘uses’ relationship, and ‘contains’ candidate LogicalAssets that could be at the other end of the relationship. This is how we know which ‘uses’ relationship it applies to. There should only be one ServicePool per ‘uses’ relationship, although if the Client uses multiple Services there would be multiple ServicePools supporting it, and if the Service is used by multiple Clients there would be multiple ServicePools containing that type of Service.

The other two relationships of a ServicePool describe how the Client-Service relationship is managed. Figure 6 represents the simplest case, in which the Client defines which candidate Services are in the pool and selects which one it will use when it needs the Service. This captures a typical late binding scenario often found in dynamic service oriented applications. All the decisions are made locally at the Client, which is classified as a Consumer of this Service. Note that even if there is only one candidate Service, a ServicePool should still exist if the Consumer has the option of not using a Service at all. Threats against Consumers (as opposed to Clients) should take account of the fact that the Client has a choice of Services, e.g. by specifying redundancy within the ServicePool as a control requirement if that would help to mitigate the threat.

Two other similar patterns are encoded in our generic ontology, covering situations where a logical asset is both a Client and a Service. The Client in this case uses the Service on behalf of its own Client, acting as an intermediary. In one case, the intermediary selects a Service from the ServicePool (as in Figure 6), but the members of the ServicePool are specified by its Client, i.e. it takes decisions on behalf of another within limits specified by them, and is classified as a Delegate. In the other pattern, the intermediary doesn’t even select the Service it will use, as that too is specified by its Client, i.e. it carries out an action on behalf of another as specified by them, and is classified as an Agent. In each case, the classification is handled by a SPIN rule, the rule corresponding to Figure 6 is shown in Table 3.

ISSN 2052-8604

TABLE 10: CONSUMER CLASSIFICATION PATTERN SPIN RULE.

```
IF Assets {Client, Service, ServicePool} exist

AND Relations {Client uses Service, Client
specifies ServicePool, Client selectsFrom
SericePool, ServicePool supports Client,
ServicePool contains Service} exist

THEN Client is a Consumer
```

Threat models

It is not reasonable to insist that generic trustworthiness model cover all possible threats to assets, as in any type of system, some types of threats will be more important than others. However, one should decide consciously which types of threats to include and which to leave out. This implies a need for a methodical (and ideally standards based) approach to decide what threats should be included in the generic model. During initial attempts to derive the threat model, we used ISO27001 [13] and ISO27005 [12] as the starting points as these are well established standards for risk identification and management and are similar to the SESAR approach [31]. However, this approach failed, because while ISO 27001 provides checklists indicating the types of threats that should be considered, these are expressed in terms of security control objectives at system level. Our risk management methods requires a classification-based approach, and this led us to use the IETF RFC 4949 standard [24]. This standard is intended as an information security vocabulary, and the portion directly related to threats is quite brief. The main advantage of RFC 4949 is its decomposition of threats into threat actions (the event or situation that compromises the system) and threat consequences (the nature of the resulting compromise). The examples in RFC 4949 don't always rigorously maintain the distinction between action and consequences, but they show that factorizing threats along these lines can simplify the problem of identifying distinct classes of threats and avoiding unintended overlaps and gaps in the threat model.

ISSN 2052-8604

The process used was to enumerate possible threat consequences for each of the asset classes from Figure 3, so obtaining a complete set of possible threat consequences (e.g. data disclosure, corruption, underperformance, etc.). Then a set of threat actions that could produce these consequences in a threatened asset was identified, taking account of interactions with other assets (e.g. impersonation of one asset to another, deception, physical destruction, unrestricted access, etc.). A complete threat model can then be found by taking every meaningful combination of threatened asset, consequence and threat action. Not all combinations are meaningful (e.g. there is no sense in considering physical destruction of non-physical assets), but these can easily be eliminated. This leads to a very large set of generic threat classes, so in the OPTET application we restricted attention to threat consequences that mattered most: disclosure of information, corruption of information or processes creating information, and underperformance or unavailability of services or resources.

Threat actions and control strategies

Our ontology uses the generic design pattern from Table 4 to model threat actions and control strategies to block the action or mitigate its consequences.

TABLE 11: CORE MODEL DESIGN PATTERN

```
Threat threatens some Asset max 1

Threat involves only Asset

Control protects only Asset

ControlRule (SPIN) : Mitigates or Blocks Threat
```

For example, Figure 7 shows a typical example of this pattern, modeling a threat involving interruption of communication with a host over a network by directing a malformed message to exploit a known bug in the host operating system:

ISSN 2052-8604

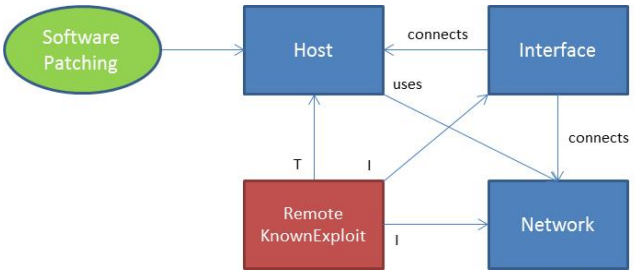


FIGURE 14: THREAT-ASSET DESIGN PATTERN IN THE TRUSTWORTHINESS MODEL

This diagram shows the relationships between the threat and the involved assets, including the fact that the threatened asset is the Host. The diagram also shows that the threat can be mitigated if the Host is protected by a software patching procedure, fixing bugs as soon as they are discovered. This is captured in the Generic Ontology by a SPIN control rule. Note that there may be multiple control rules that can be used – in this case one might also use a firewall protecting the Interface asset to block the malformed traffic before it reaches the Host.

Threat Consequences and Secondary Effects

Similar diagrams are used to describe the threat consequences. In Figure 8 the Host becomes unable to receive messages from the affected interface, which is modeled by saying both are unavailable, as shown in Figure 8. These consequences are not modeled directly in the ontology, but are added later as inputs to the Bayesian inference system for analyzing system behaviour.

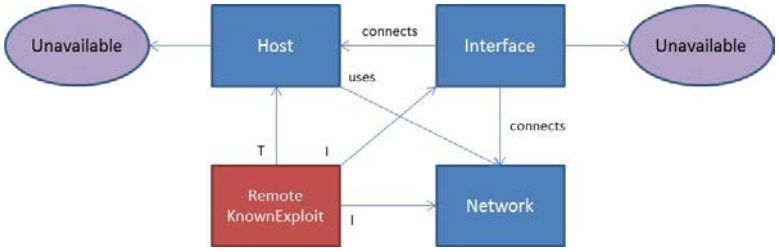


FIGURE 15: THREAT CONSEQUENCES

However, for Bayesian inference to deduce that a threat is active based on evidence from monitoring system asset behaviour, it is necessary that all of the consequences of the

ISSN 2052-8604

threat are included. In this example, if the Host supports a Service, the fact that the Host is unavailable means the Service would also become unavailable This in turn may compromise any Clients of that Service (e.g. if there were no alternatives available in the relevant ServicePool). Users of these LogicalAssets may also be affected, e.g. their level of trust in the system may become degraded. To capture all of these consequences of the initial threat, we would need a model far more extensive and complex than the one shown in Figure 8.

One of the most novel aspects of our modelling technique is the inclusion of secondary effects, which provide an elegant solution to this problem. A secondary effect is a threat that is caused by an asset misbehaviour. This is captured in the generic model ontology via a secondary effect rule. We can then keep the simple threat model exemplified by Figure 7 and Figure 8, and model knock-on consequences as secondary threats. In this case, one such effect would be the unavailability of a network which causes the host to become unavailable, as shown in Figure 9. (Note that for simplicity the threat relationships are omitted except to the threatened asset).

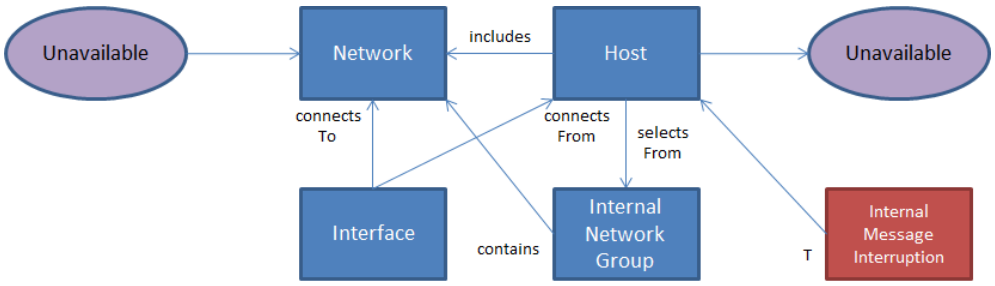


FIGURE 16: SECONDARY EFFECT CAUSED BY ASSET BEHAVIOR

This diagram represents the secondary effect rule shown in Table 5, which says that if the Network is unavailable then the host also becomes unavailable as a secondary effect.

ISSN 2052-8604

TABLE 12: SECONDARY EFFECT RULE IN SPIN

```
If

    there are instances of the following
classes: {Host, Network, Interface,
InternalNetworkGroup}

    and they are connected like this:

            the host includes the network

            the host selects from the internal
            network group

            the interface connects from the
            host

            the interface connects to the
            network

            the internal network group contains
            the network

    and the network is unavailable

    and there is a threat that involves all
    these assets and threatens the host

Then

        the existing threat is aclassified as
a secondary effect

Endif
```

ISSN 2052-8604

Figure 10 shows the possible counter measures which need to be implemented in order to mitigate the secondary effect shown in Figure 9.

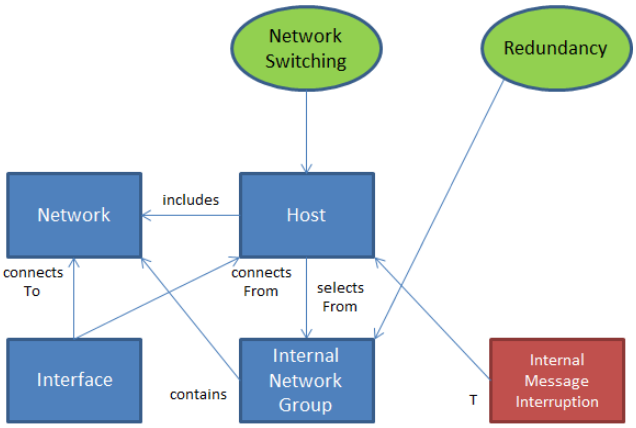


FIGURE 17: SECONDARY EFFECT CONSEQUENCES

Adding secondary effects to the modeling approach allows us to simplify the models of individual threats (actions and direct consequences), and capture knock-on consequences by modeling them separately. It also allows us to identify at run-time which threats are secondary (caused by observed asset behaviour). This is a huge bonus when we come to analyse observed asset behaviour to infer which threat(s) could be the cause. Any secondary threat can be excluded as a possible primary cause, and its effects taken into account when considering other possible causes. In the above example, the Bayesian inference algorithm considers that the Internal Message Interruption is active (but secondary), and the fact that the Host is unavailable is explained by that, making it easy to identify the root cause (a remote exploit) from the remaining behaviour.

Design-Time Models: a System Composer

The design-time trustworthiness model is produced by further sub-classing the assets and threats from the generic trustworthiness model, representing system-specific assets and their relationships, and variations on the threats involving different combinations of system-specific assets. The main challenge is that design-time trustworthiness system model development should be carried out by an expert in the system, who may not be an expert in security and almost certainly won't be expert in semantic modelling and reasoning technology. A typical system modeller therefore cannot be expected to develop

ISSN 2052-8604

models using semantic modelling tools like Protégé [21]. Hence, we introduce the system **System Composer**, which allows the creation of these design-time trustworthiness models and the automatic generation of system specific threats based on the knowledge encoded in the generic trustworthiness model.

The System Composer provides a means to design new multi-asset system models, comprising of classes and relationships specialised from a generic ontology model. Having created the structure of a new system model, this software allows system designers to generate an ontological representation (in OWL) of the Design-Time Trustworthiness Model. Inferences and templates (defined in SPARQL + SPIN) can then be run on this created ontology, which will generate a collection of possible threats to the designed model, as well as class level reasoning for each of the assets in the Design-Time Trustworthiness Model. This aids in performing a risk based analysis on the composed system and the counter measures which need to be taken into consideration before moving to the implementation stage.

The System Composer GUI

The graphical user interface exposed to the user is shown in Figure 11. It consists of 4 main parts.

1. The generic asset types which can be added onto the canvas to compose system topologies.
2. The central canvas which allows users to drag and drop assets and to add relationships between different asset types.
3. Asset classification plane: which informs the user about the type of asset (and any inferred roles it may have) and the in-coming and out-going relationships associated with this asset.
4. Threats identified for each asset and the possible countermeasures which can be implemented to either mitigate or block this threat according to the control rule defined for every particular threat (see Section 3.1 for an example of a generic control rule)

The System Composer runs on a desktop client and can be easily deployed on any platform. The ontologies required for the tool to function (core, generic models) are bundled along with the installation.

ISSN 2052-8604

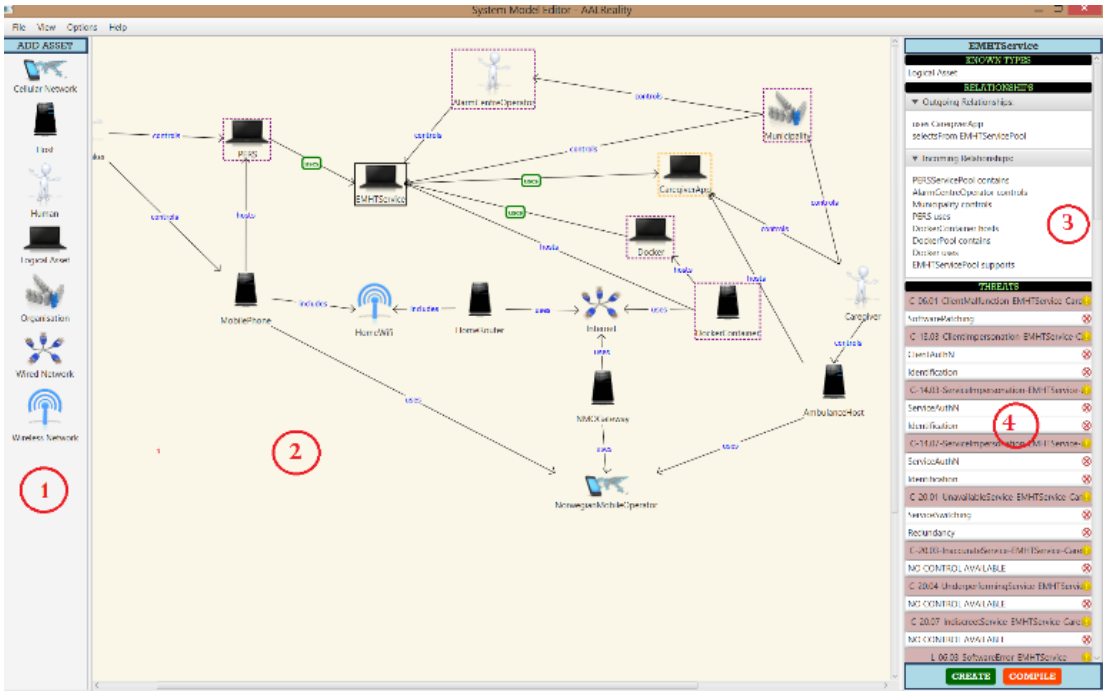


FIGURE 18: SYSTEM COMPOSER GUI

The available asset types for composing a system model are: (1) Cellular Network – a typical mobile phone network (2) Host – any physical asset which can host logical assets (3) Human – human user/stakeholder/operator (4) Logical Asset – processes/software running on the system (5) Organisation (6) Wired Network (7) Wireless Network. These assets were chosen after consultation with system designers from different OPTET validation use cases where socio-technical system design was involved. These use cases covered ambient assisted living, cyber security using distributed attack detection and visualiusation, and secure web chat for cyber crisis management. These generic classes can be subclassed by simply dragging and dropping them into the System Composer's canvas. Once dropped, they can be connected to each other where the relation type is read directly from the generic model - only valid relations can be created. Figure 12 shows the assets available in the System Composer.

ISSN 2052-8604

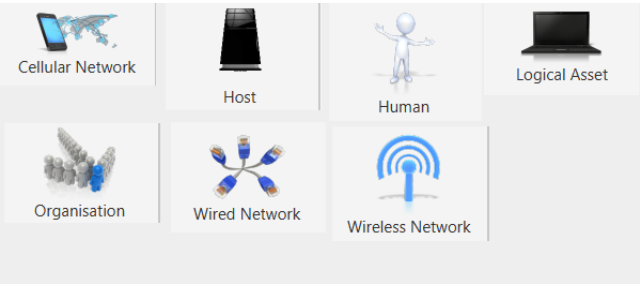


FIGURE 19: SYSTEM COMPOSER: AVAILABLE ASSETS.

Relationships between assets can be added by clicking on a particular asset on the canvas and the system automatically highlights allowed relationships (e.g. Host *hosts* LogicalAsset) to other assets within the system. This information has already been encoded in the Generic Ontology. The System Composer automatically fills in the relationships if there is only one possible option or presents a choice menu to the user in case more than one type of relationship is possible. Figure 13 shows an example of a Host, Logical Asset, Human and Cellular Network assets.

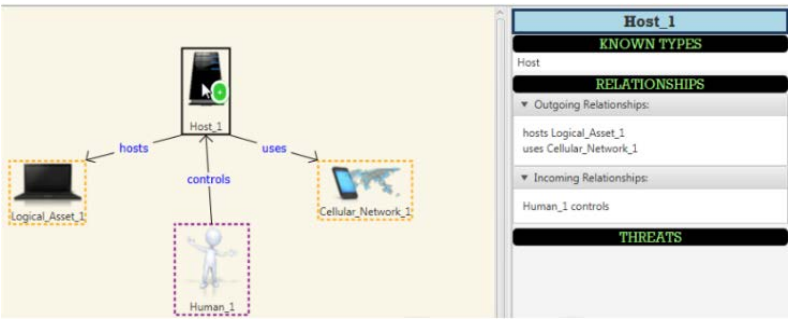


FIGURE 20: ADDING RELATIONSHIPS.

This model can now be "compiled", which means to invoke the System Model Compiler to (1) auto complete the model and (2) generate the system-specific threat classes. This is fully automated and doesn't require any additional input from the user. Automatic completion of the model requires both a combination of class level inference and SPIN rules which encode the patterns for automatic classification of Consumer, Agent, Delegate, patterns (see Section 3.2.1.1.). Additionally, some additional classes such as Interface (a class which connects Hosts to Networks) are auto generated. These asserted

ISSN 2052-8604

classes are not presented in the user interface to avoid unnecessary clutter and because their generation does not require human intervention and is rule based. Once the model has been completed, the second stage is the automatic generation of all system specific threats for the composed system. This will be explained in the next sub-section.

Automated Threat Generation and Analysis

The goal of the automated threat generation is to create system specific subclasses of generic threats depending on the design-time configuration. Such a threat subclass is generated by scanning the system asset model for matching threat definitions defined in the generic model and then subclassing these generic threats. Figure 14 shows how a threat class is axiomatically encoded in OWL in the Generic Model.

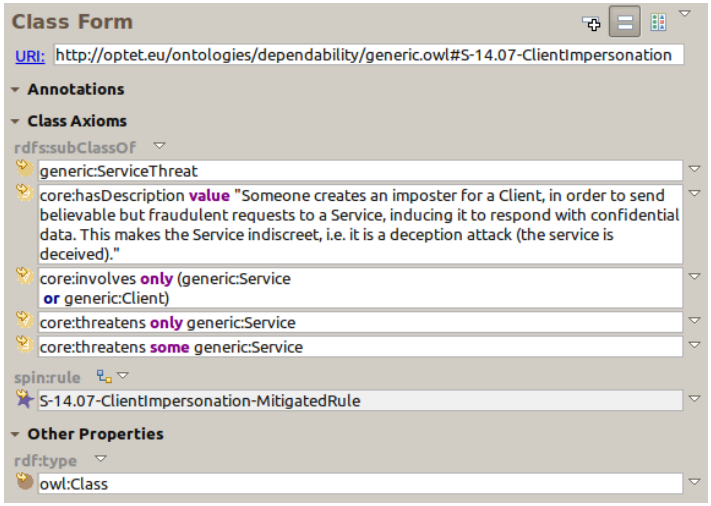


FIGURE 21: EXAMPLE OF A GENERIC THREAT CLASS

The threat shown in Figure 14 applies to the Client-Service pattern from the generic model. It threatens the Service in this pattern and involved both Client and Service and has a mitigation control rule attached to it.

Every generic threat class has a corresponding threat generation template encoded in SPIN which is part of the generic model. All rules are based on the same algorithm:

- Find a pattern in the system specific design-time model ignoring blank nodes

ISSN 2052-8604

- Create a new system-specific subclass of the generic threat for which this is a rule
- Apply new system-specific restrictions to the newly created class

This rule for the specific threat shown in Figure 14 (Client Impersonation threat) works in the following way (Table 6):

TABLE 13: CLIENT IMPERSONATION THREAT RULE LOGIC

```
If

    •   there are a system-specific
        client and service
        subclasses
    •   and the client subclass has
        a restriction to use the
        service subclass
    •   and no client impersonation
        threat exists for those two
        classes

Then

      •   create a new threat
          subclass, which has
          restrictions to threaten
          the service subclass and
          involve both the client
          and service subclass.

Endif
```

The system model compiler makes this process invisible to the user. After the system designer has built the system-specific model, he runs a reasoner first to determine the roles of the new assets within the model (e.g. Client-Service, Delegate etc.). Then system model compiler runs the threat generation templates and adds the newly created threat classes to the compiled design-time model. This model is then queried for all contained system-specific threats including the assets they affect and the possible controls that could be implemented. The resulting threats for each system specific asset are displayed in the Threat Panel of the GUI (see Section 4.1.1). The system also suggests possible counter measures which can be implemented to block or mitigate these threats and offers them as a checklist (see Figure 15). The system designers can now create a full checklist

ISSN 2052-8604

of the threats and countermeasures which need to be taken into consideration during system implementation and chose the option within the System Composer to generate a PDF report which will contain the system topology diagram, list of full assets and their relationships, threats and control measure checklist. This PDF document can be readily handed over to the implementation phase developers.

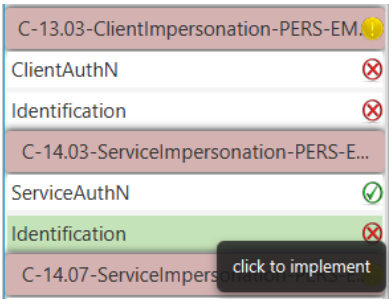


FIGURE 22: THREAT DISPLAY AND CONTROL OPTIONS.

Our approach is based on the assumption that the high-level system structure is composed by the system designer. If this high-level model is not capture correctly, it will not give the best results (overlooking of possible threats to assets). If the system is modelling correctly and the structure changes, we would need a new model. But in that case the fact that we use automated machine reasoning approaches means we should be able to apply the new model relatively easily.

Model validation

The modelling approach along with the System Composer were validated with use case owners for all the three use cases within the OPTET project (Ambient Assisted Living, Distributed Attack Detection and Visualisation and Secure Web Chat). The mini-workshops arranged with these three use cases showed that the System Composer was effective in hiding the complexity of the underlying semantic ontologies and processes and aided the user significantly in designing a socio-technical system, automatically generating system specific threats and perform risk analysis over the composed system. Ease of use was highlighted as one of the main plus points of the system. Factors taken into consideration to validate the semantic models included support for design time processes and activities, threat coverage and ease of integration into other software

ISSN 2052-8604

components. Full details of this validation and evaluation experiment will be presented elsewhere as it is out of the scope of this paper and due to space considerations.

Conclusions and future work

In this paper, we presented an approach for using semantic modelling techniques to support the creation of secure and trustworthy socio-technical systems. Encoding relationships between assets, threats and security countermeasures in a machine understandable format allows us to define a systematic method for the derivation of threats to assets based on existing standards, and enables the use of machine reasoning for auto completion these risk models with minimal human intervention. Further reasoning can identify options for mitigating threats at design time. At run-time, misbehaviour can be characterised and related to threat activity, root cause analysis performed by using chains of related secondary threats.

Next, we focused on making these models more usable through the help of an easy and intuitive graphical use interface (the OPTET System Composer), which allows system designers to effortlessly compose systems and perform risk analysis using the background knowledge encoded in the semantic models, the complexity of which is hidden away from the user of the system.

In terms of future work, we would like to further enhance the usability aspects of the System Composer based on the feedback received from the use case validation workshops. The semantic models will also need to be revisited; the addition of Human assets introduces a link between system trustworthiness and user trust levels during system run-time. The inclusion of user attributes and roles (e.g. age, sex, level of expertise, IT literacy, organisational role) at design phase could aid in estimating impact on user trust level as the trustworthiness of the system varies during run-time operation. We will also investigate the integration of our semantic models with trustworthiness application factories (where the software assets are developed) such that an economic analysis of mitigating controls can be done to avoid exposing significant vulnerabilities during the development phase.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme for the EU funded OPTET project.

ISSN 2052-8604

References

[1] Barnum, S., and Sethi, A. (2007). Attack patterns-knowing your enemies in order to defeat them. In *OMG Software Assurance Workshop: Cigital*.

[2] Blanco, C., Lasheras, J., Fernandez-Medina, E., Valencia-Garcia, R. and Toval, A. 2011. Basis for an integrated security ontology according to a systematic review of existing proposals. *Comput. Stand. Interfaces* 33, 4 (June 2011), 372-388. DOI=10.1016/j.csi.2010.12.002 http://dx.doi.org/10.1016/j.csi.2010.12.002

[3] Chakravarthy, A., Surridge, M., Chen, X., Hall-May, M. and Leonard, T.A. 2013. SERSCIS Deliverable D2.2: System Modelling Ontology and Tools: Final Implementation v1.5.

[4] Christopher, J. A. and Dorofee, A. 2002. Managing Information Security Risks: The Octave Approach. *Addison-Wesley Longman Publishing Co.*, Inc., Boston, MA, USA.

[5] Fenz, S. and Ekelhart, A. Formalizing information security knowledge". 2009. *in International Symposium on Information, Computer, and Communications Security*, Sydney, Australia.

[6] Fray, I.L. 2012. A comparative study of risk assessment methods, MEHARI & CRAMM with a new formal model of risk assessment (FoMRA) in information systems. In *Proceedings of the 11th IFIP TC 8 international conference on Computer Information Systems and Industrial Management* (CISIM'12), Agostino Cortesi, Nabendu Chaki, Khalid Saeed, and Sławomir Wierzchoń (Eds.). Springer-Verlag, Berlin, Heidelberg, 428-442. DOI=10.1007/978-3-642-33260-9_37 http://dx.doi.org/10.1007/978-3-642-33260-9_37

[7] Glimm, B., Horrocks, I., Motik, B. and Stoilos, G. 2010. Optimising Ontology Classification. *In Proc. of the 9th Int. Semantic Web Conf. (ISWC 2010), volume 6496 of LNCS, pages 225--240,* Shanghai, China.

[8] Hogganvik, I. and Stølen, K. 2006. A graphical approach to risk identification, motivated by empirical investigations. In *Proceedings of the 9th international conference on Model Driven Engineering Languages and Systems* (MoDELS'06), Oscar Nierstrasz, Jon Whittle, David Harel, and Gianna Reggio (Eds.). Springer-Verlag, Berlin, Heidelberg, 574-588. DOI=10.1007/11880240_40 http://dx.doi.org/10.1007/11880240_40

[9] Howard, M., & Lipner, S. (2009). *The security development lifecycle*. O'Reilly Media, Incorporated

[10] Herzog, I., Shahmehri, N. and Duma, C. An ontology of information security. 2007. *Internation Journal of Information Security and Privacy, pp. 1-23.*

[11] Ingoldsby, T. R. (2009). Attack tree-based threat risk analysis.

ISSN 2052-8604

[12] ISO/IEC 27005. 2011. Information technology -- Security techniques. *Information security risk management.*

[13] ISO/IEC 27001. 2005. Information technology – Security Techniques – Information security management systems – Requirements.

[14] ISO27001 (2005) Information Security Management System (ISMS) standard *Online: http://www.27000.org/iso-27001.htm (last accessed October 2014)*

[15] IT Grundschutz Manual. 2004. www: http://trygstad.rice.iit.edu:8000/Government%20Documents/Germany(BSI)/BSI%20IT-Grundshutz%20Manual%202004%20Introduction%20&%20Modules.pdf (last accessed October 2014).

[16] Kim, A., Luo, J., and Kang, M. Security ontology to facilitate web services description and discovery". 2007. *Journal on Data Semantics*, pp. 167-195.

[17] Matulevi, R., Mayer, N., Mouratidis, H., Dubois, E., Heymans, P. and Genon, N. 2008. Adapting Secure Tropos for Security Risk Management in the Early Phases of Information Systems Development. In *Proceedings of the 20th international conference on Advanced Information Systems Engineering* (CAiSE '08). Springer-Verlag, Berlin, Heidelberg, 541-555. DOI=10.1007/978-3-540-69534-9_40 http://dx.doi.org/10.1007/978-3-540-69534-9_40

[18] Meland, P. H., Spampinato, D. G., Hagen, E., Baadshaug, E. T., Krister, K. M., & Velle, K. S. (2008). SeaMonster: Providing tool support for security modeling. *Norsk informasjonssikkerhetskonferanse, NISK*.

[19] OPTET Project. 2013. www: http://www.optet.eu/ (last accessed October 2014).

[20] OWASP.org (2013), 2013 Top 10 List, online: https://www.owasp.org/index.php/Top_10_2013-Top_10 (last accessed October 2014)

[21] Protégé Tool. 2013. www: http://protege.stanford.edu/ (last accessed October 2014).

[22] Saitta, P., Larcom, B., & Eddington, M. (2005). Trike v1 Methodology Document. Online: http://dymaxion.org/trike/Trike_v1_Methodology_Documentdraft.pdf. (last accessed October 2014)

[23] SERSCIS. 2013. www: http://serscis.eu (*last accessed October 2014*).

[24] Shirey, R. 2007. RFC 4949: Internet Security Glossary. www: http://www.ietf.org/rfc/rfc4949.txt (last accessed October 2014)

[25] Shostack, A. (2014). *Threat Modeling: Designing for Security*. John Wiley & Sons.

[26] Sommerville, I.: Software Engineering, 9th edn. Pearson, Boston, pp. 286-287 (2011).

[27] SPARQL + SPIN. 2013. www: http://spinrdf.org/ (last accessed October 2014).

ISSN 2052-8604

[28] Surridge, M., Nasser, B., Chen, X., Chakravarthy, A., Melas, P. Run-Time Risk Management in Adaptive ICT Systems. In proceedings of the 8th International Conference on Availability, Reliability and Security. 2013.

[29] Swiderski, F. and Snyder, W. (2004) Threat modelling. Microsoft Press.

[30] ThreatModeller, http://myappsecurity.com/ (last accessed October 2014)

[31] Touzeau, J et al. 2011. SESAR DEL16.02.01-D03: SESAR ATM Preliminary Security Risk Assessment Method.

[32] Vorobiev, A. and Bekmamdova, A. An ontology-driven approach applied to information security. *Journal of Research and Practice in Information Technology*, pp. 61-76, 2010.

ISSN 2052-8604

The Need for Modelling and Experimentation with Decentralised Networks: a Case Study using Bitcoin

Simon Robinson[1] and Tim Watson[2]
[1]Cyber Security Centre, De Montfort University
csc@dmu.ac.uk

[2]Cyber Security Centre, University of Warwick
tw@warwick.ac.uk

Abstract

Decentralised networks are increasingly used to share computer resources for a range of tasks. Arguably the most popular, and recently covered in the general media is the crypto currency Bitcoin. Created in 2009, Bitcoin has grown in popularity and is used around the world with a daily trade value of 2.7 billion USD. It is used to purchase a range of services which require the transfer of currency from one place to another – from funding criminal activities to making donations to charitable causes. Current media and economic opinion is that Bitcoin is in the early stages of adoption and development. There is speculation and hype surrounding its importance, with no significant experimentation to substantiate the claims. In this position paper we propose the need for a network modelling solution that can be used to determine answers to questions concerning decentralised networks with a crypto currency as a case study. Questions such as: How stable is the network? What risks would adapting the scripting mechanism introduce? What are the long term impacts on other economic networks? Can this system be adapted to deliver other benefits? It will have value to look at other crypto currencies but within the constraints of this position paper Bitcoin will be the focus. First, we describe the principles and technology that underpin the system. This is followed by a description of what is and is not known about the network and what the risks are. Is society is at risk of adopting and trusting an unproven technology without in-depth analysis? Research to date is insufficient to demonstrate the impacts and risks of participating in and developing improvements for the Bitcoin network.

ISSN 2052-8604

Introduction

As society evolves it identifies new ways of converting human toil into something of value. This can then be exchanged for other things of value and, as described by Zorpette (2012), this has evolved and become more elaborate over time. This is liquidity and a foundation of economics. Cash, almost worthless on its own, is based on this concept of value trade and is an IOU from the issuing authority. A remarkable product of the digital era is that society is now not only converting human effort into currency but computational effort is considered a tradable commodity. There is no clearer example than that of the cryptographic currency Bitcoin, devised by Satoshi Nakamoto (2008). This poses computers as part of a distributed and decentralised peer-to-peer network, competing to be the first to solve a computational problem. Once a proof to the puzzle is found it is broadcast and validated by the peers (other computers). If validated this value will be added as the most recent entry into the public ledger known as the blockchain. The successful computer is rewarded with a dividend of bitcoins for its efforts. Bitcoin is but one application of a computational revolution and Governments are having to understand what societal impacts it will have and how they regulate something without centralised control, according to Frisby (2014). This is in its infancy and long term impacts, application, stability and viability are not clear. This position paper will demonstrate that there is a need to better understand this technology and by combining network, economic and agent-based models this can be achieved. This position paper will use Bitcoin as a case study. Section 2 will provide a brief technical appraisal of Bitcoin. Section 3 will speculate on the main properties of Bitcoin that make it popular. Section 4 will identify the most prominent impacts of the Bitcoin system and Section 5 will provide an overview of the main assumptions and concerns over this technology. Section 6 will discuss the themes of the position paper and Section 7 will propose some future research.

Bitcoin

A decentralised network is one where the processing and resources are spread over the network rather than centrally. Bitcoin applies the same concept to currency. The established method relies on banks and partners providing a service. They issue, process and store currency on behalf of their customers. Bitcoin offers the opportunity for coin owners to contribute to the operation of the network. By allowing their computer to become a node on the network they participate in communicating, validating and committing Bitcoin transactions to the public ledger. This implies that ownership and survival of the network relies on the continued support of its users.

ISSN 2052-8604

A direct benefit of Bitcoin is the ability to establish trust between two anonymous parties (or, more precisely, to allow a trusted transaction in the absence of trust between transacting parties). The Bitcoin transaction, which will be described later, is a contract between two or more secret identities. This contract is validated by the participants in the network ensuring that the paying party has the funds. If accepted the transaction is committed to the public ledger.

Bitcoin can be broken down into 3 main component parts:

2.1. Blocks and the Blockchain
2.2. A PKI identity scheme
2.3. A transmission protocol

Blocks and the Blockchain

A block is a collection of Bitcoin transactions that have been received, validated and processed into the ledger. Each block references the previous block as a source so each block is linked in a chain back to the source. The blockchain is the public ledger that holds a record of all transactions. Contributors to the Bitcoin network compete to be the first to commit a block to the blockchain. To achieve this they have to be the first to find a solution to a proof of work challenge.

Based on Hashcash, Back (2002), the proof of work challenge is designed in a way that it requires a large amount of computational power to solve compared to a small amount to validate. The level of complexity is flexible to allow the network to self adjust based on an expected performance threshold. The mechanism uses a hashing algorithm which can process any digital input and produce a fixed length output. The output cannot be predicted from knowing the input values. if there were multiple values and put into a different order the output would be completely different. Predicting the output or producing an output that had a specific structure is difficult. A node hashing a block attempts to produce a hash with a preceding number of zeros. The number of zeros is known as the target. This is the complex challenge: the Bitcoin proof of work. If the first attempt does not succeed then an integer in the block is incremented and it is hashed again. This is repeated until the desired minimum result is found.

ISSN 2052-8604

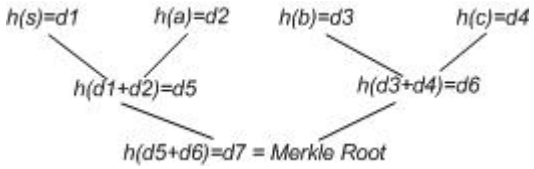


FIGURE 23- PRINCIPLE OF A MERKLE TREE

Each transaction to be processed is hashed in a similar way to the block. The result of the first two transaction hashes are concatenated together and then hashed. This is repeated until all root transactions and respective results are hashed into a single Merkle Root, Merkle (1980) as shown in figure 1. This ensured that the number of transactions in a block did not impact on the size of the block. This value has the same fixed length irrespective of the number of transactions. The Merkle Root is appended to the header of the block and hashed together. This process ensures each node is processing the same size of data To find a valid proof would require Nodes to produce thousands of hashes therefore the level of conmplexity required to find a proof is determined by a value known as the target.

The target changes with time enabling the difficulty of the problem to be adjusted. By calculating the time it took to process the past 2016 blocks. If the average is longer than 10 minutes then the difficulty is reduced, shorter then increased. The average time per transaction is set to regulate at 10 minutes per block.

The first transaction in each block is a mining transaction designed to credit the node with the free dividend of coins. The initial transaction will be different from node to node therefore all blocks will have a different combination of transactions and generate a different Merkle Root. This variation means that each node has an equal chance of finding the proof on the fewest number of attempts, levelling the playing field where processing power would typically provide an advantage.

When a proof is accepted into the network a hash of this block is included into the header of the next block creating the next link in the chain. A proof that is successful in being added to the blockchain earns the winner of the race a dividend of coins due to the self crediting first transaction. This process is known as mining and is defined within the network protocol. Private Identities

ISSN 2052-8604

To any external party observing transactions in the blockchain they would not be able to attribute the identities of any of the transacting parties due to the use of Public Key Infrastructure (PKI). Using elliptic curve cryptography [ELC] Caelli (1999) the Bitcoin network relies on a user having public and its associated private keys when making transactions. These keys are paired mathematically whereby if one is used to encrypt information only its corresponding key from the pair can decrypt it. One key is made public, hence known as a public key, and is used as an identifier on the network for transactions to be made to. In Bitcoin a transaction is made declaring the public key of the coins that are being paid to. If verified these will be locked to that public key. To spend the coins the owner must sign the transaction message with their private key. Due to the mathematical link the network can quickly verify that the signature is linked to the public key the coins are locked to. This commitment to the blockchain and network acceptance provides non-repudiated transactions that cannot be reversed or double spent, providing the network remains honest. The ability to verify the authenticity of coin ownership in a proposed transaction and committing it to an irreversible ledger allows trust to be established in an anonymised network. Therefore this can be considered a decentralised trust network, ensuring transactions only continue if the trustworthy spender can be established.

Protocol

Prior to broadcasting a transaction to the network the spender must format the structure of the transaction. This consists of a version declaration for format checks, and a declaration of the inputs and outputs for the transaction. In order for a transaction to have value the spender has to declare the previous transaction(s) the coins were received from. The output(s) should balance the input with any change from the transaction being declared as a payment back to the spender. If change is not accounted for then this can be claimed by the node that processes the transaction into a block.

When a user's machine connects to the Bitcoin network it receives an update of nodes within a certain proximity that are active. This forms an address book and is used to declare a transaction to the network. A basic message is sent to the nodes stating a new transaction is being declared. Nodes that are active respond for details and the transaction message is sent. The recipient nodes validate the message. If it has not been received before they will broadcast a new transaction declaration to the nodes in their address book. This soon propagates the transaction through the network for the mining nodes to include into their blocks for processing. This demonstrates how coins are never

ISSN 2052-8604

physically transferred and the concept of a wallet is a misnomer where coins have no physical representation and are transient as a total or previous transaction histories.

The incentivised concept of creating money from computational resource could be considered as an ideal currency for a digital world but it also has application in the physical world.

Popularity is pervasive

The current output of mining a block is 25 bitcoins which at current market rates is an equivalent of $9,590.78 USD[9]. This is more than a $200 increase on the value a year ago. This high value return on a process that requires no physical or mental effort for the user has helped contribute to its success. This cannot be the only reason as the value two years ago was about $10 therefore another factor had to contribute to its popularity.

Bitcoin offers a payment mechanism that preserves the anonymity of the transacting parties yet provides non-repudiable transactions. A large volume of digital cash systems proposed in recent years focused on trying to prevent the issue of double spending. The concept is that a mechanism is needed to prevent a digital coin from being copied and spent over and over again. This would defraud vendors of their goods and devaluing the coin system. The majority of controls needed a central authority to act as the broker. In most cases a mechanism that can revoke anonymity if a double spending event was believed to occur was proposed. This combination of cash issuer and faux anonymity could be a reason why none of these technologies had become popular. Where Bitcoin differs is that it provides the ability for anonymous parties to transact without the concern that the coins could be copied. With no need for a central authority this cash system could be used to perform trading in an environment where anonymity and a lack of trust are common. The Dark Web, as it is popularly becoming known, is a place on the internet that is not indexed by conventional search engines and can only be accessed through anonymised routing services. The Dark Web hosts markets for the sale of illegal products and services from drugs to credit card fraud services. The regular trade needed a system of coinage that was untraceable to a real identity and irreversible. This meant that trade can occur without the fear of being caught through or the transaction being reversed after goods had been shipped.

---

[1] [9] http://www.coindesk.com/price/ 21-10/2014

ISSN 2052-8604

Due to the anonymity of the Dark Web it is no surprise that a system of coinage that supports the lack of identification has become prevalent there. Barratt (2014) indicated that although Bitcoin is the currency of choice for many vendors on sites such as the Silk Road there were conflicting views over its lack of price stability and ease of access and use for bitcoins. One point made was that Bitcoin added an additional concern for being ripped off. This is the flipside of preserved anonymity: you can make a spend but on sites where you are not able to obtain the goods in a "fair transaction" you will need to rely on secondary trust mechanisms such as vendor rating to ensure you will get the goods as expected.

The popularity of Bitcoin is thought to be due to the benefits of anonymity, low-cost transactions and lack of regulation, resulting in Bitcoin rapidly growing in popularity to become a concerning entity for governments Frisby (2014) and popular tool for the privacy conscious.

Regulation or Revolution

The impact of a decentralised trust network, such as Bitcoin, is not yet known. Various parties have speculated on the potential for such a system yet there are no comprehensive studies looking at the social, financial or technological impacts of benefits of such a revolutionary system. Bitcoin does a number of things that has people concerned. It removes the power and control of a central trusted authority. There is no controller dictating what you can and cannot do, the bias is not controlled by an authoritarian who knows better than you. There are no costs incurred to join to merchants or customers. Authority and control is granted to everyone and to no one individual. You can only pay if the network validates that you can. Fees or charges, if present at all, are paid to those that contribute to the upkeep of the network. The responsibility and ability for securing one's coins are easily within the capability of the individual. All you need is to keep your private keys secure. A task that can be achieved by writing a series of 51 alphanumeric characters on a piece of paper and hiding it in a secret place.

The protocol and code that defines Bitcoin is an open-source project that anyone can contribute to. This both establishes a sense of ownership in the community and creates a more robust peer reviewed platform. This has also led to a number of variants on the Bitcoin theme which are being publicly traded. These all have different properties and applications. Not all are your conventional cash systems.

ISSN 2052-8604

With the increased popularity, low maintenance cost, difficulty to trace and ability to adapt it was little wonder that the regulators started to ask questions. In 2014 the US congress issued a statement of concern for the use of Bitcoin within the community without proper regulation. The result of this initial enquiry was to conduct another study with the need to understand what can be done. One of the key questions is how to regulate its use and how to tax a financial system that has no central authority and accountability?

The prevalence of smart phones in western societies have made the challenge of using a digital currency in the physical world surmountable. The use of apps like Blockchain Wallet[10] have meant that vendors no longer need to pay expensive bank charges to create a pay point for transactions. The value of transactions has remained low due to convenience. Rather than waiting 10 minutes for a transaction can be committed into the block, vendors will accept the transaction once they have validated that the transaction is well formed. This acceptable risk to the vendor indicates the low cost and transparency of transactions makes Bitcoin an attractive alternative to credit card vendors that add significant cost to businesses.

The US Federal Bureau of Investigation (FBI) arrested and charged Ross Ulbricht on 01/10/2013 for, among other things, allegedly being the current owner/manager of The Silk Road. This was one of the most popular drug trafficking sites on the Dark Web. The Silk Road charged a percentage fee from Bitcoin transactions earning Ulbricht thousands of bitcoins. One year after the arrest the only bitcoins the FBI were able to sell were 29,000 coins as 'proceeds of crime'. Ulbricht allegedly still controls a high proportion and so far the FBI and the American court system are powerless in their pursuit to divest Ulbricht of his fortune. Due to the nature of bitcoins and the public ledger and decentralised control ownership cannot be reassigned. Unless they have the private keys linked to the coins they are unusable. This demonstrates that an owner of bitcoins does not need to trust the strength and layers of controls of a bank to protect their money. As long as their practice of protecting the private keys of Bitcoin accounts are good then the coin security is strong.

When people follow less than best practice, however, they can be susceptible to financial loss. In the last quarter of 2013 there were a number of online Bitcoin "banks" that had

[10] Can use QR codes and smart phone data connection for easy mobile transactions: https://blockchain.info/wallet/android-app

ISSN 2052-8604

their virtual vaults emptied. MTGox, as an example, was handling around 70% of global Bitcoin trades in 2012, according to Frisby (2014) and was the largest exchange providing services to administrate wallets on customers behalf. In february 2014 all trading ceased and quickly speculation and the media reported a figure of around 350 million dollars worth of bitcoins had been stolen Bitcoin is built on strong cryptographic controls and keys are meant to be private. When trust is placed in a third party to manage these private keys behalf of the owner control is essentially lost as the keys are no longer private. Essentially whoever has access to those servers has access to those keys and thus the bitcoins those keys unlock. The considered benefit of combining non-repudiation and anonymity meant the transactions were untraceable and irreversible.

Assume we know nothing

Some properties of the Bitcoin network are taken at face value without serious consideration for their validity. Other new ideas are conceived and promoted without being thoroughly tested for their impacts. The following are some examples where further study would be beneficial:

Economic Stability

Financial and technology experts are either naive or uncertain in what the future holds for this financial system. Many experts disagree on how this new method of trading in value could have an impact on modern society.

Solomon (1999) raised concerns over the inability to control electronic currencies. They proposed that systems like Bitcoin could weaken or erode a government's economy if not controlled. Over 10 years later the Guardian (2013) reported how the U.S. Senate held hearings on the concerns and methods to regulate Bitcoin. If a government or regulators were not able to directly influence the economy with tactics such as quantitative easing then they would not be able to control undesirable behaviours such as inflation (which is also sometimes used by governments to evaporate debt). Fluctuations and instability in an economy based on such a currency could have devastating impacts on the wider markets.

Economic experimentation

Some countries have tried to adopt or develop a digital currency as part of its national monetary system. A post in the Wall Street Journal (2014) indicated that Canada, who

ISSN 2052-8604

were at the forefront, closed down their programme**.** Parties within the Isle of Jersey are lobbying the government, according to CoinReport (2014), to request the legalisation of Bitcoin for investment and international trade. Using a currency such as Bitcoin could possibly make it easier for large corporations to move money overseas or for criminals to launder money without being properly scrutinised.

Ametrano (2014) posed a Bitcoin spinoff dubbed "Hayek Money" posing a solution to the instability in market value of Bitcoin. Stabilising the value of Bitcoin could enable it to be integrated in to national currency systems. To fully evaluate the effectiveness to integrate and support government economies it would need to be evaluated in an effective way.

In March 2015 Dominica, in the Caribbean, will be the subject of a mass Bitcoin Experiment. According to PanAm Post (2014) 70,000 residents will be given small quantities of bitcoins creating the largest high density community of Bitcoin owners. The interest is to understand how these coins are traded and what influence they have on the behaviours of the island's economy. Can this type of experiment be considered ethical or possibly have negative outcomes on the social and national currency of the island? This should be evaluated under laboratory conditions before experimentation on people.

Changes to the engine

Sing (2013) proposed a new way to improve Bitcoin efficiency using code verification for faster transactions in the real world. Miers (2013) proposes Zerocoin, a Bitcoin adaption, which uses stronger cryptographic controls improving on the anonymity model of Bitcoin. Both of these examples claim improvements or benefits but should be thoroughly tested. They may not have addressed or dispelled any weaknesses or faults in their proposals. They remain theoretical technological improvement only until more detailed trials can be developed.

Another component of the Bitcoin network that changes is the scripting mechanism. Each transaction can have an element of scripting built into it. Most of these mechanism are disabled but some have been found to leave users vulnerable to forms of attack. Some script attributes have only been tested on the dummy network before they were either disabled or promoted to the live network. It is unknown what would be the impact in the real digital world when they are enabled. Will they make the network unstable or leave people vulnerable to coin thieves?

ISSN 2052-8604

Anonymous but exposed

In a recent paper by Biryukov (2014), it poses that the use of Bitcoin through the TOR network reduces the anonymity of Bitcoin transactions dramatically. Nodes in the TOR network can gain control of the information that they distribute performing a man in the middle attack and use this information to piece together users' transactions and trace them to the point of origin.

Trust in anonymous nodes

A proof is often cited that as long as the network has 51% of nodes acting honourably then it would be too difficult to compromise the block validation process. This is a mathematically sound principle but it does not identify the probability that the number of honest nodes will drop below 51%. There are situations that could arise to tip the favour in an malicious node or collection of nodes favour.

There will only ever be a finite number of nodes active at any time. If there was a discovered vulnerability[11] allowing a Denial of Service (DoS) attack against a portion of nodes, what would the impact be? As there needs to be a majority of honest nodes the question can be considered in two ways: 1) What percentage of nodes would need to be attacked to reduce the honest nodes to a point where an attacker can subvert the blockchain? 2) How long would an attack need to be sustained in order to subvert the blockchain? The answers to these questions are important to determine how cost effective such an attack would be to a malicious node. Would there be any situation where the adversary would be able to make such an activity profitable? Having an effective way of analysing such impacts on the network would be of practical use in not only vulnerability analysis but understanding the strength of proposed remediation.

Discussion

This position paper has provided evidence of open or unanswered questions in relation to the use, impact and trust in Bitcoin. The points raised vary in their complexity and application from the analysis of the impact on fiscal systems and society for adopting such a technology. Changes to the technology are proposed claiming benefit without proving

11 Vulnerabilities for DOS have been reported on the Bitcoin Foundation site such as: CVE-2012-2459: Critical Vulnerability (denial-of-service): https://bitcoin.org/en/alert/2012-05-14-dos

ISSN 2052-8604

evidence of sufficient testing. Vulnerabilities have been discovered on the scripting mechanisms but only after they have been in the public for a prolonged period of time, so there is a need to know how stable the current version is. Understanding and performance analysis on how changes to the network or the network changes society are needed.

There are different methods that can be deployed to answer these questions individually, but only one that could answer them together. To provide answers to these questions and test the hypotheses concerning potential regulation, social impact or technology changes attributes of the network cannot be considered in isolation. The network has wide ranging impacts and applications and a method to model these networks and their transactions in a scalable and manner that can be considered applicable is required. A tool that can be built on a tiered modular architecture of differing network models would be able to model decentralised networks within a contextual environment. The adaptability will enable different scales of experimentation to be conducted. Models can be used to test hypothesis in isolation or to place them in a contextual scenario applying economical or social constraints on the interaction. The following types of tests can be conducted:

- Agent based studies - looking at how user behaviour or miss-use cases can identify risks within the network and its use
- Observations of information propagation across the network
- Impacts of crypto wallet software solutions and vulnerabilities in cloud based solutions of client based
- Block chain and transaction studies adapting scripts and mining behaviours
- Long term sustainability and scalability of currency systems

Future work

Following this study the research team have designed and develop is ongoing for a multi tiered simulation laboratory that enables the study and experimentation of:

1. Decentralised Networks (Bitcoin)
2. Agent modelling of behaviour + misuse
3. Economic modelling and impact

ISSN 2052-8604

This will be used to test specific questions including those mentioned within this position paper. The tiers can then be swapped and replaced by others to provide a wider context for experimentation. The first layer will use the Bitcoin framework and libraries to model behaviour. After Bitcoin has been tested that code library will be replaced by other crypto currencies to broaden the study. This will be replaced with other emerging decentralised models where similar experiments to determine impact and risk can be conducted.

References

[1] https://blockchain.info/stats 22/11/2014

[2] Zorpette, G. (2012). The beginning of the end of cash [Special Report]. Spectrum, IEEE, 49(6), pp.27 –29.

[3] Nakamoto, S. (2008), Bitcoin: A peer-to-peer electronic cash system. Consulted, 1, p.2012.

[4] Frisby, D. (2014), Bitcoin The Future Of Money, London: Unbound

[5] Back, A. (2002). Hashcash-A Denial of Service Counter-Measure.

[6] Merkle, Ralph C. "Protocols for public key cryptosystems." 2012 IEEE Symposium on Security and Privacy. IEEE Computer Society, 1980.

[7] Caelli, W. J.; Dawson, E. P. & Rea, S. A. (1999), 'PKI, elliptic curve cryptography, and digital signatures ', Computers & Security 18(1), 47 - 66.

[8] Coindesk (last modified 16th October 2014) *What Can You Buy With Bitcoins* Available from: http://www.coindesk.com/information/what-can-you-buy-with-bitcoins/ [Accessed 30th October 2014]

[9] Barratt, M. J.; Ferris, J. A. & Winstock, A. R. (2014), 'Use of Silk Road, the online drug marketplace, in the United Kingdom, Australia and the United States', Addiction 109(5), 774--783.

[10] Solomon, ElinorHarris. "What should regulators do about consolidation and electronic money?." Journal of Banking & Finance 23.2 (1999): 645-653.

[11] The Guardian (2013) *US regulations are hampering Bitcoin's growth* Available at: http://www.theguardian.com/commentisfree/2013/nov/18/bitcoin-senate-hearings-regulation [Accessed 09th September 2014]

[12] The Wall Street Journal (2014) *Canada Puts Halt to MintChip Plans; Could Sell Digital Currency Program* Available at: http://blogs.wsj.com/canadarealtime/2014/04/04/canada-puts-halt-to-mintchip-plans-could-sell-digital-currency-program/ [Accessed 09th September 2014]

[13] CoinReport (2014) *Isle of Jersey considering adopting Bitcoin as official currency* Available at: https://coinreport.net/isle-of-jersey-adopt-bitcoin-official-currency/ [Accessed 09th September 2014]

ISSN 2052-8604

[14] Miers, I. et al., 2013. Zerocoin: Anonymous distributed e-cash from Bitcoin. In Security and Privacy (SP), 2013 IEEE Symposium on. pp. 397–411.
[15] PanAm Post (2014) *Dominica Will Be First Nation With Universal Bitcoin Possession* Available at: http://panampost.com/belen-marty/2014/08/28/dominica-will-be-first-nation-with-universal-bitcoin-possession/ [Accessed 09th September 2014]
[16] Singh, P. et al., 2013. Performance Comparison of Executing Fast Transactions in Bitcoin Network Using Verifiable Code Execution. In Advanced Computing, Networking and Security (ADCONS), 2013 2nd International Conference on. pp. 193–198.
[17] Ametrano, F.M., 2014. Hayek Money: The Cryptocurrency Price Stability Solution. Available at SSRN 2425270.
[18] Biryukov, A. & Pustogarov, I. (2014), 'Bitcoin over Tor isn't a good idea', arXiv preprint arXiv:1410.6079./

ISSN 2052-8604

The social psychology of cybersecurity

John McAlaney[1], Jacqui Taylor[1] and Shamal Faily[2]
[1]Department of Psychology, University of Bournemouth
{*firstinitiallastname*}@bournemouth.ac.uk

[2]School of Design, Engineering and Computing, University of Bournemouth
{*firstinitiallastname*}@bournemouth.ac.uk

Abstract

As the fields of HCI, cybersecurity and psychology continue to grow and diversify there is greater overlap between these areas and new opportunities for interdisciplinary collaboration. This paper argues for a focus specifically on the role of social psychology in cybersecurity. Social psychological research may help explore the dynamics within online adversary groups, and how these processes can be used to predict and perhaps prevent cybersecurity incidents. In addition the issue of motivations of cyber adversaries and the social context in which they operate and will be discussed. Finally the benefits of the shared experience of psychologists and cyber security practitioners in addressing issues of methodology and conceptual development will be explored.

Scope of this Document

The scope of this document is to discuss and evaluate the role of social psychology in understanding the actions of cyber adversaries, and to evaluate how collaborative research might be used to improve approaches to prevention and mitigation.

Background

Cybersecurity incidents extend beyond the technological aspects of the attack. Recent incidents involving large organisations such as Sony serve as examples of both the wider social causes and social consequences of cybersecurity incidents. The growth of social media provides cybersecurity actors, both adversaries and targets, with more ways to present themselves in terms of the motivations for their actions and their responses to incidents. This dialogue in turn contributes to the social and cultural context that cybersecurity actors operate within, and which in a case of reciprocal causality is also a determinant of their actions. The collective nature of some cybersecurity incidents and the social roles of those involved in cybersecurity incidents has become the focus of study and comment by anthropologists[1] and social media analysts[2], yet there remains a lack of research. A better understanding of the social factors of those who instigate cybersecurity incidents is important in a number of ways for the development of prevention and mitigation techniques.

ISSN 2052-8604

Social psychology research focuses on how the behaviour and cognition of individuals is influenced by the real, imagined or implied presence of others[3]. As such it is one area of study that can be used to begin to explore the social psychological factors of cyber-adversaries. There is of course already a history of collaboration between psychology and computing through the interdisciplinary research conducted within Human-Computer Interaction (HCI), however it could be argued that the focus of this work has been more on the cognitive aspects of psychological processes rather than the social aspects. This paper will discuss and evaluate how social psychology research is currently incorporated into cybersecurity, and what further contributions the field may make for cybersecurity practice. This discussion will be arranged to reflect the conceptual model of the role of social psychology of cyber adversaries in cybersecurity that is shown in Figure 1, the case for which will be argued in the following sections. This will be followed by a discussion on directions for future research, and how the collaborative work of social psychologists and cybersecurity practitioners may further complement each field.

Group processes

When examining cybersecurity incidents it would appear that the actions of many cyber adversaries are group based in nature, as in the case of well-known hacktivist collectives such as Anonymous[4]. The activities of these groups often appear to be the result of conversations held on message boards such as 4chan or Internet Relay Chat (IRC)[1]. However it is important to note that as demonstrated in social psychology research there does not need to be actual contact between individuals for group processes to influence behaviour. As commented the imagined or implied presence of others can also influence individual behaviour[3].This may be particularly relevant to anonymous online discussions or the posting of messages on websites such as 4chan, where it may not be immediately clear to an individual if their actions are in fact being observed by others. In contrast to an offline situation such as a group activity in a physical room an individual who is acting online may have very little sense of how much of an audience they have, and what status within a group they have. In these situations the imagined or implied presence of others may become particularly pertinent. Overall it could be argued that there are very few cybersecurity incidents that are instigated by entirely an individual without there being any influence of group processes, even when the individual is primarily responsible for the incident.

ISSN 2052-8604

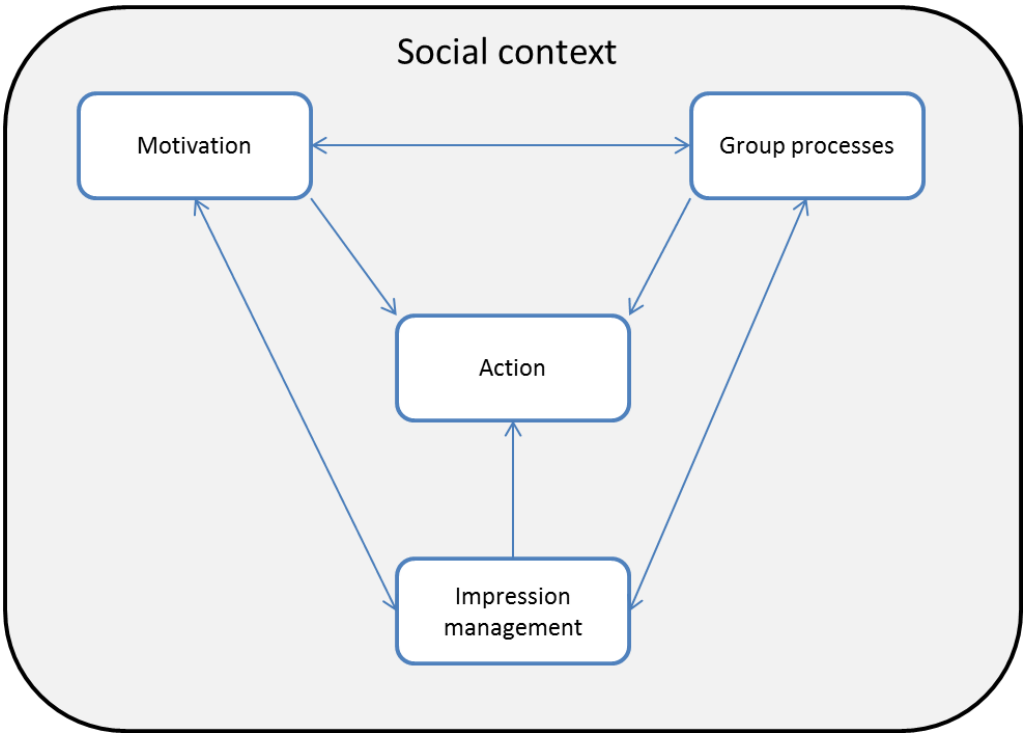


FIGURE 1: A CONCEPTUAL MODEL OF THE SOCIAL PSYCHOLOGY OF CYBER ADVERSARIES

When considering group processes social psychological research on social influence, attitude and behavioural dynamics is particularly relevant. In the case of Anonymous it has been stated that the majority of harm associated with some of the incidents was caused by a small number of technologically skilled individuals, though use for example of botnets[4]. There may have indeed only been a handful of people in this technologically skilled, smaller group but they were acting within a social context where group members were praising them and encouraging them to attack new targets. This type of positive reinforcement would be expected to increase the likelihood of individuals of the more technologically skilled group engaging in further, similar acts, as predicted by a multitude of social psychological theories of behaviour[3]. At the same time it is claimed[4] that members of the wider collective were manipulated by those leading the group action into believing that their actions using LOIC software was in fact what was primarily responsible for the incidents. In other words, social engineering was used within the group. By giving people the perception of having a role in the achievement of a goal the individual's sense of membership will be solidified, as predicted by social psychology research [5]. It would be of interest to explore how members of these groups would respond to the knowledge that they may have been manipulated by in-group members. As noted in psychology

ISSN 2052-8604

research people can respond negatively to the suggestion that they are being manipulated in some way, a response known as reactance[6]. There have been some examples of this type of reaction within hacktivist groups. For instance the revelation that one especially prominent and respected member of Anonymous was in fact working undercover for the FBI appeared to cause serious distress to other group members, as well as bringing disruption to their activities[4].

In keeping with intergroup attribution research [7] the success of the group actions of collectives such as Anonymous could also have been expected to strengthen individual members' beliefs that they are highly skilled, and that any successes of opposing groups such as law enforcement are more attributable to external circumstances and luck. This process could lead to decision making biases within the group, and could be argued to have emboldened the group to take further actions against other organisations, in the erroneous belief that their risk of being individually identified by law enforcement was lower than it actually was. Indeed many of the main individual adversaries that orchestrated the incidents associated with Anonymous in the early days of the collective have now been arrested and prosecuted[4]. Linked to these decision making biases is the effect that media reporting could have on such groups. It has been commented that early news reports about Anonymous generally overstated both the level of cohesiveness between group members and organisational structure of the group[4]. The category differentiation model of social psychology[8] suggests that the simple act of an external entity identifying a group as being a group can increase the likelihood of individuals identifying themselves as group members. In addition it has been observed that self-esteem is in part derived from membership within groups [9], particularly when that group has been engaged in conflict with what is seen to be a larger oppressor. In order to protect the self-esteem gained from these group memberships individuals may react strongly to exclude anyone who is seen to be threatening the group norms or group cohesion. This may explain some of the tensions and intra-group conflicts that invariably seem to appear within any kind of online group or hacktivist collective, where it is common for splinter groups to form and target one another[1]. Monitoring these types of reactions could be used as an indicator of how cohesive a group is becoming, which in turn helps inform how likely they are to take collective action against a target. In order to help prevent future cybersecurity incidents the media could also, as argued by Rogers[10], take more responsible approach to the reporting of cybercriminals so as to avoid glamourizing individuals and setting them up as role models.

Impression management

As with any online relationship individuals may also engage in what is termed impression management, in which an individual may attempt to construct what they see to be a desirable image of themselves[11]. There can be several motivations behind impression management, including the desire to be liked and to appear competent, and of particular relevance perhaps to cyber adversaries, the desire to appear dangerous[12]. It has been

ISSN 2052-8604

noted that the depth to which individuals engage in online impression management is linked to how likely it is they think they will meet someone offline[13]. Given that those involved in the instigation of cybersecurity incidents are already motivated to conceal their identity due to the risk of being pursued by law enforcement it could be argued that such individuals are therefore particularly likely to engage in impression management. However, there are added complications to understanding the role of impression management within online collectives. One of the websites associated with the growth of Anonymous and other cyber adversaries[1], 4chan, operates on a principle of anonymity. Users are not generally able to identify themselves when posting content or comments, and indeed users who do attempt to bypass this restriction are often met with harsh criticism for doing so[4]. This prevents individuals from building up a personal reputation or seeking fame or leadership roles. From a social psychological perspective this system of interaction is surprising, particularly in Western cultures which are characterised by individualism as opposed to collectivism[14]. As such it is an area which could be argued to be unchartered territory for social psychologists, and one which needs to be researched in much greater depth.

Groups can also engage in forms of collective impression management. It has been claimed for example that Anonymous engaged in impression management by overstating their capabilities to journalists[4]. The group also used sophisticated impression management techniques when targeting the Church of Scientology. The ‘Message to Scientology’ video that was posted on YouTube by the group stressed the severity of the threat they posed and how likely it would be that they would successfully shut down the Church of Scientology. This is consistent with Protection Motivation Theory[15], which states that individuals decide how to respond to a threat based on how severe that threat is perceived to be and how vulnerable they perceive themselves to be. The message also claimed that attempts to counter the actions of the group would be ineffectual. This reflects work into fear appeals that suggests that people are less likely to take action to protect themselves if they do not believe that they have the ability to do so[16]. Finally the group also made use of expectation management. It is stated in the video that they realise that they will not bring about an end to Scientology overnight, adding credibility to their claims of what they will achieve. Combined with the ominous background music and the voice synthesised narration the overall effect is a psychologically sophisticated video which aims to intimidate the opponent.

Linked to impression management is doxing, which refers to revealing an individual’s real life identity, as well as possibly personal contact information such as their home address. The act of doxing someone is used as a weapon within these online communities that are based on anonymous participation[4]. The effort that is put into doxing another individual can be extensive, and in some cases involving collectives associated with cyber security incidents stems from intra-group conflict about the ideology, group identity and actions of the group[4]. Doxing raises a number of interesting and challenging questions from a psychological perspective. There can obviously be a number of real life consequences of

ISSN 2052-8604

being doxed, such as being targeted at home or being pursued by law enforcement. Yet there are also potentially psychological consequences. In offline forms of conflict a common goal can be dehumanise and depersonalise an opponent, such as for example in the oppression of dissidents in dictatorships[17]. In the case of doxing however the opposite is achieved, with the target's offline identity revealed. When this happens the person effectively has their ability to engage in impression management severely curtailed, since they no longer have the ability to control and alter what information about their identity they want to be disseminated. In light of the way in which the internet allows people to create an alternative identity it can be seen why robbing someone of this ability is perceived as one of the worst possible actions in some online communities. When planning how to dissuade cyber adversaries it may be that highlighting the risk of being doxed could be an effective strategy. It may be the case that the potential loss of an online identity is so threatening to an individual that this is an effective strategy even when there is no possibility of the legal action being taken against the individual. Of course, this approach could raise a number of ethical questions.

Motivation

Group processes and impression management may determine the characteristics of a group or hacktivist collective and how they present themselves to society, but they do not in themselves predict the actions of the group. For this an understanding of motivation is needed. There are obvious financial motivations to cybercrime, but the reasons behind other cybersecurity incidents are not as apparent. Cyberwarfare, hacktivism and online social protest can all produce similar results and are not always easily to differentiate from one another. It can also be difficult to predict what will drive a group to move towards acts that are focussed on social protest. As has been commented the change in Anonymous from a group that was characterised by random actions and anarchy to one that engaged in active social protest and aided in supporting political protestors around the world was highly unexpected[1]. A better understanding of the motivations of those involved in these activities may be useful in distinguishing between cybercrime and online social protest, as well anticipating future actions. Alberici et al[18] argue that there are four motivations that drive people to collective action:

- Identification with a group which is involved in a conflict with a larger organisation
- Negative emotions arising from perception that the situations of one's own group is unfair
- A shared belief that through joint efforts the group will be able to achieve its goals
- The perception that core moral principles have been violated and that these must be defended and reinstated

These motivations would appear to be consistent with a number of cybersecurity incidents that could be termed hacktivism or online social protest. They may also be useful in developing a productive dialogue with online adversaries as to why an organisation is being targeted, and what actions might be taken to resolve the conflict between the

ISSN 2052-8604

adversaries and the target. This is not an approach that has been adopted particularly often in the past. Instead organisations such as the Church of Scientology have responded to situations involving cyber adversaries with a more confrontational approach[4], which could be argued to have fuelled further action by the cyber adversaries by reinforcing the motivations of the type identified by Alberici et al[18]. Referring back to the topic of reactance discussed above it has been noted that reactance is particularly pronounced when there is a perceived threat to personal freedom, which is known as the boomerang effect[19]. This fits with the motivations identified by Alberici et al[18], particularly if the freedom of information is viewed by the individual as being a core moral principle.

When viewing interviews of members of Anonymous and similar online groups one common theme appears to be a sense of anger[20]. At times this is directed towards specific organisations such as the aforementioned Church of Scientology, at other times it appears to be a more diffuse sense of anger towards society in general. Whether or not the actions against an organisation are morally acceptable or not is a matter of the perspective of the individual. Whilst Anonymous have been implicated in cybersecurity incidents involving apparently random targets they have also taken part in actions such as providing internet access to protestors in Tunisia during the 2011 uprising[4], after the Tunisian government attempted to block all internet traffic within the country. Examples such as this suggest that there is more to some cybersecurity incidents than simply financial gain or criminal intent. Whilst the insights of forensic psychology and criminology will undoubtedly continue to be of great importance to the field of cybersecurity there is a need to better understand the social context and social psychology of cyber adversaries. One particular psychological phenomena which may be of relevance is cognitive dissonance[21], which refers to the tendency of people to avoid holding contradictory views or attitudes. By focussing on the greater good of battling perceived social injustice members of online groups may be able to justify to themselves the act of committing criminal acts. If this is the case then attempts to dissuade individuals from acting as cyber adversaries by highlighting the criminality of their behaviour may not be effective, as the individual has already processed and discounted that information.

There would though appear to be overlap between a genuine desire to achieve social change and to acting only for personal enjoyment, or for the lulz to use the language of some online groups. As previously commented this difference in the motivations of individuals has been a source of intra-group conflict[4], with disagreements over what the ideology and goals of the group should be. If as previously discussed individuals do derive their sense of self-esteem and identity from membership of such groups then it is understandable that a lack of agreement on the purpose of that group could lead to conflict. Attempts to deliberately create conflict within groups by provoking discussions around the goals of the group also appear to be evident within the communications of some groups. This could be trolling behaviours by individuals, or it may be more organised and deliberate efforts by other groups to create tensions. As has been

ISSN 2052-8604

observed a limitation of research into information security behaviours of end-users is a lack of understanding of the social context in which these end-users operate[22] – the same comment could perhaps be applied to cyber adversaries.

Future directions

It has been argued in this paper that a better understanding of the social psychological processes behind cybersecurity incidents will help inform prevention and mitigation approaches. However it has to be acknowledged that social psychological processes are not merely something which act upon cyber adversaries. As evident in many cybersecurity incidents cyber adversaries actively use social psychological principles in the form of social engineering as a tool with which to gain access to secure systems[23]. There are numerous examples of those who are extremely skilled social engineers and books on the topic of how to apply social engineering principles[23], although it could be commented that much of this is based on anecdotal evidence, case studies and observational research. There is less work which has investigated social engineering using an experimental approach. This could be a reflection of the challenges inherent in securing ethical approval for studies that use deception or other forms of participant manipulation. Similarly studies into security behaviour often rely on measurements of intended future behaviours, rather than the actual behaviours themselves[22]. Direct observation of behaviour can lead to demand characteristics such as the Hawthorne effect, in which research participants alter their behaviour simply due to the fact that they know they are being observed by researchers. To avoid these effects it may be necessary to observe participants covertly, which can prompt ethical questions around informed consent.

Social psychologists may be able to aide in these methodological and ethical challenges. Deception and manipulation are part of many psychology studies, and as such the field has developed extensive guidelines on how these issues should be addressed[24]. Indeed, many of the ethical approval processes used in the UK could be said to have stemmed from psychological research, particularly those relating to the potential for psychological harm to participants. Being able to demonstrate that planned research is consistent with the recommendations of the British Psychological Society, the professional accreditation body for Chartered Psychologists, may help facilitate approval at the institutional level. The need to understand the psychological impact of social engineering has also been an unintended consequence of attempts to incorporate social engineering into `ethical hacking' methodologies. For example, Dimkov et al.[25] found that debriefing deceived security staff on social engineering tests was more stressful than carrying out the test itself. The risk was not identified when planning the test, despite the fact their methodology had been warranted as ethically sound.

One area of particular relevance to the understanding of social engineering is social marketing, which represents the interface between social psychology and consumer

ISSN 2052-8604

psychology. As in what could be termed commercial marketing the goal of social marketing is to bring about change, although for a social good rather than commercial profit. It has been noted that social marketing can be utilised to bring about behaviour change within organisations, specifically for cybersecurity related behaviours by end users. As Ashenden and Lawrence[22] comment simply raising awareness of security issues or changing attitudes, as has often been the goal of more traditional behaviour change strategies, does not necessarily result in behaviour change. Similarly the efficacy of attempts to modify cybersecurity behaviours through the use of fear appeals is inconsistent [7]. This has been the experience of psychologists working in the areas of health and social psychology[26], who have in turn also attempted to utilise social marketing to achieve long term behaviour change. The technique is related to the Nudge approach[27], which aims to encourage individuals towards sensible choices without actually removing options from them. Despite the adoption of the approach by a number of UK government bodies there are different views on how effective the Nudge approach actually is, although as has been observed both it and social marketing have the advantage of being relatively easily applied by those without expertise in social science[27].

It may be that by working jointly the fields of social psychology and cybersecurity are able to make a unique contribution to these types of approach. Social marketing is based largely upon the principles of commercial marketing, which were themselves informed by trial and error experience of what is successful and cost effective in the business world. Similarly it may be that further exploration of the experiences of social engineers could help inform better ways of implementing social marketing campaigns. Ultimately after all the goal of both companies and social engineers is to develop a relationship with the target and use this to prompt certain behaviours; just as for companies there are costs in terms of resources and potential risk to the social engineer if they misjudge how best to go about these activities. The experience of psychologists in conducting interviews on sensitive and potentially illegal activities could be used to complement the work already being undertaken in this field. Of course it must be commented that in light of the issue of ideology and reactance that have been discussed cyber adversaries who are experienced in social engineering may not be inclined overall to work with cybersecurity practitioners. However even with these differences there are areas where collaboration may occur. Despite the fact that websites such as 4chan are almost defined by the practice of producing the most shocking content possible it has been observed there is zero tolerance amongst users for child pornography, and indeed those attempting to obtain or disseminate child pornography material on the website often become the targets of social engineering based attempts by others users to identify and entrap them[4]. Using the experience of users who have applied social engineering to trick and deter paedophiles could aide cybersecurity practitioners and educators in the development of techniques to promote online safety in young people.

ISSN 2052-8604

There is also a need for a better understanding of the social context of cyber security. As conceptualised in Figure 1 all cybersecurity incidents occur within the context of wider society, which depending on the situation may occur at multiple levels from the local to the international. Researchers such as Holt have provided in-depth explorations of the behaviour of hackers, including less technologically skilled individuals such as script kiddies[28]. However as online technology becomes increasingly social in nature it could be argued that the social context of hackers may have widened. The users on 4chan and other sites who became involved in the online protests against the Church of Scientology were not all hackers, and may not have even been script kiddies. Yet they played a part in these protest through supporting those who did have the technological skills, and by taking part in the offline protests that continue today. Rogers[10] suggests that increasing contact between cybercriminals and more mainstream internet users may result in a change to the social environment that would discourage participation in cybersecurity incidents. This is consistent with the contact hypothesis from social psychology research, which suggests that contact between groups can reduce the conflict between them[29], particularly when cross-group friendships area created[30]. Indeed there is evidence that even asking people to imagine contact with another group can reduce intra-group hostility [31].

Social psychological research has demonstrated however that certain requirements must be met if this type of contact is to be effective. First, there must be a wider social climate which encourages integration between the opposing groups. Secondly, the contact must take place under conditions of equal social status. Finally, the contact must involve cooperation towards a shared goal. It is difficult to envisage how some of these principles could be applied to real life cybersecurity situation. For instance as discussed members of some online groups may derive their social identity and self-esteem from being members of a persecuted group that it is acting against a larger organisation, and therefore the engaging with another group under a sense of equal social status may not be consistent with their sense of group identity. Similarly organisations who have been a victim of cyber a cybersecurity incident may be unwilling to engage in a dialogue as equals. One way to start this dialogue could be to consider social psychological research that demonstrates that people often hold negative misperceptions about others, even their own peers[32]. Challenging these negative stereotypes and misperceptions and instead focussing on positive change has been found to be an effective form of behaviour change[32], perhaps because it is based on empowerment rather than fear appeals. An interesting comment is made by several of the participants in the documentary film *We Are Legion: The Story of the Hacktivists*[20], which is that they were surprised by the diversity of the people who attended the street protests against the Church of Scientology. To paraphrase one of the film participants these people were not all socially awkward male adolescents, as perhaps the stereotype would predict, but instead men and women of a range of backgrounds and of different ages.

ISSN 2052-8604

In conclusion there is potential for collaborative research in social psychology and cybersecurity to benefit both disciplines. Cybersecurity researchers and practitioners can aide social psychologists in accessing and understanding online social groups of a type which are vastly under-studied. The social dynamics of these groups may represent novel processes that could have paradigm shifting implications for the field of social psychology. Social psychologists can in turn provide cybersecurity experts with evidence based approaches on how to predict and if necessary attempt to mitigate group based cybersecurity incidents, as well as aiding in the methodological and ethical challenges inherent in studying some of the human factors of cybersecurity. Through such collaboration new ways of promoting online safety and empowering individuals to make informed decisions about their participation in cybersecurity incidents may be reached.

References

1. Coleman, G., *Hacker, Hoaxer, Whistleblower, Spy : The Many Faces Of Anonymous*. 2014, London: Verso.
2. Bartlett, J., *The Dark Net*. 2014: William Heinemann.
3. Fiske, S.T., *Social Beings: Core Motives In Social Psychology*. 2010, Hoboken, NJ: John Wiley & Sons.
4. Olson, P., *We Are Anonymous*. 2012, New York: Back Bay Books.
5. Bettencourt, B.A. and K. Sheldon, *Social roles as mechanisms for psychological need satisfaction within social groups.* Journal of Personality and Social Psychology, 2001. **81**(6): p. 1131-43.
6. Tormala, Z.L. and R.E. Petty, *What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty.* Journal of Personality and Social Psychology, 2002. **83**(6): p. 1298-1313.
7. Hewstone, M. and J.M.F. Jaspars, *Intergroup relations and attribution processes*, in *Social Identity And Intergroup Relations*, H. Tajfel, Editor. 1982, Cambridge University Press: Cambridge. p. 99 - 133.
8. Doise, W., *Groups And Individuals: Explanations In Social Psychology*. 1978, Cambridge: Cambridge University Press.
9. Marques, J.M., D. Abrams, and R.G. Serodio, *Being better by being right: Subjective group dynamics and derogation of in-group deviants when generic norms are undermined.* Journal of Personality and Social Psychology, 2001. **81**(3): p. 436-447.
10. Rogers, M.K., *The psyche of cybercriminals: A psycho-social perspective*, in *Cybercrimes: A Multidisciplinary Analysis*, G. Ghosh and E. Turrini, Editors. 2010.

ISSN 2052-8604

11. Goffman, E., *The Presentation of Self in Everyday Life*. 1959, New York: Anchor Books.
12. Jones, E.E. and T. Pittman, *Toward a general theory of strategic self-presentation*, in *Psychological Perspectives on the Self*, J. Suls, Editor. 1982, Erlbaum: Hillsdale, NJ.
13. Underwood, J.D.M., L. Kerlin, and L. Farrington-Flint, *The lies we tell and what they say about us: Using behavioural characteristics to explain Facebook activity.* Computers in Human Behavior, 2011. **27**(5): p. 1621-1626.
14. Hofstede, G.H., G.J. Hofstede, and M. Minkov, *Cultures And Organizations : Software Of The Mind : Intercultural Cooperation And Its Importance For Survival*. 3rd ed. 2010, New York: McGraw-Hill. xiv, 561 p.
15. Rogers, R.W., *A protection motivation theory of fear appeals and attitude change.* Journal of Psychology, 1975. **91**(1): p. 93.
16. Johnston, A.C. and M. Warkentin, *Fear appeals and information security behaviors: An empirical study.* Mis Quarterly, 2010. **34**(3): p. 549-566.
17. Haslam, S.A. and S. Reicher, *Debating the psychology of tyranny: Fundamental issues of theory, perspective and science - Response.* British Journal of Social Psychology, 2006. **45**: p. 55-63.
18. Alberici, I.A., et al., *Comparing social movements and political parties' activism: The psychosocial predictors of collective action and the role of the internet.* Contention, 2012. **0**(0).
19. Brehm, S. and J. Brehm, *Psychological Reactance: A Theory of Freedom and Control*. 1981, New York, NY: Academic Press.
20. Knappenberger, B., *We Are Legion: The Story of the Hacktivists*. 2012.
21. Festinger, L., *A Theory Of Cognitive Dissonance*. 1957, Evanston, Ill.,: Row. 291 p.
22. Ashenden, D. and D. Lawrence, *Can we sell security like soap?: a new approach to behaviour change*, in *Proceedings of the 2013 workshop on New security paradigms workshop*. 2013, ACM: Banff, Alberta, Canada. p. 87-94.
23. Hadnagy, C., *Social Engineering: The Act of Human Hacking*. 2011, Indianapolis: Wiley Publishing Inc.
24. British Psychological Association, *Code of Human Research Ethics*. 2010, British Psychological Society: Leicester.

ISSN 2052-8604

25. Dimkov, T., W. Pieters, and P.H. Hartel, *Two methodologies for physical penetration testing using social engineering*, in *Annual Computer Security Applications Conference*. 2010: Austin, Texas.

26. Foxcroft, D., et al., *Longer-term primary prevention for alcohol misuse in young people: A systematic review.* Addiction, 2003. **98**: p. 397 - 411.

27. Thaler, R.H. and C.R. Sunstein, *Nudge: Improving Decisions About Health, Wealth and Happiness*. 2009: Penguin.

28. Holt, T.J., *Examining the role of technology in the formation of deviant subcultures.* Social Science Computer Review, 2010. **28**(4): p. 466-481.

29. Allport, G., *The Nature of Prejudice*. 1954, Reading, MA.: Addison-Wesley.

30. Pettigrew, T.F. and L.R. Tropp, *A meta-analytic test of intergroup contact theory.* Journal of Personality and Social Psychology, 2006. **90**(5): p. 751-83.

31. Crisp, R.J. and R.N. Turner, *Can imagined interactions produce positive perceptions? Reducing prejudice through simulated social contact.* American Psychologist, 2009. **64**(4): p. 231-240.

32. McAlaney, J., B. Bewick, and C. Hughes, *The international development of the 'Social Norms' approach to drug education and prevention.* Drugs: Education, Prevention, and Policy, 2011. **18**(2): p. 81-89.

ISSN 2052-8604

The dilemma of cyber security and privacy: On the role of value sensitive design

Balbir S. Barn[1] and Ravinder Barn[2]
[1]Middlesex University
b.barn@mdx.ac.uk

[2]Royal Holloway, University of London
r.barn@rhul.ac.uk

Abstract

The design of systems that can address both security and privacy in the context of societal sustainability is a critical challenge and there is limited evidence of research activity that can draw upon the multi-disciplinary approaches that are required to address this concern. This paper examines the literature of value sensitive design and argues that it can be integrated within software practice so that a more nuanced explication of the so-called one dimensional continuum security and privacy can be better developed. The paper's key proposition is that the Quadrant framework developed by Conti et al. (2014) can be extended by value sensitive approaches so that security applications can be assessed and evaluated in a consistent manner. The approach presents opportunities for methods development in this area.

Introduction

Today's hyper-connected world is a paradox in that it creates both opportunities or benefits but also considerable risks for individuals and organisations. This in turn leads to a challenge in addressing the dilemma of security and privacy in cyber space.

Privacy is a notoriously difficult concept to define and in the context of the information society becomes doubly so. As Introna (1997) indicates, privacy both arises from and influences human autonomy derived from widespread prevalence of ICT. In this computing sense, certain aspects of privacy are particularly pertinent: privacy is a relational concept, it comes to the fore through interactions; claiming privacy is the right to

ISSN 2052-8604

limit or control access to a personal domain and finally privacy is a relative concept, it is a continuum in that total privacy may be as undesirable as total transparency.

The nation-state also concurs with this tempting and possibly simplistic view that security and privacy are opposite ends of a single continuum. It argues that trade-offs are both effective and essential for the overall well-being of society and this view is illustrated widely in public statements. In November 2014, Robert Harrigan, Head of GCHQ warned of the criminogenic properties of social networking sites such as Twitter and Facebook to argue that: "… privacy has never been an absolute right and the debate about this should not become a reason for postponing urgent and difficult decisions."

The above sentiment illustrates the contested nature of privacy in our new information society. It makes a claim that by offering security of information to citizens (through encryption of data), social networking sites are becoming havens for criminals and cyber-terrorists. Arguably, this perceived trade-off between security and privacy occurs in a society that is aspiring to be economically, environmentally and socially sustainable where privacy is very much a contested concept whose contested nature should not be bemoaned or derided but rather, seen as the tapestry of the sustainability agenda (McKenzie, 2004). In general, societal sustainability includes a focus on processes, systems and structures to support healthy and liveable communities. Societal sustainability, furthermore, requires interventions from computer science research. Here, an underlying assumption is that most sustainability issues require inter-connection and human interactions with systems - problems that are central to computing science research (Millett et al., 2012). The same report from the National Research Council of USA also noted that research in socio-technical systems is needed to encompass society, organisations and individuals and their behaviour as well as the technical infrastructure used rather than remain focused on a narrow environment related to definition of sustainability. This ultimately means addressing issues of security and privacy in the design of socio-technical systems.

Having set the scene, this paper puts forward a position that the dilemma of security versus privacy in the context of societal sustainability is essentially one of design. It contributes to the existing canon of literature that focuses on ICT design and information control (Borning and Muller, Van den Hoven, 2007). Systems that support security or maintain privacy are fundamentally socio-technical in the sense that information protection takes place using a combination of technological functionality and social

ISSN 2052-8604

practice and hence a deeper understanding of issues of privacy and security and their "trade-off" can only be possible if these (moral) values are surfaced, negotiated and agreed during the design of systems.

This position is elaborated in the paper in the following manner: Section 2 introduces the framework proposed by Conti et al (Conti et al., 2014). Section 3 examines current literature on values and the design of systems. Section 4 proposes how value sensitive design can help operationalize the framework outlined by Conti et al. By evaluating the limitations of the framework, extensions are suggested and the opportunities that are opened up are discussed. Finally we present concluding remarks that summarize our main position.

Quadrants of Security and Privacy

In a recent paper, Conti et al. (2014) argue that security and privacy trade-offs in a single continuum are too simplistic. One reason proposed is that any system implementation is not always effective hence it is not a single plane of discussion. They conceptualise the effectiveness of security solutions and their impact on privacy and civil liberties by a two-dimensional graph resulting in four quadrants. Here we present an overview of the framework and provide new examples of solutions that fall into each quadrant that are used to illustrate the framework in Figure 1a.

Quadrant I: In this quadrant, privacy is degraded while no real gain in security is achieved. Examples include: password recovery systems based on easily guessed personal information. Facebook and its ongoing adjustments to the privacy needs of its users is also an example.

Quadrant II: In this quadrant, security is improved while, privacy is degraded. Examples include: the long-term retention on-line activity by ISPs and much of what has been exposed as a result of the revelations of Edward Snowden.

Quadrant III: In this quadrant, security is reduced while privacy is achieved. This quadrant will not typically include security solutions but policy statements or court judgments may cause transition into this quadrant. The recent update of the Fourth Amendment by the United States Supreme Court as a result of the Riley vs California, and USA vs Wuries cases is an example. Here, the Supreme Court Justices concluded

ISSN 2052-8604

that searches of digital data contained on smart phones are not “reasonable” in the absence of a warrant and constitutes a major intrusion of privacy[12].

Quadrant IV: In this quadrant, optimal solutions improve both security and privacy. Such solutions have the potential to contribute to societal sustainability as they will ultimately engender trust in the processes, systems and structures to support healthy and liveable communities. Examples include: the secure sockets layer for encryption on the internet supporting all sensitive transactions such as finance. Most recently, major corporations such as Apple and Google have sought to rebuild users’ trust by the introduction of additional encryption on cloud services as a response to the Snowden revelations. Conversely, the knowledge that encryption was being violated by State vectors have moved these systems potentially back into the Quadrant III[13].

Designs of systems for Quadrant IV are clearly a desirable position and as Conti et al. (2014) indicate, such solutions begin with proper communication, planning and education. Furthermore, organizations tasked with development of systems should create opportunities for dialogue amongst the systems designers, end-users and experts in privacy and civil liberties. This is right. However, the dominant issue is how to embed this practice into existing software engineering processes underpinning potentially security solutions. Hence, methodological innovations are required. Moreover, we would contend that all systems that are part of the social fabric that contribute to societal sustainability should adhere to design approaches that lead to Quadrant IV outcomes. Further, the role of information systems especially in their new guise of “apps” delivered through sensor rich smartphones is particularly pertinent. Organizations responsible for these apps should recognize that their corporate actions with respect to design and deployment of such systems have a profound impact on all aspects of societal welfare.

The next section reviews current software engineering practice that attempts to address these concerns of privacy and security. Our lens for examining this literature places security and privacy as ultimately human moral concerns exemplified as values.

[12] http://www.supremecourt.gov/opinions/13pdf/13-132_8l9c.pdf

[13] (http://www.theguardian.com/world/2013/sep/05/nsa-gchq-encryption-codes-security

ISSN 2052-8604

Value Sensitive Design

The concept of *value* has been investigated in its role to systems design. Friedman (1996) is generally credited with introducing the study of values to IS practice through the Value-sensitive design (VSD. Value is defined as: *what a person or group of people consider important in life*. Values that are particularly pertinent to information systems include: ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, identity and others. In systems design, *values* have, to-date, been integrated mostly with participatory design approaches (Bjerknes et al., 1987) or more recently *Co-Design*. Co-design involves potential (un-trained) end users working jointly with researchers and designers to jointly create artifacts that lead directly to the end product (Sanders, 2000) and, as Yoo et al. (2013) state this has ``become a dominant user study methodology in the fields of product design, service design, interaction design and HCI (Muller, 2003)". Several researchers have commented that whilst participatory design is a dominant mode of technological design, end-users still struggle to influence the direction of design within the participatory process. Furthermore, end-users may not fully understand the ecology of available technologies. It may be that the reductionist principles of software engineering could be argued to have hindered the integration of values approaches into mainstream practice and so making it harder to monitor such concerns.

Value-sensitive design (VSD) emerged to integrate moral values (and more broadly ethics) with the design of systems. A key premise of VSD is that it seeks to design technology that accounts for human values throughout the design process (over and beyond the identification of functionality and visual appearance) of systems.

Thus VSD has a stated goal that there should be freedom from bias in systems. That is: computer systems should not systematically and unfairly discriminate against certain individuals or groups of individuals in favour of others (Friedman et al., 2006). VSD has developed both methods and theory that incorporate particular values into technologies through conceptual, empirical, and technical investigations.

In software engineering, Goguen's latter work in requirements engineering has emerged from formal computer science but presents a similar discourse on the importance of values that could be seen as a plea to move from a reductionist principle to one that proposes ``semi-formal approaches that take account of social processes can be valuable. Values are the key to unlocking ...the enormous dangers of contemporary

ISSN 2052-8604

technologies" (Goguen, 2003). Goguen recognized that there are limitations to formalisms but also noted that ethno-methodologies present challenges in collection of data and the subsequent grounding of data in the design of the target technical artefact. Thus ``methods based on ethnomethodology cannot be applied directly to systems that have not yet been built''. We are aiming for an approach to managing values sensitive issues in requirements elicitation that supports rich contextual information but is sufficiently formal to contribute to the building of the technical artefact.

A Proposition on using Value Sensitive Design for elaborating security and privacy

How does value sensitive design get properly incorporated into software design processes? Our investigations suggest several lines of research: Value Sensitive Design as a Language Model and Value Sensitive Co-Design discussed below.

Value Sensitive Co-Design

Building on the work of Yoo et al. (2013), we have developed a co-design workshop method that exposes value concerns such as security and privacy. The approach has been evaluated on a research project (Mobile Apps for Youth Offending Teams – MAYOT) aimed at developing social technology for use by young people and their caseworkers in youth offending teams in the UK. The project raised requirements on design methods to incorporate the voice of stakeholders with respect to privacy issues. A key outcome of the method is how gaps in values from the perspectives of different stakeholders is managed through the co-design process. Resolution of value gaps are then possible to address. Although the co-design approach is highly interactive we have provided formalisms for expressing the resolution of value gaps (Barn et al., 2014).

ISSN 2052-8604

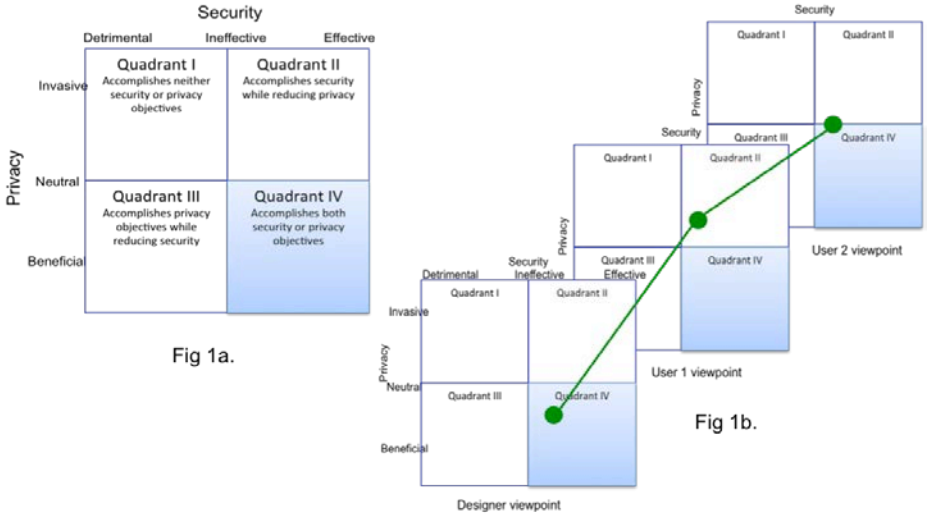


FIGURE 1A AND 1B: QUADRANTS FOR THE EXCLUSION ZONE FEATURE

Our approach to value sensitive design via semi-formal modelling provides an additional critique of the Quadrants discussed in Section 2. The quadrants can be viewed as a state machine where states are pairs (E.g. Priv-Beneficial/Sec-Detrimental = Quadrant 1, Priv-Beneficial/Sec-Effective = Quadrant 2 etc.). Viewing the framework as a state machine presents opportunities to understand what events/triggers can lead a security solution or a system to undergo a transition from one quadrant to another. Examination of the examples presented earlier suggests that a key function of a transition is the notion of trust. A trust function at a given threshold causes a solution to move from one quadrant to the other.

A second aspect of value sensitive design is the need to incorporate views from more than one stakeholder. Any given stakeholder may have a perspective that is represented by one positioning of a solution in the quadrant framework. Thus multiple perspectives of multiple stakeholders need to combined and aggregated by some function. What are the rules of such a function? We conceptualise multiple quadrant views by the representation in figure 1b.

Figure 1b illustrates an example of how the quadrant framework could have been used to assess a particular feature of our social technology developed from the MAYOT project. The Exclusion Zone feature is a function that is available on the MAYOT app that allows a case worker to define

ISSN 2052-8604

geographic region from which a young person is prohibited (with a potential risks to violating their youth order with obvious detrimental effects). The feature alerts the young person in possession of the smart phone hosting the app that they are in an exclusion zone. The feature went through a number of design changes representing the designer's view, a case worker's view (user 2), and a young person's view (user 1) who believed that the prototype app feature violated their privacy. Such concerns were considered seriously by the research team to ensure that a balance was struck between relative privacy and security of information.

In this way, the quadrant framework provides a powerful visual mechanism for looking at individual features of a security application and the overall summary of an application.

Conclusion

In this paper, we have proposed that a simplistic continuum from security to privacy is not sufficient for articulating the tensions in the design process of systems that have aspects of security and privacy to consider. While the Quadrant framework proposed by Conti et al goes some way to addressing this concern it does not provide a mechanism for designing systems that can be target to the ideal win-win Quadrant IV. By drawing upon the literature of Value Sensitive Design, we propose how to use the Quadrant Framework as a means of assessing the design of values such as privacy and security. Our view is that this is a viable position and we are currently working on further evidence and evaluation that supports this position.

References

BARN, B. S., BARN, R. & PRIMIERO, G. 2014. The Role of Resilience and Value for Effective Co-design of Information Systems. *Proceedings of the Conference of the International Association for Computing and Philosophy (IACAP 14).* Greece: To appear in: Synthese Library, Springer.

BJERKNES, G., EHN, P., KYNG, M. & NYGAARD, K. 1987. *Computers and democracy: A Scandinavian challenge*, Gower Pub Co.

BORNING, A. & MULLER, M. Next steps for value sensitive design. 2012. ACM, 1125-1134.

CONTI, G., SHAY, L. & HARTZOG, W. 2014. Deconstructing the Relationship Between Privacy and Security [Viewpoint]. *Technology and Society Magazine, IEEE,* 33**,** 28-30 %@ 0278-0097.

FRIEDMAN, B. 1996. Value-sensitive design. *Interactions,* 3**,** 16--23.

ISSN 2052-8604

FRIEDMAN, B., KAHN JR, P. H. & BORNING, A. 2006. Value sensitive design and information systems. *Human-computer interaction in management information systems: Foundations,* 5**,** 348--372.

GOGUEN, J. A. 2003. Semiotics, compassion and value-centered design. *Virtual, Distributed and Flexible Organisations: Studies in Organisational Semiotics, Reading, UK***,** 11--12.

INTRONA, L. D. 1997. Privacy and the computer: why we need privacy in the information society. *Metaphilosophy,* 28**,** 259-275 %@ 1467-9973.

MCKENZIE, S. 2004. *Social sustainability: towards some definitions*, Hawke Research Institute, University of South Australia.

MILLETT, L. I., ESTRIN, D. L. & OTHERS 2012. *Computing Research for Sustainability*, National Academies Press.

MULLER, M. J. 2003. Participatory design: the third space in HCI. *Human-computer interaction: Development process***,** 165--185.

SANDERS, E.-N. 2000. Generative tools for co-designing. *Collaborative design.* Springer.

VAN DEN HOVEN, J. 2007. ICT and value sensitive design. *The information society: Innovation, legitimacy, ethics and democracy in honor of Professor Jacques Berleur SJ.* Springer.

ISSN 2052-8604

Cyber Security Awareness Campaigns: Why do they fail to change behaviour?

Maria Bada[1], Angela M. Sasse[2] and Jason R.C. Nurse[3]
[1] Global Cyber Security Capacity Centre, University of Oxford
maria.bada@cs.ox.ac.uk

[2] Department of Computer Science, University College London
a.sasse@cs.ucl.ac.uk

[3] Cyber Security Centre, Department of Computer Science, University of Oxford
jason.nurse@cs.ox.ac.uk

Abstract

The present paper focuses on Cyber Security Awareness Campaigns, and aims to identify key factors regarding security which may lead them to failing to appropriately change people's behaviour. Past and current efforts to improve information-security practices and promote a sustainable society have not had the desired impact. It is important therefore to critically reflect on the challenges involved in improving information-security behaviours for citizens, consumers and employees. In particular, our work considers these challenges from a Psychology perspective, as we believe that understanding how people perceive risks is critical to creating effective awareness campaigns. Changing behaviour requires more than providing information about risks and reactive behaviours – firstly, people must be able to understand and apply the advice, and secondly, they must be motivated and willing to do so – and the latter requires changes to attitudes and intentions. These antecedents of behaviour change are identified in several psychological models of behaviour. We review the suitability of persuasion techniques, including the widely used 'fear appeals'. From this range of literature, we extract essential components for an awareness campaign as well as factors which can lead to a campaign's success or failure. Finally, we present examples of existing awareness campaigns in different cultures (the UK and Africa) and reflect on these.

Introduction

Governments and commercial organizations around the globe make extensive use of Information and Communications Technologies (ICT), and as a result, their security is of

ISSN 2052-8604

utmost importance. To achieve this, they deploy technical security measures, and develop security policies that specify the 'correct' behaviour of employees, consumers and citizens. Unfortunately, many individuals do not comply with specified policies or expected behaviours [1]. There are many potential reasons for this, but two of the most compelling are that people are not aware of (or do not perceive) the risks or, they do not know (or fully understand) the 'correct' behaviour.

The primary purpose of cyber security-awareness campaigns is to influence the adoption of secure behaviour online. However, effective influencing requires more than simply informing people about what they should and should not do: they need, first of all, to accept that the information is relevant, secondly, understand how they ought to respond, and thirdly, be willing to do this in the face of many other demands [2][3].

This paper engages in a focused review of current literature and applying psychological theories to awareness and behaviour in the area of cyber security. Our aim is to take a first step towards a better understanding of the reasons why changing cyber security behaviour is such a challenge. The study also identifies many psychological theories of behavioural change that can be used to make information security awareness methods significantly more effective.

This paper is structured as followed. Section 2 reviews current information on security-awareness campaigns and their effectiveness. In Section 3, we examine the factors influencing change in online behaviour, such as personal, social and environmental factors. Section 4 reflects on persuasion techniques used to influence behaviour and encourage individuals to adopt better practices online. In Section 5, we summarise the essential components for a successful cyber security awareness campaign, and consequently, factors which can lead to a campaign's failure. Finally, Section 6, presents examples of existing awareness campaigns in the UK and Africa and initially reviews them in the light of our study's findings.

Cyber security awareness campaigns

An awareness and training program is crucial, in that, it is the vehicle for disseminating information that all users (employees, consumers and citizens, including managers) need. In the case of an Information Technology (IT) security program, it is the typical means used to communicate security requirements and appropriate behaviour. An awareness and training program can be effective, if the material is interesting, current and simple enough to be followed. Any presentation that 'feels' impersonal and too general as to apply to the intended audience, will be treated by users as just another obligatory session [4].

Security awareness is defined in NIST Special Publication 800-16 [4] as follows: "*Awareness is not training. The purpose of awareness presentations is simply to focus attention on security. Awareness presentations are intended to allow individuals to*

ISSN 2052-8604

recognize IT security concerns and respond accordingly". This clearly highlights where the main emphasis on awareness should be. It identifies the fact that people need not only to be aware of possible cyber risks but also, behave accordingly.

In terms of the public more generally, governments encourage citizens to transact online and dispense advice on how to do so securely. However, major cyber security attacks continue to occur [5]. Although a likely reason for this could be the fact that attackers are becoming more skilled, there is also the reality that security interfaces are often too difficult for the layman to use.

Another relevant point that has arisen from the literature is the fact that people know the answer to awareness questions, but they do not act accordingly to their real life [6]. It is proposed that it is essential for security and privacy practices to be designed into a system from the very beginning. A system that is too difficult to use will eventually lead to users making mistakes and avoiding security altogether [7]. This was the case in 1999 [8] and is still the case today [9].

The fact today is that security awareness as conceived is not working. Naturally, an individual that is faced with so many ambiguous warnings and complicated advice, may be tempted to abandon all efforts for protection, and not worry about any danger. Threatening or intimidating security messages are not particularly effective, especially because they increase stress to such an extent that the individual may even be repulsed or deny the existence of the need for any security decision.

Factors influencing change in online behaviour

The increased availability of information has significant positive effects, but simply providing information often has surprisingly modest and sometimes unintended impacts when it attempts to change individuals' behaviour [10]. A considerable amount of investment is being spent by governments and companies on influencing behaviour online [11], and the success in doing so would be maximised if they draw on robust evidence of how people actually behave.

Various research articles have investigated the factors which influence human behaviour and behaviour change but one of the most complete is the Dolan, et al. [12]. In their article the authors present nine critical factors, namely: (1) the messenger (who communicates information); (2) incentives (our responses to incentives are shaped by predictable mental short cuts, such as strongly avoiding losses); (3) norms (how others strongly influence us); (4) defaults (we follow pre-set options); (5) salience (what is relevant to us usually draws our attention); (6) priming (our acts are often influenced by sub-conscious cues); (7) affect (emotional associations can powerfully shape our actions); (8) commitments (we seek to be consistent with our public promises, and reciprocate acts); (9) ego (we act in ways that make us feel better about ourselves).

ISSN 2052-8604

These factors hint at the key ingredients for an overall approach to influencing behaviour change, since the psychological mechanisms which they refer to, are core in making any type of decision. Furthermore, these mechanisms can influence the user's motivation to actually adopt the knowledge offered by a security campaign and behave accordingly. In order to enact change, the current sources of influence (conscious or unconscious, personal, environmental or social) need to be identified. The following section describes these aspects.

Personal factors

Reflecting on literature, it is well recognised that an individual's knowledge, skills and understanding of cyber security as well as their experiences, perceptions, attitudes and beliefs are the main influencers of their behaviour [13]. Of these, personal motivation and personal ability, are two of the most powerful sources of influence. Specifically, it is the difference between what people say and what people do that needs to be addressed. In many cases, people will have to overcome existing thought patterns in order to form new habits.

People can sometimes get tired of security procedures and processes, especially if they perceive security as an obstacle, preventing them from their primary task (e.g., being blocked from visiting a music download website because the browser has stated that the site might have malware). It can also be stressful to remain at a high level of vigilance and security awareness. These feelings describe the so called 'security fatigue', and they can be hazardous to the overall health of an organization or society [14][15].

In the security domain, the so called 'Security, Functionality and Usability Triangle', describes the situation of trying to create a balance between three, usually conflicting, goals [16]. If you start in the middle and move toward security, you also move further away from functionality and usability. Move the point toward usability, and you are moving away from security and functionality. If the triangle leans too far in either direction, then this can lead to a super secure system that no one can use, or an insecure system that everyone can use (even unwanted individuals, such as hackers). Security fatigue becomes an issue when the triangle swings too far to the security side and the requirements are too much for the users to handle. Therefore, there has to be a balance between system security and usability [9].

Moreover, perceived control is a core construct that can also be considered as an aspect of empowerment [17]. It refers to the amount of control that people feel they have, as opposed to the amount of their actual control [18][19][20]. The positive effects of perceived control mainly appear in situations where the individuals can improve their condition through their own efforts. Also, the greater the actual threat, the greater the value that perceived control can play. When we apply this theory to cyber security, we could assume that home-computer users often experience high levels of actual control

ISSN 2052-8604

over their risk exposure. This is because they can choose which websites to visit, whether to open email attachments and whether to apply system updates [21].

In Psychology, the Regulatory Focus theory [22] proposes that in a promotion-focused mode of self-regulation, individuals' behaviours are guided by a need for nurturance, the desire to bring oneself into alignment with one's ideal self ('ideal self' is what usually motivates individuals to change), and the striving to attain gains. In a prevention-focused mode of self-regulation individual's behaviours are guided by a need for security, the need to align one's actual self with one's ought self by fulfilling duties and obligations and the striving to ensure non-losses. Thus, the effectiveness of advertising campaigns for adolescents may be enhanced either by using two types of messages (prevention and promotion focused) or by priming one type of regulatory focus through the advertising vehicle.

Cultural and environmental factors

Culture is also an important factor that can have a positive security influence to the persuasion process. Messages and advertisements are usually preferred when they match the cultural theme of the message recipient. As a result, cultural factors are one of the most important factors for consideration when designing education and awareness messages [23].

The cultural systems of a society shape a variety of their psychological processes. Intrinsically motivated behaviours emanate from the self and are marked by the enjoyment and satisfaction of engaging in an activity. Conversely, extrinsic motivation refers to motivation to engage in an activity in order to achieve some instrumental end, such as earning a reward or avoiding a punishment. Messages tend to be more persuasive when there is a match between the recipient's cognitive, affective or motivational characteristics and the content of framing of the message. Also, messages are more persuasive if they match an individual's ought or self-guides, or self-monitoring style [24]. People might be motivated to follow a cyber security campaign's advice. But if that causes certain limitation on the sites they can visit online, then this can automatically result in emotional discomfort, thus leading to ignorance of a suggested 'secure' behaviour.

Perception of risk can be a collective phenomenon and it is crucial for awareness raising specialists to be aware of the different cultural characteristics. The values that distinguish country cultures from each other could be categorised into four groups [25]: (1) Power Distance; (2) Individualism versus Collectivism; (3) Masculinity versus Femininity; and (4) Uncertainty Avoidance. In more individualistic cultures, such as the West, people tend to define themselves in terms of their internal attributes such as goals, preferences and attitudes. For example, in cyber security, a message used in a Western country would tend to avoid presenting the general risks of not being secure online and rather focus on the benefits of being secure.

ISSN 2052-8604

In more collectivist cultures, such as those typically found in the East, individuals tend to define themselves in terms of their relationships and social group memberships [26]. In this cultural context, individuals tend to avoid behaviours that cause social disruptions. Therefore, they favour prevention over promotion strategies focusing on the negative outcomes, which they hope to avoid rather than the positive outcomes they hope to approach [27]. Moreover, risk is also seen as the other side of trust and confidence, a perception being imbedded in social relations [28]. The emphasis on different risks, in different cultural contexts is another important aspect that needs to be addressed when creating cyber security awareness campaigns.

Persuasion techniques

Persuasion can be defined as the "*attempt to change attitudes or behaviors or both (without using coercion or deception)"* [29]. There are two ways of thinking about changing behaviour: (1) by influencing what people consciously think about (rational or cognitive model) and (2) by shaping behaviour focused on the more automatic processes of judgment and influence (context model) without changing the thinking. In this section we present the different persuasion and influence techniques, in an effort to examine potential challenges in the area of cyber security awareness.

Influence strategies

People do not usually simply follow advice or instructions on how to behave online even if they come from an expert or a person of authority. In many cases, end users are not fully aware of the dangers of interacting online, and to exacerbate the issue, security experts provide them with too complicated information, often evoking emotions of fear and despair [30]. The basic persuasion techniques include: fear, humour, expertise, repetition, intensity, and scientific evidence.

People base their conscious decisions on whether they have the ability to do what is required and whether the effort is worth it. Examples of messages aimed at persuading individuals to change their behaviour online, can be found in advertising, public relations and advocacy. These 'persuaders' use a variety of techniques to seize attention, to establish credibility and trust, and to motivate action. These techniques are commonly referred to as the 'language of persuasion'. They can also be found in cyber security awareness campaigns. For example, fear is often being used as a persuasion technique for cyber security.

Surveys have shown that the invocation of fear can be a very persuasive tactic to specific situations, or indeed a counterproductive tactic in others [31]. Security-awareness campaigns mostly tend to use fear invocations, by combining messages with pictures of hackers in front of the screen of a computer. Even, the word 'cyberspace', indicates something unknown to many, thus leading to fear. Typically, invocations of fear, are

ISSN 2052-8604

accompanied with recommendations that are as efficacious in preventing the threat. Thus, the three central structures in fear invocations are fear, threat and efficacy.

Various behavioural theories including the Drive Model [32], the Parallel Response Model [33], or the Protection Motivation Theory [34], consider the cost and efficiency of a reaction and have independent effects on persuasion. According to the Protection Motivation Theory for instance, the way a person responds to and carries out a cyber security awareness campaign's recommendations depends on both the cyber threat appraisal but also on the person's self-efficacy.

The attempt to change a certain behaviour is much more difficult when the person is bombarded by a large number of messages about certain issues. However, even when the design of the message is taken into account, there is a big gap between the recognition of the threat and the manifestation of the desired behaviour at regular intervals. Specifically for security awareness campaigns, the behaviour that users will need to adopt, should be as simple and easy as possible highlighting the advantages of adopting it.

Moreover, findings suggest that interventions based on major theoretical knowledge to change behaviour (e.g., social learning theory or the theory of self-efficacy) that take into account cultural beliefs and attitudes, and are more likely to succeed [35].

Factors leading to success or failure of a cyber security awareness campaign

There are several components which need to be taken into consideration in order for an awareness campaign to be successful. One of the most crucial parts is that of *communication*. Teaching new skills effectively can lead to prevention of high-risk online behaviour, since what appears to be lack of motivation is sometimes really lack of ability [36].

There is a wide discussion about security-awareness campaigns and their effort to secure the human element, leading to a secure online behaviour. In many cases, security-awareness campaigns *demand a lot of effort and skills* from the public, while measures do not provide real insight on their success in changing behaviour. Often, solutions are not aligned to risks; neither progress nor value are measured; incorrect assumptions are made about people and their motivations; and unrealistic expectations are set [6].

As previously discussed *fear invocations* have often proved insufficient to change behaviour [31]. For example, a message combined to a photo of a hacker, might prove to be funny rather than frightening or might cause the public to feel not related to the advertisement.

ISSN 2052-8604

In order for a campaign to be successful, there are also several pitfalls which need to be avoided. The *first* is not understanding what security awareness really is. *Second*, a compliance awareness program does not necessarily equate to creating the desired behaviours. *Third*, usually there is lack of engaging and appropriate materials. *Fourth*, usually there is no illustration that awareness is a unique discipline. *Fifth*, there is no assessment of the awareness programmes [37]. *Sixth*, not arranging multiple training exercises but instead focusing on a specific topic or threat does not offer the overall training needed [38].

Perceived control and personal handling ability, the sense one has that he/she can drive specific behaviour, has been found to affect the intention of behaviour but also the real behaviour [18][19]. We suggest that a campaign should use simple consistent rules of behaviour that people can follow. This way, their perception of control will lead to better acceptance of the suggested behaviour.

Cultural differences in risk perceptions can also influence the maintenance of a particular way of life. Moreover, even when people are willing to change their behaviour, the process of learning a new behaviour needs to be supported [22][23]. We suggest that cultural differences should be taken into consideration while planning a cyber security awareness campaign.

Measuring the effectiveness of information security awareness efforts for the public though, can be a very complicated process. Metrics such as the number of phishing e-mails opened or number of accesses to unauthorised pages are difficult to measure in a larger scale. This is why, defined large scale metrics are needed, to help security-awareness efforts be evaluated and assessed.

The present paper has thus far, reviewed some of the various personal, social and environmental factors influencing online behaviour change as it relates to cyber security. Also, we have tried to identify the factors which can lead to a cyber security awareness campaign's success or failure.

Case studies

This section will present existing awareness campaigns on cyber security in the UK and in Africa. These two countries were selected in an effort to explore possible core cultural differences reflected in awareness efforts. The two countries differ not only regarding cultural characteristics, but also in the amount of investment being spent on influencing secure behaviour online.

Cyber security awareness campaigns in the UK

There are various awareness efforts in UK aiming to improve online security for businesses and the public. Below, we present two of the most popular of these.

ISSN 2052-8604

A) *The GetSafeOnline Campaign* [39] is a jointly-funded initiative between several government departments and the private sector, and focuses on users at home and in businesses. The positive message of ''*Get safe online*'' itself is an intriguing one, and at its core, emphasises to individuals that they have the responsibility for getting safe online. The campaign offers a comprehensive repository of information on threats and how-to advice on protecting oneself and one's enterprise. The charge, however, is on individuals to make use of this information and properly apply it to their context.

B) *The Cyber Streetwise Campaign* [40] also concentrates on users at home and in businesses. The new Home Office Cyber Streetwise site advises businesses to adopt five basic measures to boost their security. These include, using strong, memorable passwords, installing antivirus software on all work devices, checking privacy settings on social media, checking the security of online retailers before loading card details, and patching systems as soon as updates are available. This is a campaign which tries to cause a behavioural change by providing tips and advice on how to improve online security. The campaign uses a positive message method to influence the behaviour of users, ''*In short, the weakest links in the cyber security chain are you and me*''.

Cyber security awareness campaigns in Africa

A) *The ISC Africa* [41] is a coordinated, industry and community-wide effort to inform and educate Africa's citizens on safe and responsible use of computers and the Internet, so that the inherent risks can be minimised and consumer trust can be increased. The campaign uses a positive message method to influence the behaviour of users in a more collectivist approach ''*Working together to ensure a safe online environment for all*'. Here, we can see an obvious difference to the messages used in awareness campaign in the UK, that is, the cyber security-awareness efforts in Africa have been aligned to the cultural aspects of that society.

B) *Parents' Corner Campaign* [42] is intended to co-ordinate the work done by government, industry and civil society. Its objectives are to protect children, empower parents, educate children and create partnerships and collaboration amongst concerned stakeholders. Parents' Corner tips for a safer Internet include: ''*People aren't always who they say they are, Think before you post, Just as they would in real life - friends must protect friends''.* Once again, one of the main messages refer to users protecting users in terms of their relationships and social group memberships.

Comparing cyber security awareness campaigns in the UK and Africa

In our effort to investigate potential differences in cyber security-awareness campaigns, in different cultural contexts, we considered existing campaigns in the UK and Africa.

We have to state that there are a large number of existing national campaigns in the UK, but we selected two of the most popular of these. On the contrary, in Africa the number of

ISSN 2052-8604

existing awareness campaigns is limited. This difference could indicate lack of resources, or lack of current emphasis on cyber security in Africa. Moreover, it could even indicate that Africa has a more organised and coherent security-awareness plan, with a small number of targeted and coordinated campaigns.

As previously discussed, messages and advertisements are usually preferred when they match a cultural theme of the message recipient [23]. While reviewing the main messages used by campaigns in the UK, it became clear that most of them refer to the individual [25]. For example The *GetSafeOnline Campaign uses the message ''Get safe online''* by emphasising to individuals and their responsibility for getting safe online. On the contrary, the messages used by campaigns in Africa, refer to users in terms of their relationships and social group memberships, as well as the need to fulfil duties and obligations (Parents' Corner Campaign includes a message saying: *Just as they would in real life - friends must protect friends*).

The cultural aspects have been reflected in the awareness campaigns, in both cases, using a more individualist approach in UK and a more collectivist approach in Africa [23][25][27]. It is important to decide the target group of a campaign and try to match a cultural theme of the message recipient but also, match the recipient's cognitive, affective or motivational characteristics with the content of framing of the message [27][26][29].

Usually, most of official awareness-campaign sites include advice which usually comes from security experts and service providers, who monotonically repeat suggestions such as 'use strong passwords'. Such advice pushes responsibility and workload for issues that should be addressed by the service providers and product vendors onto users. One of the main reasons why users do not behave optimally is that security systems and policies are often poorly designed [9]. There is a need to move from awareness to tangible behaviours.

Another important aspect is that most of the official awareness-campaign sites in UK and Africa do not offer the possibility to users to call a help-line, not only to report cybercrime but also to receive help. Less skilled users could find this feature useful.

Conclusions

This paper presents a review of current literature based on the psychological theories of awareness and behaviour in the area of cyber security, and considers them to gain insight into the reasons why security-awareness campaigns often fail.

Simple transfer of knowledge about good practices in security is far from enough [6]. Knowledge and awareness is a prerequisite to change behaviour but not necessarily sufficient, and this is why it has to be implemented in conjunction with other influencing strategies. It is very important to embed positive cyber security behaviours, which can result to thinking becoming a habit, and a part of an organisation's cyber security culture. One of the main reasons why users do not behave optimally is that security systems and

ISSN 2052-8604

policies are poorly designed – this has been presented time and time again throughout research [9].

Behaviour change in a cyber security context could possibly be measured through risk reduction, but not through what people know, what they ignore or what they do not know. Answering questions correctly does not mean that the individual is motivated to behave according to the knowledge gained during an awareness programme. A campaign should use simple consistent rules of behaviour that people can follow. This way, people's perception of control will lead to better acceptance of the suggested behaviour [18][19][20].

Based on our review on the literature and analysis of several successful and unsuccessful security-awareness campaigns, we suggest that the following factors can be extremely helpful at enhancing the effectiveness of current and future campaigns: (1) security awareness has to be professionally prepared and organised in order to work; (2) invoking fear in people is not an effective tactic, since it could scare people who can least afford to take risks [30]; (3) security education has to be more than providing information to users – it needs to be targeted, actionable, doable and provide feedback; (4) once people are willing to change, training and continuous feedback is needed to sustain them through the change period; (5) emphasis is necessary on different cultural contexts and characteristics when creating cyber security-awareness campaigns [35].

In future work, we will aim to conduct a more substantial evaluation of several cyber security-awareness campaigns around the world, especially in North America and Asia, to examine the extent to which they have implemented the factors mentioned above and their levels of campaign success.

References

[1] Humaidi, N., Balakrishnan, V.: Exploratory Factor Analysis of User's Compliance Behaviour towards Health Information System's Security. J Health Med Inform 4(2), (2013).

[2] Rogers, R.W. Attitude change and information integration in fear appeals. Psychological Reports, 56, (1985) 183–188.

[3] Witte, K. Message and conceptual confounds in fear appeals: The role of threat, fear and efficacy. The Southern Communication Journal, 58(2), (1993) 147-155.

[4] National Institute of Standards and Technology - NIST: Building an Information Technology Security Awareness and Training Program. Wilson, M. and Hash, J. Computer Security Division Information Technology Laboratory. October 2003. http://csrc.nist.gov/publications/nistpubs/800-50/NIST-SP800-50.pdf

ISSN 2052-8604

[5] Kirlappos, I., Parkin, S., Sasse, M. A.: Learning from Shadow Security: Why understanding non-compliance provides the basis for effective security. *Workshop on Usable Security,* 2014.

[6] Information Security Forum (ISF): From Promoting Awareness to Embedding Behaviours, Secure by choice not by chance, February 2014. https://www.securityforum.org/shop/p-71-170

[7] Coventry, D.L., Briggs, P., Blythe, J., Tran, M.: Using behavioural insights to improve the public's use of cyber security best practices. Government Office for Science, London, UK, 2014. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/309652/14-835-cyber-security-behavioural-insights.pdf

[8] Whitten, A., Tygar, J.D.: SSYM'99 Proceedings of the 8th conference on USENIX Security Symposium - Volume 8, (1999) 14-14. USENIX Association Berkeley.

[9] Nurse, J.R.C., Creese, S., Goldsmith, M., Lamberts, K.: Guidelines for usable cybersecurity: Past and present, in The 3rd International Workshop on Cyberspace Safety and Security (CSS 2011) at The 5th International Conference on Network and System Security (NSS 2011), Milan, Italy, 6-8 September.

[10] Smith, M.S., Petty, E.R.: Message Framing and Persuasion: A Message Processing Analysis. Pers Soc Psychol Bull 22(3) (1996) 257-268.

[11] UK - Cabinet Office. The UK Cyber Security Strategy, Report on Progress and Forward Plans, December 2014.https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/386093/The_UK_Cyber_Security_Strategy_Report_on_Progress_and_Forward_Plans_-_De___.pdf

[12] Dolan P., Hallsworth, M., Halpern, D., King,D., Vlaev, I.: MINDSPACE Influencing behaviour through public policy, Institute for Government, Cabinet Office, (2010). http://www.instituteforgovernment.org.uk/sites/default/files/publications/MINDSPACE.pdf

[13] Hogan, J.: Motivation, In J.J. Bolhuis (Ed.), The behaviour of animals: Mechanisms, function and evolution (2005) 41-70. Malden, MA: Blackwell Publishing.

[14] O'Donnell A. How to Prevent IT 'Security Fatigue. http://netsecurity.about.com/od/advancedsecurity/a/How-To-Avoid-IT-Security-Fatigue.htm

[15] NTT Com Security Inc. Risk: Value Research Report 2014. https://www.nttcomsecurity.com/us/landingpages/risk-value-research/

[16] Waite, A. InfoSec Triads: Security/Functionality/Ease-of-Use. June 12, 2010.

ISSN 2052-8604

[17] Eklund, M., & Backstrom, M.: The role of perceived control for the perception of health by patients with persistent mental illness. *Scandinavian Journal of Occupational Therapy, 13,* (2006) 249-256.

[18] Bandura, A. Self-efficacy: The exercise of control. W.H. USA: Freeman and Company Pbl. (1997).

[19] Ajzen, I. Perceived Behavioral Control, Self-Efficacy, Locus of Control, and the Theory of Planned Behavior. Journal of Applied Social Psychology, 32, (2002) 665-683.

[20] Wallston, K.A. Control beliefs. In N. J. Smelser & P. B. Baltes (Eds.), International encyclopedia of the social and behavioral sciences. Oxford, UK: Elsevier Science (2001).

[21] More J.: Measuring Psychological Variables of Control, In Information Security, (2011). http://www.sans.org/reading-room/whitepapers/awareness/measuring-psychological-variables-control-information-security-33594

[22] Higgins, E.T.: Promotion and prevention: Regulatory focus as a motivational principle. *Advances in Experimental Social Psychology, 30* (1998) 1-46.

[23] Kreuter, M. W., & McClure, S. M.: The role of culture in health communication. *Annual Review of Public Health, 25*, (2004) 439-455.

[24] Uskul, A.K., Sherman, D.K. Fitzgibbon J.: The cultural congruency effect: Culture, regulatory focus, and the effectiveness of gain- vs. loss- framed health messages. *Journal of Experimental Social Psychology, 45(3),* (2009) 535-541.

[25] Hofstede, G., Hofstede, J.G., Minkov, M. Cultures and Organizations: Software of the Mind. 3rd Edition, McGraw-Hill USA, 2010.

[26] Triandis, H.C.: The self and social behaviour in differing cultural contexts. *Psychological Review, 96* (1989) 506-520.

[27] Lockwood, P., Marshall, T., & Sadler, P.: Promoting success or preventing failure: Cultural differences in motivation by positive and negative role models. *Personality and Social Psychology Bulletin, 31* (2005) 379-392.

[28] Dake, K. Myths of Nature: Culture and the Social Construction of Risk. Journal of Social Issues 48(4) (1992) 21–37.

[29] Fogg, B. J.: Persuasive Technology: Using Computers to Change What We Think and Do. Morgan Kaufmann (2002).

[30] Witte, K. Fear control and danger control: A test of the Extended Parallel Process Model (EPPM). Communication Monographs, 61, (1994) 113-134.

ISSN 2052-8604

[31] Ahluwalia, R.: An Examination of Psychological Processes Underlying Resistance to Persuasion. *Journal of Consumer Research, 27 (2)* (2000) 217-232.

[32] Janis, I.L. Effects of Fear Arousal on Attitude Change: Recent Developments in Theory and Experimental Research, in Advances in Experimental Social Psychology, Vol. 3, ed. Leonard Berkowitz, San Diego, CA: Academic Press, (1967) 166-224.

[33] Leventhal, H.: Findings and theory in the study of fear communications. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 5, pp. 119-186). New York: Academic Press (1970).

[34] Rogers, R.W. A protection motivation theory of fear appeals and attitude change. *Journal of Psychology, 91*, (1975) 93-114.

[35] Arthur, D., Quester, P.: Who's afraid of that ad? Applying segmentation to the Protection Motivation Model. *Psychology & Marketing, 21 (9)* (2004) 671-696.

[36] Winkler I. & Manke S.: 6 essential components for security awareness programs (2013). http://www.csoonline.com/article/2133971/strategic-planning-erm/6-essential-components-for-security-awareness-programs.html

[37] Khan, B., Alghathbar, S.K., Nabi, S.I., & Khan, M.K.: Effectiveness of information security awareness methods based on psychological theories. African Journal of Business Management 5 (26) (2011) 10862-10868. http://www.academicjournals.org/article/article1380536009_Khan%20et%20al.pdf

[38] Winkler I. & Manke S.: 7 Reasons for Security Awareness Failure, CSO Magazine, (2013). http://www.csoonline.com/article/2133408/network-security/the-7-elements-of-a-successful-security-awareness-program.html

[39] GetSafeOnline Campaign [Accessed online November 2014] www.getsafeonline.org

[40] The Cyber Streetwise campaign [Accessed online November 2014] www.cyberstreetwise.com

[41] ISC Africa [Accessed online November 2014] http://iscafrica.net/#home

[42] Parents corner [Accessed online November 2014] http://www.parentscorner.org.za/

ISSN 2052-8604

Sustainable
Society Network

ISSN 2052-8604

The Problem of the P3: Public-Private Partnerships in National Cyber Security Strategies

Madeline Carr[1] and Tom Crick[2]
[1]Department of International Politics, Aberystwyth University
mac64@aber.ac.uk

[2]Department of Computing & Information Systems,Cardiff Metropolitan University
tcrick@cardiffmet.ac.uk

Abstract

Cyber security is an emerging -- and increasingly high profile -- national policy concern; not only in terms of material vulnerabilities but also in terms of conceptualising security approaches. Many states, particularly Western democracies, have situated the `public-private partnership' (P3) at the centre of their national cyber security strategies. However, there has been a persistent ambiguity around this fundamental concept: policymakers regard the state as without the capability and also without the mandate to impose security requirements beyond government-owned systems; the private sector, however, is highly averse to accepting responsibility for national security and will fund cyber security only within the parameters of the profit/risk calculation appropriate for a shareholder-based arrangement. Amidst increasing suggestions that a market-led approach to cyber security has failed, a deeper exploration at the ideas and concepts behind this approach finds that a reliance on the P3 emerges from deeply held and shared beliefs about government legitimacy and private authority which may not be easily reconciled with wider national security issues for a modern digital economy.

Digital Economy
Transforming Business and Society

Sustainable Society Network

Introduction

Cyber security is emerging as one of the most challenging aspects of the information age for policymakers, technologists and scholars of international relations. It has implications for national security, the economy, human rights, civil liberties and international legal frameworks. Although politicians have been aware of the threats of cyber insecurity since the early years of Internet technology [1], anxiety about the difficulties in resolving or addressing them has increased rather than abated [2, 3, 4]. In response, governments have begun to develop national cyber security strategies to outline the way in which they intend to address cyber insecurity. In many states where critical infrastructure such as utilities, financial systems and transport have been privatised, these policies are heavily reliant upon what is referred to as the `public-private partnership' as a key mechanism through which to mitigate the threat. In the UK and US, the public-private partnership has repeatedly been referred to as the `cornerstone' or `hub' of cyber security strategy [1, 5, 6].

While public-private partnerships have often been developed as an appropriate means to address both non-traditional and traditional security threats [7, 8], in the context of cyber security this arrangement is uniquely problematic. There has been a persistent ambiguity with regard to any clear and agreed parameters for the partnership. The reticence of politicians to claim authority for the state to legislate tougher cyber security measures coupled with the private sector's aversion to accepting responsibility or liability for national security leaves the `partnership' without clear lines of responsibility or accountability. Questions are now being raised about the efficacy of a market-driven approach to cyber security, although any alternative in liberal democratic states has yet to emerge [2]. Crucially, questions arise here about the extent to which the state can be seen to be abdicating not just authority but responsibility for national security. As Dunn Cavelty and Suter [9] point out, `generating security for citizens is a core task of the state; therefore it is an extremely delicate matter for the government to pass on its responsibility in this area to the private sector'. Essentially, this raises questions about how well the state is equipped to provide national security in this context and about how existing policies and practices of national security are being challenged by this new threat conception.

This paper develops a comprehensive understanding of how policymakers and the private sector are conceptualising their respective roles in national cyber security, where there may be disparity in these conceptions and what implications this may have for national and international cyber security. The paper moves onto the analysis of the public-private partnership from the perspectives of both partners. It should be noted here that there is a round of interviews yet to be completed for this project which will contribute further to the analysis; what is presented here is the outcome of secondary research, with some discussion of the underpinning conceptual framework.

Analysis of the Public-Private Partnership in Cyber Security

ISSN 2052-8604

There are several reasons why cyber security, particularly in the context of critical infrastructure protection, has been conceived of as some kind of collaborative project for the public and private sectors. The state is understood to be responsible for the provision of security, especially national security. Critical infrastructure, those assets and systems necessary for the preservation of national security (broadly defined), is perceived as an integral part of providing security to the state [10]. The potential implications of a large-scale cyber attack on critical infrastructure are so extensive that it follows naturally that the government would recognise some authority and responsibility here. However, because most of the critical infrastructure in both the UK and US is privately owned and operated, by definition there has to be some kind of relationship between the public and private sector in terms of the provision of security in this context.

The public-private partnership is not of course, unique to cyber security. It has been employed widely by states like the UK and US as a mechanism to deal with a range of other issues including security related ones. The practice intensified from the 1990s when the privatisation of critical infrastructure was regarded as economically beneficial to the state, freeing up capital and relying more heavily on the efficiencies and business practices of the private sector. There is an extensive body of literature that has developed in the wake of this shift that examines the public-private partnership in all kinds of contexts [11, 12, 13]. It deals with the background of these partnerships, the range of different approaches, how to measure success and failure, and how responsibility and authority are delegated. There has also been some examination of the public-private partnership in cyber security, most notably by Dunn Cavelty and Suter [9], but this focuses on ways to improve it rather than critically analysing the political implications of it. Combined, this literature provides a solid foundation in highlighting the ways in which this partnership is distinct but also by outlining common assumptions and expectations that run through public-private partnerships more generally.

What is this public-private partnership?

It is necessary to be clear about what exactly is meant by the term public-private partnership in this particular context. Perhaps not unexpectedly, there is a huge range of diverse arrangements that are referred to as public-private partnerships, ranging from the joint provision of services with some government regulatory oversight (health sectors), to closely contracted outsourcing of large infrastructure projects, (building roads and bridges, the Olympics, etc). Much of the literature on public-private partnerships revolves around identifying and classifying partnership arrangements. This often takes place within a framework

ISSN 2052-8604

of authority and responsibility -- key concepts for this study. In examining these relationships, Wettenhall [14] identifies two broad categories: (*a*) horizontal, non-hierarchical arrangements characterised by consensual decision-making and (*b*) hierarchically organised relationships with one party in a controlling role. The implication being, he argues, that true `partnerships' are of type (*a*) and not type (*b*).

This distinction has implications for the public-private partnership in cyber security. National cyber security strategies avoid suggestions of hierarchy when they refer to the public-private partnership. The language is deliberately cooperative and implies a shared purpose and shared interests. The UK Cyber Security Strategy [15] states that achieving the goal of a safe, secure Internet will `require everybody, the private sector, individuals and government to work together. Just as we all benefit from the use of cyberspace, so we all have a responsibility to help protect it'. With specific reference to the role of the private sector, it states that there is an expectation that the private sector will `work in partnerships with each other; Government and law enforcement agencies, sharing information and resources, to transform the response to a common challenge, and actively deter the threats we face in cyberspace' [16]. This non-hierarchical language belies the poor alignment of perceptions about the `common challenge' and the `threats we face in cyberspace' [17]. It assumes that those are the same for the public and private sector when in fact, they are not. The private sector regards cyber security challenges as financial and reputational -- not as a common public good (i.e. whose benefits accrue to the community at large) which is how governments regard national cyber security.

On a more granular level, Linder [18] identifies six distinctive uses of the term P3 and links them to neo-liberal or neo-conservative ideological perspectives. In doing so, he draws out questions about their intended purpose and significance as well as `what the relevant problems are to be solved and how best to solve them'. Two of these `types' can shed light on what is meant by the public-private partnership in cyber security; *partnership as management reform* and *partnership as power sharing*.

Linder argues that partnership as management reform refers to the expectation that government managers will learn `by emulating their partners' and shift their focus from administrative processes to deal making and attracting capital in a more entrepreneurial and flexible approach. Significantly, this is regarded as one of the objectives of the partnership because of the belief that the market is inherently superior and `its competitive character stimulates innovation and creative problem solving' -- a view embedded in neo-liberalism [18]. Perhaps

ISSN 2052-8604

not surprisingly, although this is reflected in the strategies of both states, it is much more pronounced in the US policies.

The [George W.] Bush Administration's National Strategy to Secure Cyberspace [5] argued that in the US “traditions of federalism and limited government require that organizations outside the federal government take the lead” in cyber security. This interpretation of the government's limited authority is combined here with an assumption of its limited capability. “The federal government could not -- and, indeed, should not -- secure the computer networks of privately owned banks, energy companies, transportation firms, and other parts of the private sector”. This is based on the belief that “in general, the private sector is best equipped and structured to respond to an evolving cyber threat” and, at a US Congressional hearing in 2000, Deputy Attorney General Eric Holder's statement that decision makers in the US “believe strongly that the private sector should take the lead in protecting private computer networks.” [19]. In testimony before a hearing on Internet security, the FBI's Michael Vatis argued that cyber security is “clearly the role of the private sector. The Government has neither the responsibility nor the expertise to act as the private sector's system administration.” [20]

So there is a rejection here of government liability for private networks that is framed in the belief that the government has neither the authority nor the capability to deal with cyber security. It is an approach in keeping with the partnership as management reform type identified by Linder -- though the government rejects the objective of change inherent within that type. Rather, it promotes two `truths' about the private sector. First, they must take responsibility and liability for their own network security and second, their superior capacity for flexibility and innovation means that they are best placed take the lead on this particular security problem. The problem of course, is that these networks are central to national security and therein lies the problem from the perspective of the private sector.

The private sector develops security strategy within a very different framework to that of the government's `public good' conception. For the private operators of critical infrastructure, decisions are made within a business model that responds to profit margins and maximising shareholder interests. This is largely incompatible with the promotion of a `public good' (especially in wider context of *Keeping the UK safe in cyber space* [21]) The private sector raises two main objections to the role that the government perceives for them in the cyber security strategies; first, they argue that the expense of ensuring cyber security to a national security level would be significant and second, that the litigious

ISSN 2052-8604

nature of (especially US) society means that industry would be very resistant to accepting liability for the security of their products or systems [22].

Stiglitz and Wallsten [23] make some important observations about this dichotomised approach to public-private partnerships in the context of technology innovation. `Theory predicts' they argue, `and many empirical studies confirm, that profit-maximising firms invest less than the socially optimal level of [technology research and development]'. What is in society's best interest with regard to cyber security, is not always in the best interests of the private sector. This is because, they argue, social benefits do not translate in terms of private profitability -- no matter how desirable the outcome.

So private sector owners of critical infrastructure accept responsibility for securing their systems -- to that point that it is profitable. That is, that the cost of dealing with an outage promises to cost more than prevention [4, 24]. However, they tend to make a distinction between protecting against the low-level threat such as “background noise, individual hackers, and possibly hacktivists” and protecting against an attack on the state (national security). In testimony at a US hearing on privately owned critical infrastructure cyber security, one witness explained that “it is industry's contention that government should protect against the larger threats -- organized crime, terrorists, and nation-state threats -- either through law-enforcement or national defense.” [25]. This was particularly pertinent in the fallout surrounding the Sony Pictures hacking in late 2014[14], even though it remains to be seen whether it was the act of a malicious nation-state.

This disjunction in perceptions is arguably at the heart of the tension in this `partnership'. Typically, the rationale articulated in the literature for partnering is that neither partner on its own can achieve their desired objectives. They must either need each other or there must be a financial arrangement that makes the partnership attractive. This, we can observe most readily in the single most emphasised practice in this partnership -- information sharing. And information sharing can be understood in the second of Linder's `types' of public-private partnerships -- partnerships as power sharing.

Linder writes that partnerships as power sharing are based on an ethos of cooperation where `trust replaces the adversarial relations endemic to command-and-control regulation' and in which there is some mutually beneficial

[14] http://www.bbc.co.uk/news/entertainment-arts-30512032

ISSN 2052-8604

sharing of responsibility, knowledge, or risk. In most instances, he writes, `each party brings something of value to the others to be invested or exchanges'. Finally, `there is an expectation of give-and-take between the partners, negotiating differences that were otherwise litigated.' [18]. The previous section explains how rather than shared responsibility, this partnership is characterised by disputed responsibility. Sharing knowledge, however, is certainly regarded by both partners as integral to this relationship and building trust and collaboration is a dominant theme running through not only the strategy documents but also the responses from the private sector.

The practice of information sharing as a partnership

There can be little doubt that the main form of cooperation within the public-private partnership is found in the shared emphasis on information sharing [9]. In July 2010, the US Government Accountability Office published a report entitled *Critical Infrastructure Protection: Key Private and Public Cyber Expectations Need to Be Consistently Addressed* [26]. The purpose of the study was to clarify the partnership expectations of both the public and private sectors and to determine the extent to which those expectations were being met. The study was limited to five key critical infrastructure sectors deemed to be most reliant on cyber security: communications, defence industrial base, energy, banking and finance, and information technology.

The provision of timely and actionable cyber threat and alert information emerges as a key expectation of the partnership from both the public and the private sectors but there are a number of obstacles to sharing information from both perspectives. The private sector reports that it is not always easy to immediately distinguish between some kind of technical problem, a low-level attack and a large-scale sustainable attack. In addition, it sometimes runs counter to their commercial interests to report vulnerabilities. Finally, for private security firms, sharing information with the government about attacks, could lead to it being shared with their competitors. Their business model is reliant on obtaining, holding and selling information, not sharing it [26].

The public sector also encounters limitations to sharing information. Classified contextual information cannot be shared with individuals who do not have adequate security clearances. Even those working in the private sector who do have security clearance can often do nothing with classified information because to take action on it would expose it. In addition, there is a high expectation that threat information shared from the public to the private sector will be accurate and this leads to extensive and stringent review and revision processes that also delay the release of time critical information [26]. This

ISSN 2052-8604

problem of sharing information has persistently been regarded as a key impediment to cyber security and in testimony before a US Congressional hearing on cyber security in 2011, a senior official highlighted this as one of two main areas that needed improvement [27]. However, the UK Government have attempted to address this issue with the creation of the Centre for the Protection of National Infrastructure[15], which aims to protect national security by providing protective security advice (particularly cyber security/information assurance), as well as CESG[16], the UK's national technical authority for information assurance, which protects the vital interests of the UK by providing advice and guidance to the UK Government on the security of communications and electronic data, in partnership with industry and academia. In March 2013, the UK Government launched the Cyber Security Information Sharing Partnership (CiSP)[17], a joint government and industry initiative to share cyber threat and vulnerability information in order to increase overall situational awareness of the cyber threat and therefore reduce the impact on UK business. These initiatives are overseen by the Office of Cyber Security and Information Assurance[18], which provides strategic direction and coordinates the cyber security programme for the government, enhancing cyber security and information assurance in the UK. Cabinet Office Minister responsible for the Cyber Security Strategy, Francis Maude MP said at the launch of CiSP~\cite{caboff:2013}:

> "*This innovative partnership is breaking new ground through a truly collaborative partnership for sharing information on threats and to protect UK interests in cyberspace. The initiative meets a key aim of our Cyber Security Strategy to make the UK one of the safest places to do business in cyberspace. As part of our investment in a transformative National Cyber Security Programme; we are pleased to provide a trusted platform to facilitate this project.*"

Howard Schmidt, former White House Cyber Security Adviser, welcomed the CiSP announcement, saying:

---

[15] http://www.cpni.gov.uk

[16] http://www.cesg.gov.uk

[17] https://www.cert.gov.uk/cisp

[18] https://www.gov.uk/government/groups/office-of-cyber-security-and-information-assurance

ISSN 2052-8604

> "*In the US, we have seen the emphasis that President Obama has placed on cyber security and in particular steps to protect our critical infrastructure. Many senior leaders in private sector companies are supporting it and recognizing it is not only a security issue but a business imperative. The launch of the UK CISP is an important step in forging an ongoing partnership between industry and government, promoting information sharing by providing the ability to analyze and redistribute information in a timely, actionable and relevant manner.*"

Key objectives and markers of success

By the late 1990s, the critical literature looking at public-private partnerships was maturing and there was a realisation that evaluating these arrangements was complex and under-researched. Essentially, there was little evidence to suggest what the success/failure rate of these arrangements was. In fact, there was not really even a conceptual framework for doing so. In 1999, *American Behavioral Scientist* published a special issue dedicated to these questions. In the introduction, Rosenau summarises [29] many of the journal arguments when she writes that `in general, partnering success is more likely if (*a*) key decisions are made at the very beginning of a project and set out in a concrete plan, (*b*) clear lines of responsibility are indicated, (*c*) achievable goals are set down, (*d*) incentives for partners are established, and (*e*) progress is monitored'. She also identifies a set of criteria for the measurement of success -- some of which are useful in considering this case, particularly accountability and possible conflicts of interest.

In terms of conflict of interest, she makes the case that partnerships do not (as many assume) necessarily reduce regulation. If the interests of the private sector are misaligned with normative goals like care for the vulnerable (for example, old age homes) then the government must monitor and regulate to ensure the profit motive does not supersede the intended delivery of service [29]. Here we see the profile of one of the central problems of this public-private partnership; the expectation that the private sector will invest in cyber security beyond their cost/benefit analysis to fully accommodate the public interest -- in other words, to ensure national security. Because market incentives are not adequate to promote this level of security, oversight and some level of regulation are necessary. A 2013 US Government Accountability Office report [30] found that many of the experts they consulted argued that the private sector had not done enough to protect critical infrastructure against cyber threats. The private sector explanation for not fully engaging in the government's cyber security strategy was that the government had failed to make a convincing business case that mitigating threats warranted substantial

ISSN 2052-8604

new investment. Dunn Cavelty and Suter argue that while public private cooperation is necessary, the way it is organised and conceptualised needs to be rethought. They propose to do so through governance theory and they find that `CIP policy should be based as far as possible on self-regulating and self-organising networks'. By this, they mean that `...the government's role no longer consists of close supervision and immediate control, but of coordinating networks and selecting instruments that can be used to motivate these networks for CIP tasks.' [9]. This may provide some forward momentum though Rosenau makes the point here that a public-private partnership cannot be regarded as a success if it `results in lower quality of public policy services, the need for more government oversight, and the need for expensive monitoring, even if it appears to reduce costs'. Perhaps more problematically for Dunn Cavelty and Suter's recommendation is the problem of accountability.

On accountability, Rosenau writes that because these partnerships often see policy decisions and practices that are normally reserved for elected officials delegated to the private sector, accountability is essential to maintaining a healthy democratic order. If responsibility and accountability can be devolved to private actors, the central principle that political leaders and governments are held to account is undermined [29]. For many scholars, to ensure effective accountability in a public-private partnership, the specifics of roles and responsibilities must be made clear at the outset and goals must be clearly articulated. In addition, Stiglitz and Wallsten [23] observe that in doing so, it becomes clear when additional incentives and resources are necessary to achieve agreed goals and these must be provided if accountability is to be sustained. In cases such as cyber security, in which the public good is the end goal for government, as with the alignment of interests discussed above, accountability does not appear to emerge from market forces alone, nor is it a trivial undertaking [31]. This is not to suggest that public-private partnerships cannot be successful when interests and objectives diverge, but in the view of Stiglitz and Wallsten, in these cases `more attention needs to be placed on the incentive-accountability structure' [23].

The 2010 US GAO report [26] referred to previously is also useful for the analysis of key objectives of this partnership and for measuring its success. The report found that in addition to information sharing, there were two main expectations that the government holds of the private sector in this partnership. First, it was expected that they would commit to execute plans and recommendations such as best practices. This is important because it is an example of the government shifting responsibility to the private sector in the understanding that if the private sector responds, then regulation can be

ISSN 2052-8604

avoided. The study reported that four of the five sectors examined were meeting government expectations to a `great/moderate' degree. The exception was the IT sector which was reported as demonstrating `little/no' commitment to execute plans and recommendations such as best practice. In fact, the IT sector meets only one out of ten services expected by the government to a `great/moderate' degree – technical expertise. On all other criteria, this sector ranked at `some' or `little/no' [26]. Given the reliance of the other sectors on the IT sector, this deficit is particularly concerning and to some degree, has to undermine the others' compliance.

The second key expectation (apart from information sharing) identified in the GAO report is that the private sector will provide appropriate staff and resources. Only banking/finance and commerce were reported to be meeting this expectation to a `great/moderate' degree with defence industrial base, energy and IT all being ranked at `some'. There is clearly a significant skills gap and lack of qualified workers in the UK, although there have been wholesale changes to the school Computing curriculum in England [32] from ages 5-16 (along with reform in Scotland and changes anticipated in Wales) to address broader digital and computational skills. There have also been a range of UK policy announcements specifically addressing cyber security education alongside the curriculum changes, all the way from funding initiatives to improve baseline digital competencies (e.g. *Get Safe Online*[19]), to national learning programmes and competitions (e.g. *Cyber Security Challenge UK*[20]) and the accreditation of appropriate Master's degrees by GCHQ [33]. It remains to be seen how successful these initiatives will be in raising the profile of cyber security skills and careers, as well as developing a sustainable and resilient national capability in this space.

Conclusions

At this stage, prior to the fieldwork interviews, it is possible to draw some preliminary conclusions. First, and somewhat surprisingly given its centrality in successive cyber security policies, exactly what this `partnership' entails has always been unclear. Unpacking it has revealed that there are inherent tensions and misaligned objectives that are not in keeping with expectations of public-private partnership arrangements. The partnership is consistently referred to in strategy documents using normative, value based language rather

[19] http://www.getsafeonline.org

[20] http://cybersecuritychallenge.org.uk/education

ISSN 2052-8604

than clear statements outlining legal authority, responsibility and rights. Although politicians subscribe to the notion that there exists (or should exist) a deeply entrenched norm of cooperation between the government and private sector this appears not to be the case. Rather, the private sector has consistently expressed an aversion to accepting responsibility for national security and regard cyber security within a cost/benefit framework rather than a `public good' framework. This is particularly pertinent given the high-profile announcements in January 2015 of increased US-UK cyber security cooperation [34], especially in strengthening cooperation on cyber defence, supporting academic research on cyber security and improving critical infrastructure cyber security.

The second conclusion to arise from this study is that we are witnessing a unique approach to `out-sourcing' national security that has implications for conceptions of governance, state power, global security and international partnerships and resource sharing. States with greater government control over critical infrastructure and also over their information infrastructure potentially have a significant advantage in that they are able to control and shape their response to cyber insecurity with greater autonomy and agency. However, there are potentially profound consequences for civil liberties: monitoring, data retention, the use of encryption and more broadly what we mean by `digital rights'. This is of particular relevance to emerging UK cyber security strategy and needs to be considered more thoroughly from a research, policymaking and national infrastructure perspective.

References

[1] Bill Clinton. Speech on the economy at Wharton School of Business, University of Pennsylvania. http://www.ibiblio.org/nii/econ-posit.html, April 1992.

[2] Barack Obama. Remarks by the President On Securing Our Nation’s Cyber Infrastructure. http://www.whitehouse.gov/the press office/Remarks-by-the-President-on- Securing-Our-Nations-Cyber-Infrastructure/, May 2009.

[3] House of Commons. Scientific advice and evidence in emergencies. Report HC 498, Science and Technology Select Committee, February 2011.

[4] Department for Business, Innovation & Skills. 2014 Information Security Breaches Survey. UK Government, 2014.

[5] George W. Bush. The National Strategy to Secure Cyberspace. The White House, February 2003.

[6] Francis Maude. Cyber Security Strategy one year on, speech on the 2012 Information Assurance Conference.

ISSN 2052-8604

https://www.gov.uk/government/speeches/francis-maude-speech-at- ia12-cyber-security-strategy-one-year-on, December 2012.
[7] Max G. Manwaring. The Inescapable Global Security Arena. Strategic Studies Institute, US Army War College, 2002.
[8] US Department of Commerce. White House Announces Public-Private Partnership Initiatives to Combat Botnets. http://www.commerce.gov/news/press-releases/2012/05/30/white-house-announces-public-private-partnership-initiatives-combat-b, May 2012.
[9] Myriam Dunn Cavelty and Manuel Suter. Public-Private Partnerships are no silver bullet: An expanded governance model for Critical Infrastructure Protection. International Journal of Critical Infrastructure Protection, 2(4):179–187, 2009.
[10] Nazli Choucri, Stuart Madnick, and Jeremy Ferwerda. Institutions for Cyber Security: International Responses and Global Imperatives. Information Technology for Development, 20(2):96–121, 2014.
[11] Stephen P. Osborne, editor. Public-Private Partnerships: Theory and Practice in International Perspective. Routledge, 2007.
[12] Piet de Vries and Etienne B. Yehoue, editors. The Routledge Companion to Public-Private Partnerships. Routledge, 2013.
[13] Mariana Mazzucato. The Entrepreneurial State: Debunking Public vs. Private Sector Myths. Anthem Press, 2013.
[14] Roger Wettenhall. The Rhetoric and Reality of Public-Private Partnerships. Public Organization Review, 3(1):77–107, 2003.
[15] Cabinet Office. Cyber Security Strategy. UK Government, November 2011.
[16] Cabinet Office. National Cyber Security Strategy 2013: forward plans and achievements. UK Government, December 2013.
[17] Amyas Morse. The UK Cyber Security Strategy: Landscape Review. UK National Audit Office, February 2013.
[18] Stephen H. Linder. Coming to Terms With the Public-Private Partnership: A Grammar of Multiple Meanings. American Behavioral Scientist, 43(1):35–51, 1999.
[19] Eric H. Holder, Jr. Statement before the Subcommittee on Communications, Senate Committee on Commerce, Science and Transportation. http://www.justice.gov/archive/dag/testimony/ holderinternet38.htm, March 2000. US Deputy Attorney General.
[20] Michael A. Vatis. Statement before the Senate Armed Services Committee. http://fas.org/ irp/congress/2000 hr/000301mv.pdf, March 2000. Deputy Assistant Director, Federal Bureau of Investigation.
[21] Cabinet Office. Keeping the UK safe in cyber space. UK Government, December 2014.

ISSN 2052-8604

[22] Alan Paller. SCADA Systems and the Terrorist Threat: Protecting the Nations Critical Control Systems. Statement before the Subcommittee on Economic Security, Infrastructure Protection and Cybersecurity, House Committee on Homeland Security. http://www.gpo.gov/fdsys/pkg/CHRG-109hhrg32242/html/CHRG-109hhrg32242.htm, October 2005. Director of Research, The SANS Institute.
[23] Joseph E. Stiglitz and Scott J. Wallsten. Public-Private Technology Partnerships: Promises and Pitfalls. American Behavioral Scientist, 43(1):52–73, 1999.
[24] Department for Business, Innovation & Skills. Cyber governance health check: 2014. UK Government, 2014.
[25] Sam Varnado. SCADA Systems and the Terrorist Threat: Protecting the Nations Critical Control Systems. Statement before the Subcommittee on Emergency Preparedness, Science and Technology, House Committee on Homeland Security. http://www.gpo.gov/fdsys/pkg/CHRG-109hhrg32242/html/CHRG-109hhrg32242.htm, October 2005. Director of Information Operations Center, Sandia National Laboratory.
[26] David A. Powner. Critical Infrastructure Protection: Key Private and Public Cyber Expectations Need to Be Consistently Addressed. US Government Accountability Office, July 2010.
[27] Gregory C. Wilshusen. Cybersecurity: Continued Attention Needed to Protect Our Nation's Critical Infrastructure and Federal Information Systems. Statement before the Subcommittee on Cybersecurity, Infrastructure Protection and Security Technologies, House Committee on Homeland Security. http://www.gao.gov/assets/130/125787.html, March 2011. Director Information Security Issues, US Government Accountability Office.
[28] Cabinet Office. Government launches information sharing partnership on cyber security. https://www.gov.uk/government/news/government-launches-information-sharing- partnership-on-cyber-security, March 2013.
[29] Pauline Vaillancourt Rosenau. The Strengths and Weaknesses of Public-Private Policy Partner- ships. American Behavioral Scientist, 43(1):10–34, 1999.
[30] Gregory C. Wilshusen and Nabajyoti Barkakati. Cyber Security: National Strategy, Roles, and Responsibilities Need to Be Better Defined and More Effectively Implemented. US Government Accountability Office, February 2013.
[31] Colin Williams. Security in the cyber supply chain: Is it achievable in a complex, interconnected world? Technovation, 34(7):382–384, 2014. Special Issue on Security in the Cyber Supply Chain.

ISSN 2052-8604

[32] Department for Education. National curriculum in England: computing programmes of study. https://www.gov.uk/government/publications/national-curriculum-in-england- computing-programmes-of-study, September 2013.
[33] GCHQ. Developing the Cyber Experts of the future – GCHQ certifies Master's Degrees in Cyber Security. http://www.gchq.gov.uk/press and media/press releases/pages/gchq-certifies- masters-degrees-in-cyber-security.aspx, August 2014.
[34] The White House. FACT SHEET: U.S.-United Kingdom Cybersecurity Cooperation. http://www.whitehouse.gov/the-press-office/2015/01/16/fact-sheet-us-united- kingdom-cybersecurity-cooperation, January 2015.

ISSN 2052-8604

Business versus technology: Sources of the perceived lack of cyber security in SMES

Emma Osborn[1], Sadie Creese[2], and David Upton[3]
[1]University of Oxford Centre for Doctoral Training in Cyber Security
emma.osborn@cybersecurity.ox.ac.uk

[2]University of Oxford Cyber Security Centre
sadie.creese@cybersecurity.ox.ac.uk

[3]University of Oxford Said Business School
david.upton@sbs.ox.ac.uk

Abstract

There is increasing concern about the standard of cyber security in SMEs, voiced by governments and the large companies who interface with them, yet many past initiatives seem to have failed to have a significant impact on the sector. In this paper, we report upon a study in which Small and Medium Enterprises (SMEs) were surveyed to establish what barriers they might face in terms of cyber security. The results were combined with publicly available information to identify how stakeholders in the SME cyber security ecosystem interact, and establish whether the perceived lack of uptake of cyber security measures in SMEs was accurate. The paper concludes by discussing how the refined understanding of the barriers faced by SMEs might influence development of future SME security solutions.

Introduction

The European Commission states that 99% of businesses in Europe are Small and Medium Enterprises (SMEs), using the definition provided in EU law (EU recommendation 2003/361 [1, 2]), as outlined in Table 1. This definition is used in the UK where SMEs accounted for 59.3% of private sector employment and 48.1% of turnover in 2013 [3]. Despite this, products, standards and the education of cyber security professionals most often focus on big budget options for securing large organisations.

ISSN 2052-8604

Company Category	Employees	Turnover or Balance Sheet
Medium-sized	< 250	≤ €50m or ≤ €43m
Small	< 50	≤ €10m or ≤ €10m
Micro	< 10	≤ €2m or ≤ €2m

TABLE 14 EU SME DEFINITION

The overall sustainability of a cyber security model that is largely inaccessible to smaller companies is questionable. While the threat and implications of a cyber security incident are considered to be higher in large organisations, a lack of downwardly scalable and affordable options for SMEs may put these businesses and the wider supply chain at risk.

While there are risks posed by the way SMEs are perceived as approaching cyber security, the sector also represents a huge marketplace for any supplier able to produce suitable user-friendly and low cost solutions. Currently only antivirus products have become widely adopted in the sector, which raises the question why comparatively few companies have managed to make the business case for developing with this sector in mind; or why SMEs appear to be so much slower in adopting the products available to them than larger companies and government.

This study began by approaching SMEs, using a questionnaire to get their perspective on cyber security issues, details of which can be found in Section 2. Information from the questionnaire has been combined with publicly available information about how other government and private sector stakeholders influence the SME cyber security marketplace, in order to describe the often disjointed dialogue between the stakeholders in this ecosystem, and its outcomes.

The point of view of each stakeholder group is described in Sections 3 – 5, and Section 6 concludes the paper. To aid the reader some high level statistics about the questionnaire dataset are outlined in the following subsection.

ISSN 2052-8604

Questionnaire Statistics

There were 33 respondents to the survey, from 19 different industry sectors. The sector with the highest number of respondents was IT and telecoms (8), and there were 11 respondents who provided professional services other than IT. Respondents were distributed across 15 UK counties, with one response from a company outside of the UK. There were 8 respondents in single person companies, 13 in micro companies of more than one person, 10 in small companies, and 2 in medium-sized companies.

Methodology

The primary source of data for this study was a questionnaire, aimed at SME owners, directors or managers and consisting of 19 questions or questions sets. Questions were reviewed by both cyber security experts and SME owners before the survey was carried out. To introduce an extra level of granularity, an extra category of SMEs, in addition to those defined in Table 1 was used — companies which had only one person. Questionnaire data was combined with observations while interacting with SMEs, additional information supplied by some respondents and existing literature.

The scope of the results is limited by the number of participants, as only 33 responses are included in the dataset. In order to obtain a representative cross-section of the SME community a variety of means were used to identify potential respondents — the researcher's professional contact list, Oxford University social media accounts and by using council local business directories to locate websites or email addresses.

This selection technique introduced selection biases due to respondents self-selecting based on their interest in cyber security, the level of knowledge some of the data collection methods tested and the uneven distribution of respondents across different sizes of company (although the distribution of respondents is representative of the respective numbers of each size of company in the UK).

Multiple linear regression was used against the numerically-coded portion of the dataset [4], and statistically significant relationships, outliers and the open response questions were combined with observational and literary data, using Grounded Theory [5] to develop the concepts presented in this paper. A number of SMEs were consulted about the interpretation of responses.

Concerns Over Cyber Security in SMEs

ISSN 2052-8604

This subsection focuses on the stakeholders within the SME cyber security ecosystem who are voicing concern over the level of cyber security achieved by SMEs. Cyber security in SMEs is an issue which seems to periodically cycle in and out of the spotlight. There are three main arguments given for concern over SME cyber security:

1. The suggestion that SMEs are being used as attack vectors for large companies or government departments higher up the supply chain [6].
2. Attacks reported by SMEs in surveys such as the UK Department for Business Innovation and Skills (BIS) Information Security Breaches Survey [7].
3. The political moral incentive — the risk of financial loss from cyber incidents versus the number of SMEs in the UK, the percentage of the GDP they account for and the amount of employment they provide [8].

The concerns stated above have been the drivers for various SME security initiatives.

Past initiatives include *ENISA's Risk Assessment and Risk Management Methods: Information Packages for SMEs* (March 2006) [9] and ISSA-UK's *Security Standard for SMEs* (March 2011) [10]. The ENISA information package is still available, and while some of the advice it contains now seems fairly dated in the face of emerging attacks, it describes a risk analysis process in accessible language.

Present UK Government initiatives are slightly more varied in their approaches and development process: the Cyber Security Voucher Scheme, Cyber Streetwise, and Cyber Essentials [11, 12, 13], were all launched in the last year. With this level of investment one may ask where are the barriers?

Concern 1

Full details of attacks are rarely published, meaning that Concern 1 could be viewed as hearsay, but even with concrete and SME-accessible evidence that this is the case, the risk owners are not the SMEs. The SMEs surveyed in this study were asked whether there was a risk of them losing business due to customers asking them to adhere to cyber security standards. Respondents were almost unanimous in denying any pressure having been put on them by the supply chain— risk is not being systematically transferred.

ISSN 2052-8604

Concern 2

This could directly impact a company's profitability but in the BIS survey [7] the respondents were mainly from much larger companies, with only one in three not holding an IT or security role. Smaller SMEs looking for evidence of cyber risk may not identify with this survey's respondents.

Concern 3

Concern 3 is a political argument, with the risk holder being UK PLC. Moral incentives of this type are documented as being poor motivators due to the underlying message that nobody is achieving what they are expected to [14], and SME owners may not welcome government interference.

While the concerns about SME cyber security are based on risks and requirements owned by large organisations the SMEs will struggle to find strong cyber security incentives. The survey responses about the source of concern were mixed, with 60% agreeing that someone they knew had made them think cyber security was important and only 36% saying that that the press had made them worried about cyber security.

The Impact of the ICO on Concern About SME Security

The Information Commissioner's Office (ICO) is different, because their clear negative incentive is not aimed at protecting SMEs; rather, the aim is to protect the general public from any organisations being negligent with the personal data they hold. The very high proportion of respondents aware of their data risk may be linked to the success in publicising the ICO's power to apply clear financial and reputational penalties. 82% of survey respondents agreed with the statement "We have customer or supplier data that we need to protect."

Lack of Cyber Security Industry Focus on SMEs

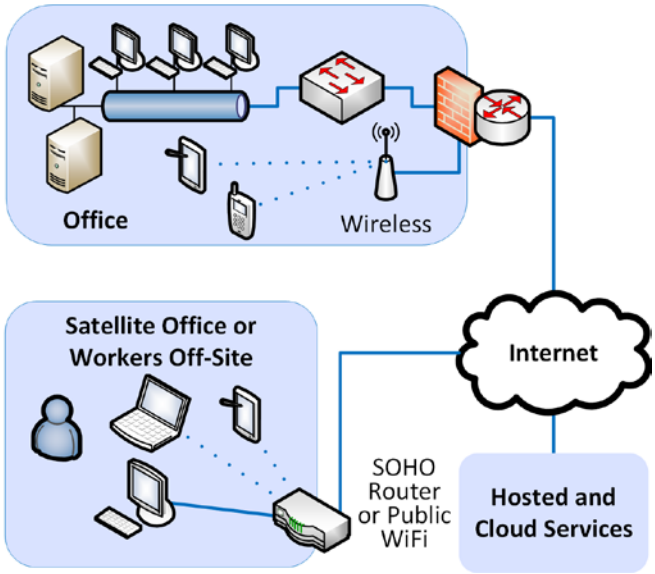
The cyber security industry has developed over time in response to risks posed by cyber threats and cyber security measures are derived from the identification of vulnerabilities. Directly or indirectly government and large companies tend to finance the development of cyber defences – if the funding, or security requirement comes from these organisations that is where vulnerability researchers will search. This raises the question of scalability – to what extent do these vulnerabilities and solutions apply to smaller organisations?

Personal devices are almost identical in a home or corporate setting. In these cases a healthy market of cheap, publicly available cyber security measures is

ISSN 2052-8604

available. The big issue is that, beyond the endpoints, SME IT infrastructures can vary enormously. Network design engineers typically create high level designs describing network infrastructure in large organisations [15]. They categorise the different types of site the customer has, aligning requirements to each site, with cyber security devices included in the design.

FIGURE 24 AN EXAMPLE OF THE INFRASTRUCTURE DESCRIBED FOR AN SME FOR CYBER SECURITY PURPOSES



A brief survey of network designs published by cyber security stakeholders, to describe either the way they have approached SME security initiatives or products aimed at small companies, typically produces a diagram similar to that of Figure 1.

The derivation of the SME network diagram from a typical network designer's 'small site' is obvious, but there are multiple indicators in this study's survey results that this is not an accurate representation of many SME infrastructures, it's the one the cyber security industry understands, and can supply some security measures for.

There are multiple questions within the dataset which aid in building a picture of SME IT infrastructures. Part of the issue is that it is impossible to define what a single 'normal' SME infrastructure model could look like, so the responses are considered given the size of the company. In all cases it is assumed that the respondents will use SOHO routers for offices of up to 10 people, for the infrastructure shown in Figure 3.B this has been ratified by the respondent.

Single Person Companies

Most single person companies do not have dedicated offices. The significance of this is that where there are no dedicated offices there is no in-house dedicated network infrastructure, meaning Figure 2 shows an infrastructure more familiar in a home environment.

None of the respondents are in the security or IT sector, and only one of

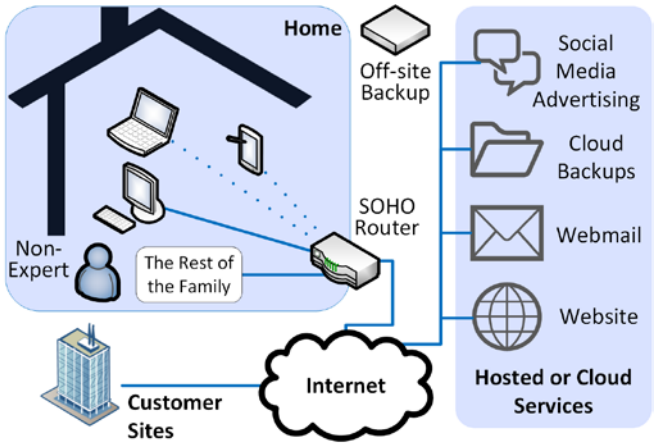


FIGURE 25 INFRASTRUCTURE IN SINGLE PERSON COMPANIES

ISSN 2052-8604

these companies gets an IT expert to set up their computer. All respondents have at least one computer containing company data; half have two. Multiple computers and a lack of company-issued machines signify company data is often held on personal machines.

All but one of these respondents have a company website, a smartphone or tablet containing company data, allow automatic updates and keep backups off-site. Half of the respondents say they use webmail, and a different set of four respondents said they use cloud services to store data or use applications. Half have suppliers or customers who provide a link into their IT systems, or who they allow to link into theirs. Five of these companies have their own social media accounts and use social media as their only or main source of advertising.

Micro Companies

Nine out of the 13 micro companies who responded to this survey had dedicated offices; this leads to possible infrastructures ranging from all employees working from home offices linked by cloud services, to a single small office.

Small and Medium Companies

All of the respondents working in companies with between 10 and 249 people have dedicated offices. Figure 4 focuses on the two respondents from medium-sized companies, to illustrate possible evolutions in infrastructure as an SME grows. There is little resemblance to the SME infrastructure described in Figure 1 in either example.

The first is in the Education sector and is felt to be the respondent with the highest level of fixed infrastructure because educational establishments have a requirement not only to provide IT services for their staff but also for their students. As a guide information has been taken from a school who list their ICT facilities on their website. This is a school which has 162 staff, and 1420 pupils:

> *"There are 8 ICT suites spread across the school. These suites have 32 Windows 7 PCs and a Promethean interactive whiteboard. . .All teachers have access to a PC in their individual teaching rooms."[16]*

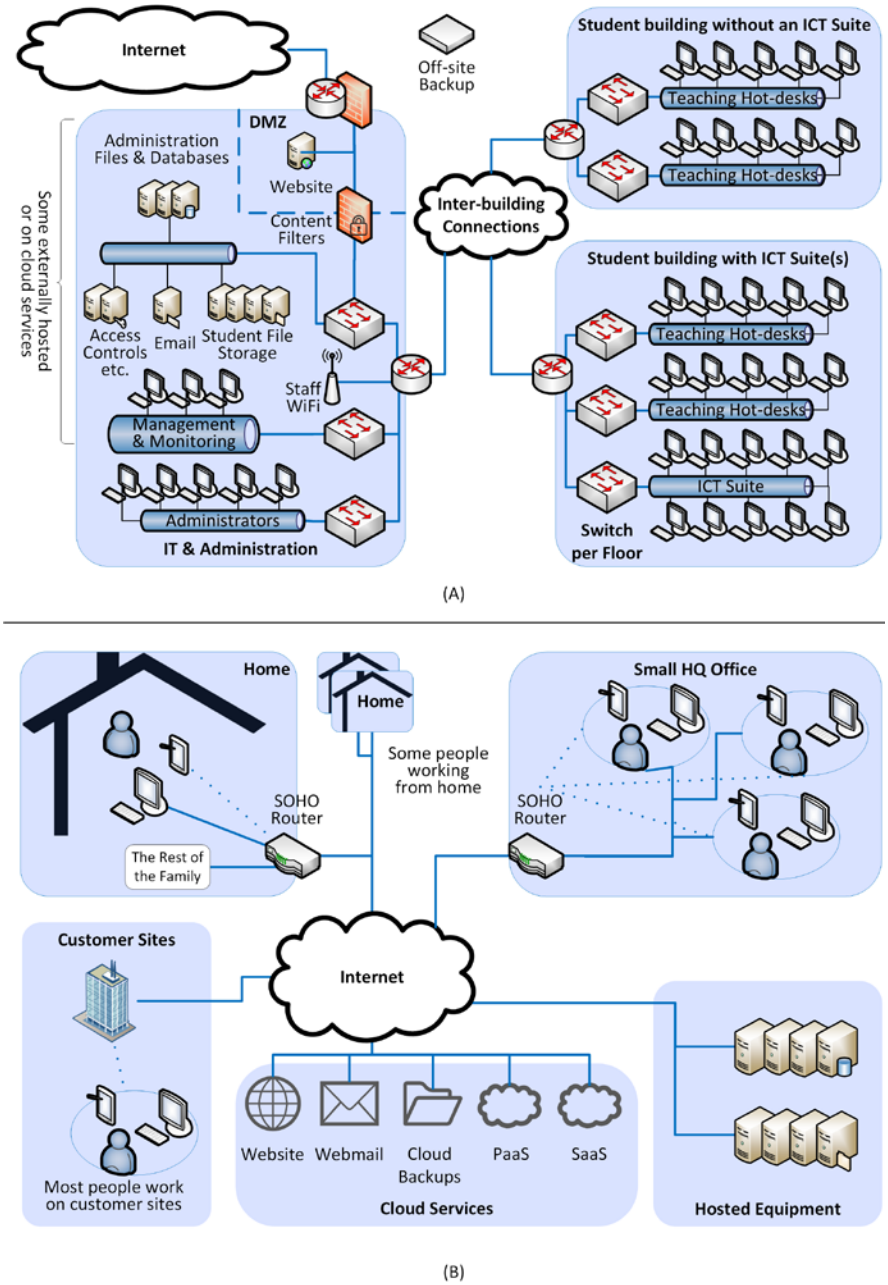
The second respondent from a medium-sized enterprise chose to describe their infrastructure in his own words:

ISSN 2052-8604

“From the start we have taken an approach of not having much, if any, on-premises IT systems. As a consequence all of our business systems are either provided as software as a service (for example the HR system) or as platform as a service (using AWS). We also use co-location facilities to host some equipment. In addition most of our employees work at customer provided facilities and we have a single small HQ office with a number of us working out of our homes. We use Office 365 as our email solution.”

ISSN 2052-8604

FIGURE 26 INFRASTRUCTURE IN MEDIUM-SIZED COMPANIES: (A) RESPONDENT FROM THE EDUCATION SECTOR; (B) RESPONDENT FROM THE GOVERNMENT & DEFENCE SECTOR



How SMEs Experience Cyber Security

Some of the reticence from SMEs to engage with cyber security cannot be explained by lack of awareness or general inaction. The following subsections discuss other influencing factors.

ISSN 2052-8604

Cyber Budgets

In the survey, respondents were asked how much they currently spend on cyber security. Their responses can be seen in Figure 5. What is apparent is that the budget range a company is willing to spend on cyber security increases with the size of the company, as would be expected. The number of companies with a total cyber spend of less than £100 per year shows the level of constraint some companies are under, but also shows why the cyber security industry might find this sector less appealing when compared to the multi-million pound secure systems extremely large organisations deploy. The sales and customer services overhead involved in dealing with thousands of customers does not make the SME sector 'low-hanging fruit'.

The per-person representation in Figure 5, where the budget ranges stated have been divided by the number of people in that sized company, shows that the cost per person rises dramatically as company size increases, with no economy of scale until a company reaches a medium size.

The graph also provides further insight into the barriers facing SMEs. In the small company group two clear bands can be seen. The first at £10-50 is matched in the micro companies, and is roughly the price of basic endpoint security. The second band, at £110-500, can therefore be assumed to show the

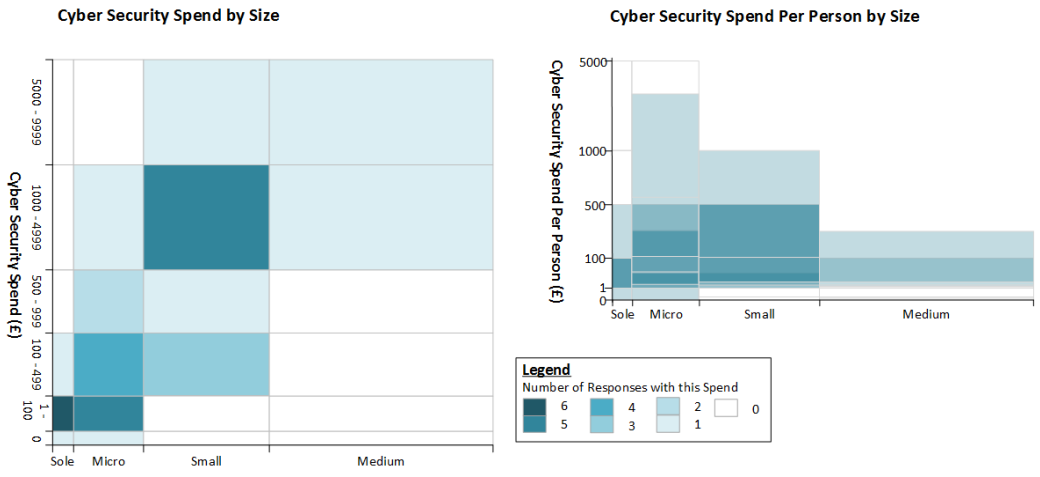


FIGURE 27. CYBER SECURITY BUDGETS

cost of transitioning from home computer security, to what the cyber industry would recognise as corporate security.

A lack of resources is clearly an issue for SMEs. Six respondents stated that a lack of time or financial resource is what they found most difficult about cyber

ISSN 2052-8604

security, but in the context of maintaining or enforcing security policies having low resource could allow the decision maker to retain an overview of the company's security.

Risks & Requirements

Risk has been mentioned on several occasions throughout this paper, and is seen as a central consideration in any decision about cyber security. The responses to questions about the formal risk analysis process showed 39% have done an in-depth risk analysis which included cyber security and 48% keep the company's risk analysis, policies and backups up to date. These figures are low when compared to the BIS information breaches survey 2013 [7], which shows 60% doing a risk analysis including information security.

Irrespective of whether a respondent had carried out a formal risk assessment they demonstrate that they are aware of reasons for implementing security measures. The barrier they face is one of a lack of knowledge. While the survey results suggest that SMEs are struggling to quantify their cyber security risk it does give some examples of treatment strategies they might be employing. Taylor defines the four risk treatment options as avoid, reduce, transfer or accept [17].

There is some evidence in the survey of risk avoidance. One participant mentioned modifying their behaviour by *"not visiting dodgy websites"* which although free comes at a personal cost —modifying their web use in both their professional and private life. The evidence of respondents employing basic cyber security measures, such as antivirus, probably indicates that companies are deploying basic measures for risk reduction or due diligence. There was no explanatory relationship between data protection risk and the cyber security requirements respondents indicated later in the survey. This may well mean that, having carried out their basic risk reduction measures, the quantifiable risk provided by clearly outlined financial penalties is allowing SME owners to accept the risk. The option often promoted by the cyber security industry as a catch-all solution for SMEs is the use of cloud services, which should allow SMEs to transfer cyber security risk. There are respondents using cloud services, but there is little evidence of respondents managing to successfully transfer risk (a more in-depth analysis of SME cloud-security requirements has been carried out by Sangani et al [18]).

The three concepts that regression analysis identified as significant in explaining requirements were: Basic knowledge, Risk (predominantly of linking

ISSN 2052-8604

to another company's IT system, but also of losing IP) and the use of cloud services, including webmail.

The first thing to note is that these influencers provide evidence that respondents are highlighting requirements based on their business risk. This indicates that SMEs are evaluating their companies for risk, even where the respondent states that a formal risk analysis is not part of their business processes. Irrespective of their industry, the respondents are applying what knowledge they have of cyber security to independently resolve their issues.

The preponderance of requirements related to the interconnection of systems over other types of risk is felt to be related to a risk to reputation. Companies have been shown to survive large data breaches [19, 20], but allowing a company to link IT systems requires a higher level of trust.

The Art of the Possible

Section 4 described SME infrastructures in the context of the suitability of existing cyber security offerings. What is also apparent from the infrastructure diagrams, is the lack of elements within these systems which are under the control of the risk owner. The most extreme case is the single person company, where the security of everything beyond the end point is in the hands of third parties, often holding no contractual requirement to provide good cyber security.

SMEs also have to develop their cyber footprint in order to advertise their companies. The result is that out of necessity to trade these individuals make themselves extremely easy to target. This intersection of a large cyber footprint and a small IT infrastructure means that many of their highest impact cyber risks are in systems managed by organisations they interact with, for example: *"People are hacking accountant's login details to HMRC to submit false tax repayment returns."* This argument is emulated in Section 3 by stakeholders impacted by an SME's lack of cyber security, bringing the study full circle.

Conclusions & Future Work

Existing business models for cyber security are unsustainable — the level of interactions and interconnections between traditional cyber security stakeholders and those requiring small-scale cyber security solutions are too complex.

ISSN 2052-8604

What has been shown in this study is that the respondent group all employ some kind of cyber security measures, attempt to make judgements about requirements based on the risks that they understand, and face significant barriers. There is a lack of focus from the cyber security industry on the types of measures SMEs think they need within an SME budget — “simple effective measures that are not too time consuming and require a great in depth knowledge of IT systems.” Due to the size of the dataset the results of this study can only be seen as preliminaries to a wider study, there are several potential strands of development:

- Respondents show a need for unbiased advice and threat intelligence that would enable SMEs to do DIY security.
- The home cyber security market is user-centric, configuration is simplified and cyber products are offered as part of larger purchases, or as software for evaluation. Where SMEs need the cyber security sector to bridge the gap between home and corporate security this business model may be a more effective than the traditional product-centric corporate cyber security model.
- A reduction in the barriers for SMEs wanting to use cloud services.

References

[1] European Commission. What is an SME? http://ec.europa.eu/enterprise/policies/sme/factsfigures-analysis/sme-definition/index en.htm.
[2] European Commission. EU recommendation 2003/361 concerning the definition of micro, small and medium-sized enterprises, May 2003.
[3] Department for Business Innovation and Skills. Business population estimates for the UK and regions 2013, October 2013.
[4] Mark Saunders. Research methods for business students. Pearson Education, Harlow, 6th edition, 2012.
[5] Christina Goulding. Grounded theory: a practical guide for management, business and market researchers. Sage, London, 2002.
[6] ADS. Defence cyber protection partnership (DCPP), April 2014. www.adsgroup.org.uk
[7] Information security breaches survey 2013: technical report - publications - GOV.UK. https://www.gov.uk/government/publications/informationsecurity-breaches-survey-2013-technical-report.
[8] Warwick Ashford. Small firms lose up to 800m to cyber crime, says FSB. ComputerWeekly.com, May 2013.
[9] Information packages for small and medium sized enterprises (SMEs) - ENISA, 2006.

ISSN 2052-8604

https://www.enisa.europa.eu/activities/riskmanagement/files/deliverables/information-packagesfor-small-and-medium-sized-enterprises-smes.
[10] Tim Holman. ISSA 5173 the security standard for SMEs, March 2011. http://www.2-sec.com/2011/03/22/issa-5173-the-security-standardfor-smes/.
[11] Technology Strategy Board. Innovation vouchers for cyber security, July 2014. https://vouchers.innovateuk.org/cyber-security.
[12] HM Government. Cyber Street, 2014. www.cyberstreetwise.com.
[13] Cyber essentials scheme: overview - publications - GOV.UK. https://www.gov.uk/government/publications/cyberessentials-scheme-overview.
[14] Robert B. Cialdini. Crafting normative messages to protect the environment. Current Directions in Psychological Science, 12(4):105–109, August 2003.
[15] Priscilla Oppenheimer. Top-down network design. Cisco Press, Indianapolis, Ind, 3rd edition, 2011.
[16] The Mountbatten School website. http://www.mountbatten.hants.sch.uk/home/.
[17] Andy Taylor. Information security management principles. BCS, Swindon, second edition. edition, 2013.
[18] N.K. Sangani, T. Vithani, P. Velmurugan, and M. Madiajagan. Security & privacy architecture as a service for small and medium enterprises. In 2012 International Conference on Cloud Computing Technologies, Applications and Management (ICCCTAM), pages 16–21, December 2012.
[19] Target data theft hit 70 million. BBC, January 2014. http://www.bbc.co.uk/news/technology-25681013.
[20] Sony fined over PlayStation hack. BBC, January 2013. http://www.bbc.co.uk/news/technology-21160818.