Wetenschappelijk Onderzoek- en
Documentatiecentrum
Ministerie van Veiligheid en Justitie

Memorandum 2017-4

Cybermetrics 2016: DDoS en malware

R. Cornelisse
M.S. Bargh
R. Choenni

Memorandum
De reeks Memorandum omvat de rapporten van onderzoek dat door en in opdracht van het WODC is verricht.
Opname in de reeks betekent niet dat de inhoud van de rapporten het standpunt van de Minister van Veiligheid en Justitie weergeeft.

Alle rapporten van het WODC zijn gratis te downloaden van www.wodc.nl

Inhoud

Bijlage

1 Achtergrond Cybermetrics

Vanuit de overheid, het bedrijfsleven en de burger is er behoefte aan inzicht in de stand van zaken op het gebied van cyber security in Nederland. Het jaarlijkse Cyber Security Beeld Nederland (CSBN), dat door het NCSC wordt samengesteld, geeft inzicht in recente ontwikkelingen, belangen, dreigingen en weerbaarheid op het gebied van de nationale cyber security. Het merendeel daarvan betreft echter kwalitatieve beschrijvingen, aangezien kwantitatieve gegevens slechts beperkt beschikbaar zijn.
Tussen juli 2014 en april 2015 heeft het WODC, op verzoek van het NCSC en met inbreng van NCSC en Capgemini, gewerkt aan de ontwikkeling van een Cyber Security Dashboard (CSD). In deze periode is de informatiebehoefte van verschillende (potentiële) doelgroepen van een CSD in kaart gebracht, is een informatiemodel vastgesteld, het ontwikkelproces uitgewerkt en ingericht, zijn gegevensbronnen geïnventariseerd, is een begin gemaakt met de daadwerkelijke verzameling en bewerking van gegevens en zijn verschillende mogelijkheden voor analyse en weergave via indicatoren in een dashboard onderzocht. Tevens is een eerste versie van een wenselijke beheerstructuur van een CSD uitgewerkt, en is een Privacy Impact Assessment (PIA) in gang gezet. Een uitgebreidere beschrijving van de opbrengsten van het onderzoeksproject is te vinden in *Cyber Security Dashboard – eindrapport fase 1* (WODC/NCSC/Capgemini, 2015).
Cybermetrics 2016: DDoS en malware is de nieuwe fase van het oorspronkelijke CSD dat als doel heeft het ontwikkelde informatiemodel toe te passen op een tweetal specifieke fenomenen op het gebied van cyber security: DDoS-aanvallen en malware.

2 Doelen

Centraal in het project *Cybermetrics 2016: DDoS en malware* stonden een vijftal doelstellingen:

1. Wat zijn relevante databronnen voor kwantitatieve informatie over DDoS en malware?
2. Hoe kan het CSD-informatiemodel worden toegepast om met behulp van deze bronnen snel een zo volledig en consistent mogelijk beeld te geven van deze fenomenen?
3. Wat zijn de meest relevante indicatoren om een kwantitatief beeld te schetsen van de stand van zaken m.b.t. DDoS-aanvallen en malware?
4. Wat zijn de onderlinge relaties tussen de indicatoren?
5. Hoe kunnen deze indicatoren en relaties gevisualiseerd worden voor gebruik in o.a. het Cyber Security Beeld Nederland?

Naast het beantwoorden van de bovenstaande onderzoeksvragen zouden aan het einde van het project de volgende deliverables worden opgeleverd:

1. Een overzicht van bronnen voor beide topics;
2. Een overzicht van indicatoren en relaties daartussen voor beide topics;
3. Een voorstel voor statische visualisaties voor beide topics, bijv. een infographic;

4 Een voorstel voor dynamische visualisaties voor beide topics, bijv. topic-specifiek dashboard;
5 Een overzicht van bereikte inzichten in de vorm van een onderzoeksrapport.
6 Ten minste één wetenschappelijk artikel.

3 Aanpak

Voor het behalen van de doelstellingen was het noodzakelijk om kwantitatieve gegevens te verzamelen over malware en DDoS aanvallen. Echter, zowel tijdens het haalbaarheidsonderzoek als tijdens de eerste fase van het CSD was al gesignaleerd dat de terughoudendheid van potentiële partijen om daadwerkelijk relevante data te leveren één van de grootste uitdagingen zou worden. Daarom is er voor het verzamelen van de gegevens tijdens Cybermetrics 2016 voor gekozen om organisaties te benaderen die in het verleden al bereid zijn geweest om gegevens te verstrekken en organisaties waarvan de verwachting was dat deze eenvoudig te overtuigen waren om gegevens te verstrekken. Er zijn in totaal dertien organisaties benaderd voor het leveren van data.

Omdat de geformuleerde doelstellingen te abstract waren om direct tot nut te zijn bij het benaderen van de verschillende organisaties, zijn eerst de onderstaande onderzoeksvragen geformuleerd om de doelstellingen van het project concreter te maken.

DDoS

- Wat is de frequentie van DDoS-aanvallen?
- Wat is de duur/totaal volume/piek intensiteit van DDoS-aanvallen?
- Welke soort van aanvallen worden gezien? Wat is bijvoorbeeld de verhouding tussen UDP en TCP of tussen NTP-reflectie en DNS-reflectie? Wat is bijvoorbeeld de verhouding tussen OSI laag-3/4 aanvallen en laag-7?
- Welke patronen worden gezien in het aanvalsverkeer? Is een aanval bijvoorbeeld constant, oplopend, periodiek of volledig onvoorstelbaar? Welke TCP-flags worden bijvoorbeeld gebruikt, beginnen aanvallen eerst klein, nemen ze langzaam af, enzovoort?
- Wat is de herkomst van DDoS-verkeer? Wat is bijvoorbeeld de verhouding tussen botnets, booters en open DNS-resolvers? Wat is bijvoorbeeld de verhouding tussen AS'en of landen?
- Welke apparaten worden gebruikt om DDoS-aanvallen uit te voeren? Wat is bijvoorbeeld de verhouding tussen pc's, servers en diverse Internet-of-Things-devices?

Malware

- Wat is de frequentie van malware aanvallen (als bijlage of via malafide URL)?
- Welke soort van malware wordt gesignaleerd? Wat is bijvoorbeeld de verhouding tussen malafide PDF of Word-document met malafide macro's?

4 Bottlenecks

Bij het benaderen van elke organisatie is gekozen voor een bottom-up benadering door personen te contacteren die direct betrokken zijn bij de operationele cybersecurity. Deze werden tijdens een mondelinge pitch geïnformeerd over het doel, de achtergrond, en het soort gegevens waaraan het CSD behoefte had. De pitch slaagde er over het algemeen in het nut van het project duidelijk te maken, en nagenoeg altijd werd een toezegging gedaan om de gegevens binnen een afzienbare periode (meestal binnen een maand) te verstrekken. Echter, ondanks deze toezeggingen bleek het daadwerkelijk verkrijgen van de gegevens zeer lastig te zijn. Uit de reacties van de organisaties kon geconcludeerd worden dat één van de volgende twee redenen nagenoeg altijd een rol speelde.

Complexiteit

Bij veel organisaties die zijn benaderd is de digitale architectuur vaak zeer complex. Dit komt voornamelijk omdat bij deze organisaties de afgelopen jaren veel onafhankelijke ICT diensten zijn samengevoegd tot één centrale afdeling die nog te maken heeft met een grote verscheidenheid aan systemen. Hierdoor is het vaak lastig voor de organisatie zelf om een uniform overzicht te hebben van de cybersecurity status, laat staan om de gevraagde aspecten te delen met derden.

Gegevensgebrek

Tijdens het daadwerkelijk inventariseren van de gegevens die een organisatie in huis had, bleek dat de gedane beloften niet waar konden worden gemaakt. Vaak werd er toch minder bijgehouden of bleken gegevens minder lang bewaard te worden dan gedacht. Dit probleem was niet geheel onverwacht, want ook het CyberDew onderzoek van het WODC had al laten zien dat het ondoenlijk is om alle gegevens langdurig op te slaan. De Hogeschool Rotterdam was partner binnen het project CyberDew en heeft daadwerkelijk ook Netflow data geleverd. De data had betrekking op een periode van 3 maanden. Data dat ouder zijn dan 3 maanden worden weggegooid, simpelweg omdat er te veel data wordt gegenereerd.

Privacy

De gevoeligheid van de data en privacy factoren waren bezwaren die door alle organisaties weren aangekaart tijdens de pitch. Met name het delen van privacy gevoelige attributen (zoals IP adres) of organisatie specifieke informatie, maar ook het feit dat de gegevens potentieel gedeeld zouden worden met derde partijen of de publieke sector werden gezien als een probleem. Ondanks toezeggingen dat gegevens alleen op geaggregeerd niveau worden gepresenteerd zodat resultaten niet meer naar individuele organisaties zijn te herleiden, en dat (privacy) gevoelige attributen direct worden geanonimiseerd is het mogelijk dat dit niet als voldoende garanties worden beschouwd.

Vertrouwen/baten

Twee verschillende concepten die bij het uitwisselen van gegevens beide een belangrijke rol spelen zijn vertrouwen en het feit dat de organisatie die gegevens verstrekt er ook baat bij moet hebben. Echter, het zijn ook concepten meestal niet direct worden genoemd als reden om geen gegevens te kunnen leveren. Dit maakte het lastig om te bepalen welk van de twee domineerde, en daarom is besloten ze hier onder dezelfde noemer te scharen. Desalniettemin bestaat de indruk dat vertrouwen een noodzakelijk voorwaarde is voor het delen van data en een element in het wegnemen van genoemde bottlenecks. Als er een vertrouwensrelatie bestaat tussen organisaties is de kans groter dat zij best practices ten aanzien van de bottlenecks, bijvoorbeeld beheersen van de complexiteit van systemen, zullen delen.

5 Lessons learned

Ondanks initiële toezeggingen van ongeveer een dozijn organisaties is het een grote uitdaging gebleken om daadwerkelijk gegevens over malware en DDoS voor het CSD te verkrijgen. Hiervoor kunnen de vier oorzaken die hierboven zijn genoemd worden aangewezen. Privacy en vertrouwen/baten zijn in de cybersecurity literatuur bekende obstakels bij het delen van gegevens. Echter, de complexiteit en gegevensgebrek zijn niet eerder onderkend, en zijn ook nog eens fenomenen waar nauwelijks invloed op uit te oefenen valt en die in de nabije toekomst moeizaam zullen veranderen.
Terugkijkend op het volledige onderzoekstraject, beginnende bij het haalbaarheidsonderzoek van het cybersecurity dashboard in 2013 en eindigend bij Cybermetrics 2016, is het duidelijk geworden dat het opvragen van gegevens bij partijen geen levensvatbare aanpak is. Hoewel tijdens dit volledige traject meerdere keren is getracht om verschillende partijen te betrekken bij het onderzoek, is dit nooit goed van de grond gekomen. Achteraf gezien is dit misschien niet zo heel vreemd, want veel van de uitdagingen waarvoor wordt gewaarschuwd in handboeken over data uitwisselingsinitiatieven in het cybersecurity domein hebben bij Cybermetrics 2016 (en de voorlopers) ook een rol gespeeld. Ten eerste waren er niet voldoende drijfveren voor de benaderde partijen om in Cybermetrics 2016 (en de voorlopers) te investeren. Ten tweede is er te weinig getracht om vertrouwen op te bouwen bij de verschillende partijen. Ten derde is niet voldoende geapprecieerd dat de digitale architectuur van de benaderde partijen complex is en dat het extraheren van gegevens inspanning kost. Tot slot werd vooraf niet gerealiseerd hoe weinig data en hoe kort de verschillende partijen deze nu echt bewaren.

6 Conclusies

Het is duidelijk geworden dat het verzamelen (met name door de publieke sector) van cybersecurity data bij verschillende partijen niet werkt. Het is dan ook niet mogelijk gebleken om ook maar één van de geformuleerde doelstellingen te behalen. Een logisch gevolg hiervan is dat, buiten dit onderzoeksrapport en de wetenschappelijke paper die als bijlage is toegevoegd, de overige deliverables ook niet kunnen worden opgeleverd.
De roep naar een betere kwantitatieve onderbouwing van de dreigingen van cybersecurity op zowel nationaal niveau, binnen sectoren, als bij individuele organisaties blijft echter aanwezig. Mocht er in de toekomst een nieuw initiatief zijn om hier onderzoek naar te doen dan moet er gedacht worden aan een andere opzet dan die van het CSD/Cybermetrics 2016. Met name zal vooraf moeten worden nagedacht over de drijfveren en het vertrouwen bij de partijen die de data moeten leveren. Hierbij zal ook de inspanningen in rekening moeten worden genomen om de data te extraheren en de tijd die nodig zal zijn om een minimaal bruikbare set op te bouwen. Dit betekent waarschijnlijk dat kleinschalige data delingsinitiatieven, waarbij alle leverende partijen direct profiteren en de publieke sector pas op termijn, een grotere kans van slagen hebben dan de opzet van CSD/Cybermetrics 2016.

Bijlage 1 Challenges around Measuring Cyber Security Threats in the Netherlands

Rémon Cornelisse[1], Mortaza S. Bargh[1], Kas Clark[2], Sunil Choenni[1,3]

Abstract

The main challenges in obtaining and collecting quantitative information about the size, scope and trends of cyber threats in the Netherlands are presented. These challenges are mainly due to: the complexity of the architecture of the organizations, lack of data, privacy concerns, and lack of trust from and/or benefits for the participating organizations. The privacy concerns and lack of trust/benefits were known issues, but the other two are new. Both the complexity of the architecture and the lack of data also need to be considered when requesting information about cyber security threats from large organizations. While the complexity of the digital architecture of an organization is a phenomena that will be difficult to change in the near future, explaining to organizations that keeping an aggregated set of the absolute minimum amount of attributes that are useful for tactical and strategic purposes could solve the lack of data in the future.

1 Introduction

As in many countries, cyber security is high on the social and political agenda in the Netherlands. The national government invests millions of euros in securing the cyberspace to provide a safe and reliable platform for Internet users and e-services (like e-commerce and e-government). To this extent, the Dutch government established the first National Cyber Security Strategy in 2011 [25]. This document stresses the growing need for detecting cyber threats and the urgency to increase the resilience of the cyberspace. One result of [25] is the yearly publication of the Cyber Security Assessment Netherlands (CSAN) report by the Dutch National Cyber Security Centre (NCSC-NL). This report publishes statistics on incidents and analyses new developments and trends in cyber security and often influences policy decisions, both for the government and in the private sector, in this domain.

Given the insight provided by the CSAN report, policymakers steadily request more quantitative analysis to substantiate the key findings of this report. However, measuring the size and scope of cyber threats is a notoriously difficult problem (see [9] and references therein). According to [9] no reliable, consistent, longitudinal data exists to get an accurate overview of the problem, and most surveys to date are of a dubious methodology. This is mainly due to the fact that most cyber threats are extremely concentrated, so that representative sampling of the population does not give a representative sampling of the threats [13]. Another, compounding

[1] Research and Documentation Centre, Ministry of Security and Justice, The Hague, The Netherlands

[2] National Cyber Security Centre, Ministry of Security and Justice, The Hague, The Netherlands

[3] Research Centre Creating 010, Rotterdam University of Applies Sciences, Rotterdam, The Netherlands

problem is that the intended motive/aim of the data provider, i.e. altruistic or commercially interested, is difficult to quantify [5]. Therefore, any related data should be regarded with skepticism and its validity should be questioned. These problems become even more pronounced when one is mainly interested in gaining insight in cyber threats on the national level, i.e., the Netherlands in this case, since most commercial reports do not specify the cyber threat level per country.

One way to overcome the current limitations is to survey that part of the population that definitely encounters cyber security threats on a regular basis, i.e. focus on a representative set of organizations that have these problems. One such set of organizations includes those that belong to the critical infrastructure[4], i.e. organizations that are essential for the functioning of society and the economy. However, a survey that is only limited to (cyber security) experts will introduce biases (e.g., see [18] and references therein). In the case of cyber security, it is possible to avoid these expert biases by directly accessing the relevant information stored in data logs from specific cyber security devices at the relevant organizations. However, this requires some form of data sharing, which has its own problems such as the necessity for all entities to benefit from the cooperation, loss of privacy, and the need of some form of trust management [26].

In this paper we will describe our experiences with setting up a data sharing initiative to obtain tactical and strategic information on cyber threats using data from cyber security devices of the key organizations that belong to the critical infrastructure in the Netherlands. Using our experiences with interacting with over a dozen different organizations, the focus will be on the challenges that were encountered and the lessons learned from setting up such an initiative. Section 2 provides a short background of the project, and Section 3 describes different kinds of information sharing models. In Section 4 the experiences of the initiative are presented, with Subsection 4.1 describing the challenges encountered and Subsection 4.2 presenting the results. In Section 5 the obstacles (Subsection 5.1) and a way forward (Subsection 5.2) are discussed. Finally, Section 6 finishes with some conclusions.

2 Background

To improve on the limited amount of quantitative validation of the findings in CSAN it was decided to develop a dashboard, a popular visualization tool that integrates key performance metrics with a summary of the underlying operational data into a single display (e.g. [21]). For this so-called Cyber Security Dashboard (CSD) the goal is to create meaningful visualizations for policymakers using relevant quantitative data in order to understand trends, evaluate specific mitigation policies, and predict future developments. For example, one goal of the CSD is to translate data from attack registration systems to a prioritized list of attack types to guide investment in mitigation. Furthermore, to visualize the effectiveness of past investments the data from attack registration systems can be combined with additional indicators of financial losses and average security budgets. Finally, a timeline of major cyber incidents on a global level would allow analysts to detect attack patterns that could predict future cyber security incidents and developments.

[4] en.wikipedia.org/wiki/Critical_infrastructure (last consulted May 2017).

During the first phase of the development of the CSD the information needs of the different (potential) target groups were collected using a top down approach. In other words, senior cyber security officials from different organizations provided input on the key indicators that should be present in the CSD. Furthermore, an information model was established to provide insight in the kind of information that should be collected. According to this model, information should be collected on the assets, threats, resilience and manifestations for each cyber security question [11]. For the development of the CSD, an iterative process is adopted. The process started with a simple dashboard that was improved step-by-step in multiple rounds. In each round the lessons learned during the previous rounds are used in according to the principles and approach of action design research [30]. Finally, a first set-up was created using publicly available data sources. However, often these data sources are not focused on the Netherlands and are biased due to the commercial interests of their providers.

During the second phase it was decided to focus on obtaining information on two specific cyber security threats, i.e., over the Distributed Denial of Service (DDoS) and malware attacks. DDoS attacks attempt to flood a machine that is connected to the Internet with a large amount of (superfluous) requests from multiple locations to overload it and prevent legitimate requests from being fulfilled, thereby disrupt its services temporarily or indefinitely.[5] Malware, short for malicious software, is any software used to disrupt computer or mobile operations, gather sensitive information, gain access to private computer systems, or display unwanted advertising.[6] The main questions to be visualized via the CSD are related to the typical characteristics and frequency/size of both DDoS and malware attacks, and should be answerable using only the data logs from specific mitigation and/or detection equipment (e.g. anti-virus software, anti-malware/ DDoS equipment, etc.). To obtain these logs on DDoS and malware attacks the choice was made to approach a group of representative organizations for a data sharing initiative, and in the remainder of the paper the experiences with setting up this initiative are described.

3 Information Sharing

The Internet ecosystem comprises a vast number of stakeholders who collaboratively govern the well functioning of the Internet. These stakeholders include, for example, Internet Service Providers, Domain Name Service Providers, and Application Service Providers. These stakeholders – being associated with a wide range of national and international institutions, public and private sectors, and academia and industries – constitute a community of peers who independently and collaboratively work across the globe [34].

To build a robust ecosystem for the Internet, these stakeholders contribute to the Internet governance processes in a participatory and bottom-up way. A key requirement for governing the Internet such inclusively (and for maintaining the stability, integrity, and open nature of the underlying technologies and systems [34]) is reliance on the principles of transparency and information sharing among the stakeholders. Particularly concerning the security aspects of the Internet, there

[5] en.wikipedia.org/wiki/Denial-of-service_attack (last consulted May 2017).

[6] en.wikipedia.org/wiki/Malware (last consulted May 2017).

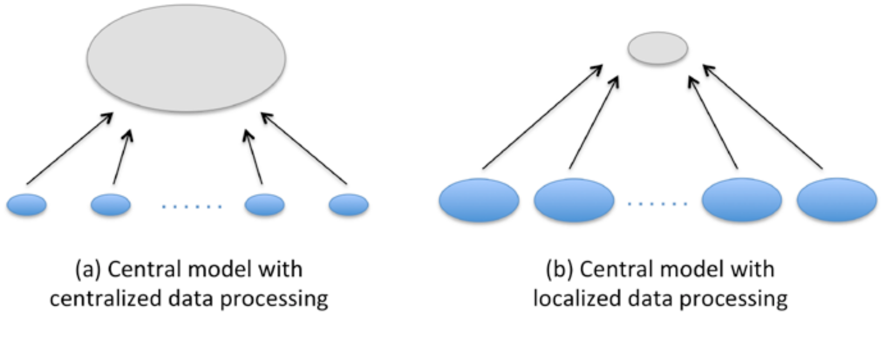
is no central entity to monitor the Internet and enforce required security measures. Therefore, various Internet stakeholders must collaborate with each other and, among others, share information in order to resolve or mitigate cyber security issues. This is called collaborative security approach, which strongly relies on the underpinning "participatory" and "multi-stakeholder" principles of the Internet [35].

3.1 Data Sharing Models

Data sharing among Internet stakeholders can create situational awareness about, in this case, cyber security issues. This awareness can be at various levels, ranging from operational level (e.g., receiving alarms about on-going cyber attacks for incident response purposes) to strategic level (e.g., receiving trend indications for strategic decision making purposes). Data sharing can be done in a peer-to-peer (i.e., fully distributed) way or via a (trusted) central party. As a generalized form of the latter, the hierarchical data sharing can also be recognized, where data are shared between those organizations that have a parent-child relation on the tree-like structure of the corresponding organizations. In this contribution we discuss and focus on the centralized data sharing. We argue that the discussion results in Section 3.2 can readily be applied to the hierarchical and the peer-to-peer models.

Figure 1 shows the centralized data sharing, for which two sub-models are recognized: centralized data processing and localized data processing, depending on the amount of the data processing done centrally or locally. In the centralized data processing model (see Figure 1a), every participating organization shares its raw data with the central (trusted) party to process the data centrally and to calculate a desired function/value of the received raw data. For example, banks share all their customer data (about their savings, investments, etc.) with a central bank to calculate the net or gross profit of all banks. In the local data processing model (see Figure 1b), every participating organization processes its data (e.g., using data mining and statistical data aggregation) before sending them to the central organization for calculating the desired function from all the aggregated data received. For example, banks share their annual profits and losses with a central bank.

Figure 1 An illustration of the centralized data sharing models with (a) centralized data processing and (b) localized data processing (which are represented by thick and thin data processing volumes at the central party, respectively)


(a) Central model with centralized data processing
(b) Central model with localized data processing

3.2 Complexity versus Trust

In this article we shall base our arguments on the amount of complexity (thus efforts) imposed on local organizations and on the amount of trust needed for these local organizations to have in the central organization when sharing their data (either raw or aggregated) with the central organization. Intuitively, we argue that the more intelligent processing is carried out locally (like applying appropriate privacy and information sensitivity protections while preserving the utility of the resulting data), the lower the amount of trust needed to have in the central organization (e.g. trusting the central organization not to misuse the acquired data). This inverse relation between trust and complexity is schematically illustrated in Figure 2.

Figure 2 An illustration of the complexity versus trust when sharing raw to processed data

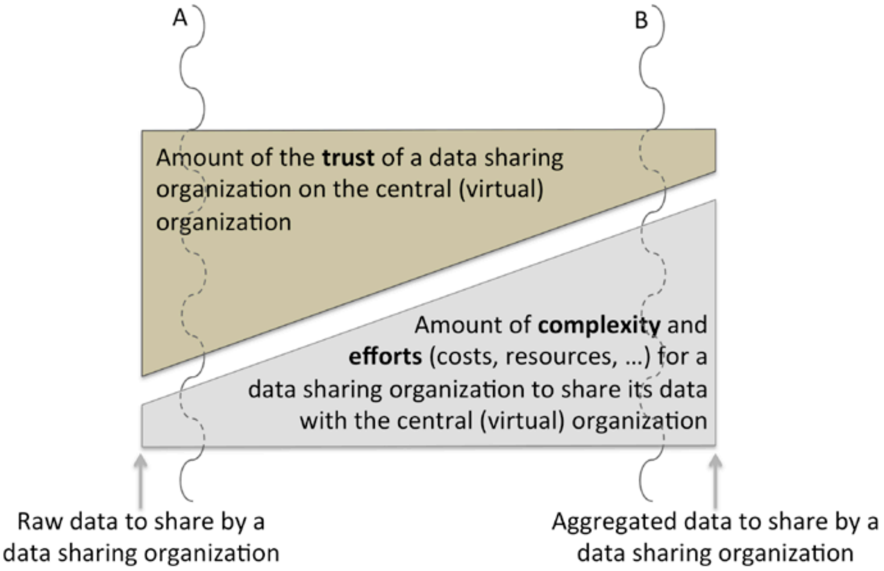

A
B
Amount of the **trust** of a data sharing organization on the central (virtual) organization
Amount of **complexity** and **efforts** (costs, resources, ...) for a data sharing organization to share its data with the central (virtual) organization
Raw data to share by a data sharing organization
Aggregated data to share by a data sharing organization

Figure 2 also illustrates two extreme cases A and B, distinguished with respect to the rawness of the shared data. At point A the shared data are rather raw and therefore the amount of the efforts spent on data processing is low. However, there is a rather high possibility that some information sensitivity issues arise after sharing the raw data. Therefore, the local organization needs to have a high amount of trust on the central organization. The situation is reverse at point B. Here the shared data are thoroughly processed and the processing complexity is high, so every local organization has enough confidence that (almost) nothing can go wrong if the data are shared. Applying such intelligent data processing locally allows a local organization to have low amount of trust in the centralized organization. In case of multi-party secure computing, for example, the amount of complexity at the local organizations is very high but instead the amount of trust in the central organization is very low, as the central entity does not even learn anything about the (aggregated) data given as input to the central entity.

4 Experiences

Since it was unclear from the outset that information contained in the logs from the mitigation and/or detection equipment was sufficient to visualize the main questions on DDoS and malware for the CSD, a bottom-up approach was used to obtain the required data. In particular, those experts were contacted that were directly involved in the operational cyber security at the organizations that in the past had already been willing to provide us with their data, and the organizations that were expected to be easily convinced of the added value of a CSD. These organizations were chosen since they were thought to be the most likely candidates to quickly obtain data from and to determine if the information in the data logs would be sufficient for the CSD. Furthermore, as an extra incentive for cooperation, the aim was to impose no or as-little-as-possible burden on the organizations, i.e., to operate as close as possible near line A in Figure 2 where an organization requires little effort but has a lot of trust in the data sharing initiative. The shared data could then be used to create a first set-up of the CSD. This could then be used to convince other organizations to join the initiative, as it would clearly demonstrate the added value.

4.1 The challenges

In total, about a dozen organizations were approached. During visits to each organization a short pitch was given on the goals, background, and the kind of information that was needed for the CSD. In general this pitch succeeded in conveying the usefulness of the project, and almost always a promise was made to provide relevant data in the near future (typically within a month). However, despite these promises it turned out to be a larger challenge to actually receive the data than expected. Based on the reactions of these organizations, it can be concluded that at least one of the following reasons always played a key role.

Complexity
Given that the organizations belong to the critical infrastructure, and are thus typically very large, they often have complex digital system architectures. This is in particular due to the many mergers undergone in the last few years, whereby multiple previously independent ICT departments were merged into a single one. Often this single department still has a large variety of systems, making it difficult for the organization itself to have a uniform overview of the cyber security status, let alone sharing the required data with a third party.
The complexity of an organization was also indicated in more indirect ways, and the provided reasons can best be summarized using the concepts of immaturity and viscosity. Although these concepts were not directly mentioned, they could be deduced from the reactions during the pitch. For example, in the case of immaturity it became obvious from the questions during the pitch that the cyber security awareness and mitigation strategies of an organization were not well known (or/and developed) yet. They could typically not indicate what kind of data were present at all, and they immediately mentioned that it would be difficult to extract any kind of data from their system. The other indirect indication that complexity was a problem was the viscosity of the organization as a whole. The problem of organizations with a high viscosity in general is that it is challenging for them to perform a non-standard task. For example, organizations with a (suspected) high viscosity typically indicated immediately that a period of a month was too short to fulfill the request, while they were often incapable to give a deadline in which they could.

Lack of data

During the pitch almost all organizations indicated that data was only kept for a limited amount of time, typically ranging from several days to several months, before everything was removed. However, during the actual inventory of the data that an organization possessed, it turned out that even these estimates were too optimistic. Often, less data was saved than expected or it turned out that data was not kept as long as was expected.

Privacy

Privacy and information sensitivity were an issue that was raised by all organizations. Most of the time this was about the fact that privacy sensitive attributes (such as IP addresses) or organization specific information would be shared, but also the fact that the data would be with third parties or the public sector (as by law the latter is supposed to make its data open in certain circumstances). Despite assurances that the data would only be presented at an aggregated level to make sure that the results could not be traced back to individual organizations, and the promise that (privacy) sensitive attributes would directly be anonymized, is it possible that these were not considered to be adequate guarantees.

Benefit

In addition to trust, another important ingredient for a successful data sharing initiative is the fact all entities need to benefit from the cooperation [26]. Although trust and benefit are different concepts, they are related. Furthermore, they are also concepts that are often not explicitly stated when reasons were given over why the data could not be delivered. This makes it difficult to judge which of these two concepts dominated, and it was therefore decided to treat them together.

The easiest reason to understand was when an organization immediately indicated to us that they could not spend time on the collection of data unless there was a clear compensation of immediate use to them. Another, often heard, reason was a lack of time, which is most likely a euphemism for considering it as a very low priority job. Since the follow-up request for the data typically occurred several weeks after the pitch, the initial enthusiasm about the CSD project would have disappeared. Coupled with the fact that providing data once feels like committing to a periodic data delivery exercise for a project for which the usefulness for the organization itself is unclear, makes it understandable that organizations were hesitant to share their data.

4.2 The results

Despite the fact that the majority of the organizations that were approached did not provide us with data, there were exceptions. A small amount of data was obtained, which provided several important lessons for the CSD. First of all, the information in a typical malware should be sufficient to obtain relevant tactical and strategic information. However, the log of a single day from an organization that belongs to the critical infrastructure appears to be 0.2-0.5 Gbyte of text, contains around 50-70 columns, and is made up of around a quarter of a million records. These numbers are more than double the 80 events per minute mentioned by [27] in their 2015 survey, and provide a hint why these logs are not stored for an extended period of time. They are simply too large to efficiently handle manually, and will only be consulted when a single (set of related) event(s) need to be isolated using dedicated cyber forensics tools.

Another important lesson that could be learned is that many columns in those logs contain ‘unnecessary’ or duplicate information. For example, in one log a column containing the date and time of an event occurred four times, while many other columns were degenerate (i.e., each record contained the same number or the same expression such as for example ‘THREAT’). All of this makes a typical log highly unreadable when a cyber security expert opens one. Furthermore, many of the columns are (obviously) strongly correlated. For example, clicking on a link to download malicious content will always be done via the same medium (i.e. web browsing) using the same protocol (i.e. TCP) and the same port number (i.e. 80) resulting in the same warnings (e.g. ‘incoming via the http server’; ‘anonymous file’). Finally, a typical log does contain privacy sensitive information. These are not only the obvious ones such as IP addresses, but also organization specific information. Although this information is most likely not sensitive enough to endanger the security of the organization, it could easily be another obstacle to share these logs.

5 Discussion

5.1 Obstacles

In what appears to be the first systematic study on the cost of cyber threats, [3] already point out that, despite 100 different sources of data on cyber threats, the available statistics are still insufficient and fragmented. Many reasons for this lack of reliable statistics are provided by [3], ranging from intentional errors (e.g. from players in the cyber security field that have a strong interest in playing up threats) to unintentional errors due to response effects or sampling bias. These findings are supported by McAfee, a for-profit security software company, that also highlighted the lack of reliable data in their report [23]. Also [5] concluded that without large improvements in quantification and measurement it will not be possible to solve the problem of cyber threats in the near future.

In order to make a new attempt on tackling the problems of quantitatively measuring the size of cyber threats and making a CSD based on those measurements, a data sharing initiative was started among those organizations that are part of the critical infrastructure and the public sector. Taking into account the relatively young age of the field of cyber security, it already has a long tradition of praising the virtues of sharing cyber security information (e.g. [29], [2]). However, it appears that not much progress has been made, and that even nowadays these calls for cooperation are still needed (e.g. [17], [19], [15]). Several reasons have been identified why it is difficult to share information, such as the necessity for all entities to benefit from the cooperation, loss of privacy and the necessity of some form of trust management [26]. Although most of these reasons were also encountered in our work, the CSD found two reasons that have not previously been identified with regard to cyber security, namely the complexity of the digital systems of organizations and the lack of data. However, it should not come as a surprise that these problems also play a role in cyber security, since they have already been identified as major challenges to overcome in other fields.

The problem of the complexity of organizations has been known for a long time [31], and has an extensive literature (e.g. [4] for an overview). Complex organizations are nonlinear systems that evolve over time due to the pressures for competitiveness, flexibility or dexterity. This is in particular true for cyber security departments. Since they have seen large changes due to the fast evolving nature of cyber threats [20], their evolution has lead to a centralization of the departments. This centralization is an ongoing process and is leading to many legacy systems at the organizations that still need to be maintained. Since there are no clear signs that this evolution is slowing down in the near future, it will be unlikely that this problem with the presence of legacy systems will soon disappear.

The lack of data is also not an unexpected problem. With the still exponentially increasing amount of network traffic [8], it should be expected that also the number of potential security incidents is still increasing rapidly and therefore the amount of data that are produced by the different detection/mitigation systems is increasing. This problem is similar to the challenges that big data encounters with its large scale and sheer volume [6]. In particular the ubiquity and dynamic nature of the various data generation devices complicate the storing and processing of cyber security data. Since the core duties of the cyber security experts are incident detection, incident response, and vulnerability assessment [33], there is typically no need to analyze all the data from mitigation/detection systems for strategic purposes or to keep this information after an incident has been resolved.

Another complicating factor is the problem that most cyber security experts do not appear to have a good overview of the larger picture [5], thereby introducing expert biases if only their knowledge was used. The survey by [5] amongst stakeholders involved in the fight against cyber threats found that “better metrics and statistics on cybercrime” was indicated as one of the most important topics that needed research. Furthermore, [5] found that most respondents declared that the main consequence of the cyber threats they encountered was ‘only some inconvenience’, although they also claimed that their country or the worldwide economy suffered enormous losses as a result of cyber threats. Similar results were also found in a survey amongst cyber security experts working in the critical infrastructures in the Netherlands, where most respondents could also not give an estimate of the financial losses due to cyber threats [32].

5.2 Way Forward

In hindsight it is easy to understand why the data sharing initiative for the CSD failed. First of all, not enough incentives were created for the participating organizations to invest some efforts and resources. Second of all, not enough trust was created to start a data sharing initiative where a public organization was involved (as by law public organizations are supposed to make all their data open in certain circumstances). Third of all, it was not properly acknowledged that the digital architecture of an organization is complex and data extraction will take effort. Finally, it was not realized that many organizations only keep (a limited amount of) data for a limited period. Therefore, any attempt of a new data sharing initiative should at the very least try to address these problems.

Creating enough incentives for the participating organizations to invest some efforts and resources is challenging. For most organizations it should be clear that the investments will have an added value. This added value should be large enough to make the participation beneficial. In the field of economics the literature on the theories of information sharing is already extensive (e.g. [14], [28], [24]). The results of these theories have also been applied to cyber security (e.g. [16], [29], [19]). In their overview, [22] enlist (amongst others) that a strong incentive that enables data sharing is that the shared information has a clear content, is actionable, is of immediate use to the recipient, and saves costs. This claim is further supported by a survey of the ENISA network that shows that the two top incentives for data sharing are timely access to valuable and relevant information, and enabling the efficient allocation of information security resources and cost savings [12]. Two keywords in both studies are actionable and immediate use, and were both absent in our original set-up. A new data sharing initiative should make these keywords central and operational.

Creating enough trust to start a data sharing initiative might actually be a simple problem to solve. An interesting result from the survey by [32] was the fact that 85% of the cyber security experts working in the critical infrastructures in the Netherlands indicated that they shared information about the most recent security event with the other organizations in the same sector. Furthermore, 80% also indicated that in the past year they increased collaboration and/or information sharing with the other organizations in the same sector. This sharing is done via a so-called Information Sharing and Analysis Center (ISAC), linking sector specific knowledge networks on cyber security. This suggests that, at least in the Netherlands, the sharing of (aggregated) information at sector level is taking place. Since [22] point out that a good practice for data sharing is to make use of existing initiatives, these ISACs appear to be the perfect organizations to approach. They already have the trust of the organizations within their specific sector, and they should also have an interest in the tactical and strategic view of the cyber threats for their sector.

The final two problems are closely related and are treated simultaneously. The current solution to the large amounts of cyber security data being produced is to discard everything after a short period. The other extreme, keeping all data ‘forever’, is neither feasible nor desirable, but a compromise where only a limited amount of data will be kept for a longer time is a good alternative. Such a compromise is standard practice in fields that truly have a big data problem, namely that of particle physics and astronomy. For example, the ATLAS experiment located at the CERN Large Hadron Collider only selects and stores the events with potentially interesting physics [1], while the astronomical satellite mission Kepler can only send 6% of the total information collected back to Earth due to the limited bandwidth [7]. Furthermore, [10] showed that also for cyber security it is not always necessary to store all network traffic, but keeping only a few key characteristics is sufficient for a good description of the original data.

The same principle, i.e. only keep the most relevant part of the original dataset, can be applied to the few logs that were obtained for the CSD. For example, all the degenerate columns, double columns, and columns with strongly correlated information can be removed with only a negligible loss of information. Furthermore, to avoid extra complications for the data sharing initiative, it is best to remove all columns with potentially privacy sensitive information such as IP addresses. Finally, there are columns that might potentially be of tactical or strategic interest. For these columns a trade-off needs to be made how much data is truly necessary for the data sharing initiative and the willingness of the organizations to provide this data.

Figure 3 (Top) shows the results of applying these considerations as rigorously as possible to a ‘typical’ log that only contains 10 records. Where the original log consisted of over 60 columns, this has now been reduced to only 5. Not only does this increase the ‘readability’ of the data that will be shared, but almost immediately suggests another reduction to the amount of data that needs to be stored. Since the data sharing initiative is only interested in a high level overview of the cyber security landscape, aggregating the records is an obvious next step. The resulting aggregated overview, which is presented in Figure 3 (Bottom), is only a small fraction of the original one and should make negligible impact on the storage capacity of an organization. However, since the original logs are only kept for a limited time, these aggregated overviews must be created on a regular basis. This is best done in an automated and stable fashion, i.e. the set of attributes that is kept and aggregated changes as little as possible in order to make it a onetime investment in both time and money.

Figure 3 (Top) The resulting dataset after removing all redundant, correlated, and privacy sensitive information from the first 10 records of a typical log
(Bottom) Resulting dataset after aggregating the log above

6-1-2017	file	web-browsing	GZIP	incoming
6-1-2017	file	web-browsing	GZIP	incoming
6-1-2017	vulnerability	ssh	OpenSSH	outgoing
6-1-2017	file	web-browsing	ZIP	incoming
6-1-2017	file	web-browsing	GZIP	incoming
6-1-2017	file	web-browsing	GZIP	incoming
6-1-2017	file	web-browsing	GZIP	incoming
6-1-2017	file	web-browsing	GZIP	incoming
6-1-2017	file	web-browsing	GZIP	incoming
6-1-2017	file	web-browsing	GZIP	incoming

6-1-2017	file	web-browsing	GZIP	incoming	8
6-1-2017	vulnerability	ssh	OpenSSH	outgoing	1
6-1-2017	file	web-browsing	ZIP	incoming	1

Combining the ideas listed above, a new approach to the CSD would combine a localized data processing model (i.e. the model in Figure 1b) with a complexity versus trust ratio closer to point B in Figure 2. In practice this means that individual ISACs in the Netherlands should be approached with the idea of a public-private initiative. The goal of the initiative is to share highly aggregated cyber security data of the organizations within their sector as a way to obtain a tactical (i.e. short term) overview of the size of the cyber threat problem. This will immediately give an indication how an organization is doing compared to its peers (useful information), and what areas of cyber security need more attention (actionable information). The role of the public sector is to somehow provide support in extracting the data (acknowledge that it is complex) and to provide the aggregated overview for the other sectors (in order to build up trust). Obviously the way the overview of the other sectors can be presented will depend on the individual ISACs, since they should act as the gatekeeper and decide on how information can be shared with third parties (to build trust). Since this information should also be interesting for strategic (i.e. long term) purposes, it should be worthwhile to create these aggregated overviews regularly and therefore in an automated fashion (convince them that keeping some data is better than none).

6 Conclusions

Despite the initial promises from over a dozen organizations, it turned out to be a major challenge to actually obtain the promised data on DDoS and malware attacks for the CSD. This can be attributed to several reasons, namely the complexity of the architecture, a lack of data, privacy concerns and a lack of trust from and/or benefits for the organizations. The privacy concerns and lack of trust/benefits were already known issues when starting an information sharing initiative. The other two reasons, the lack of data and the complexity of the architecture, are issues that have not been reported before. Both issues also need to be considered when requesting information about cyber security threats. While the complexity of the digital architecture of an organization is a phenomena that can hardly be influenced and will be difficult to change in the near future, explaining to organizations that keeping an aggregated set of the absolute minimum amount of attributes that are potentially useful for tactical and strategic purposes could solve the lack of data in the future.

Since the call for a better quantitative understanding on cyber security threats continues to exist, a data sharing initiative similar to the CSD will always be considered. Although no guarantees can be given for future success, the lessons of the CSD suggests using the ISACs of specific sectors to share actionable and directly useful aggregated information to obtain a tactical and strategic overview of the cyber threats are essential ingredients for a successful private-public collaboration.

7 References

[1] G. Aad, ATLAS Collaboration, et al., ‘The ATLAS experiment at the CERN large hadron collider’, J. Instrum., 3 2008

[2] Anderson, Ross, ‘Why Information Security is Hard: An Economic Perspective’, Proceedings of 17th Annual Computer Security Applications Conference, 2001, Dec. 10-14.

[3] Anderson, R., Barton, C., Böhme, R., Clayton, R., Van Eeten, M. J., Levi, M., ... & Savage, S, ‘Measuring the cost of cybercrime’ In The economics of information security and privacy (pp. 265-300). Springer Berlin Heidelberg, 2013

[4] Arévalo, Luz E. Bohórquez, and Angela Espinosa. ‘Theoretical approaches to managing complexity in organizations: A comparative analysis.’ Estudios Gerenciales 31.134, pp. 20-29, 2015

[5] Armin, Jart, Bryn Thompson, and Piotr Kijewski. “Cybercrime Economic Costs: No Measure No Solution.” Combatting Cybercrime and Cyberterrorism. Springer International Publishing, pp. 135-155, 2016

[6] P. Barnaghi, A. Sheth, C. Henson, From data to actionable knowledge: big data challenges in the web of things, IEEE Intelligent Systems, 28 (6), pp. 6–11, 2013

[7] Borucki, W.J., Koch, D., Basri, G., Batalha, N., Brown, T., Caldwell, D., et al., ‘Finding Earth-size planets in the habitable zone: the Kepler Mission’, Proceedings of the International Astronomical Union, IAU Symposium, Volume 249, pp. 17-24, 2008

[8] Cisco Systems 2015, ‘Cisco Visual Networking Index: Forecast and Methodology, 2014-2019’, 2015

[9] Cobb, Stephen, 'Sizing Cybercrime: Incidents and accidents, hints and allegations', Proceedings of the 25th Virus Bulletin International Conference. 2015.

[10] Cornelisse, Rémon, et al. 'Compressing Large Amounts of NetFlow Data Using a Pattern Classification Scheme', Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 2016 IEEE 2nd International Conference on. IEEE, 2016.

[11] CSAN (2013), 'Cyber Security Assessment Netherlands 2013', National Cyber Security Centre Netherlands, 2013, www.ncsc.nl/english/current-topics/Cyber+Security+Assessment+Netherlands/cyber-security-assesment-netherlands-2013.html

[12] ENISA (2010), 'Incentives and Challenges for Information Sharing in the Context of Network and Information Security', ENISA, Heraklion,Greece, 2010, www.enisa.europa.eu/act/res/policies/good-practices-1/information-sharing-exchange/incentives-and-barriers-to-information-sharing/at_download/fullReport

[13] Florêncio, D., Herley, C., 'Sex, lies and cyber-crime surveys', Economics of information security and privacy III, New York: Springer pp. 35–53, 2013, research.microsoft.com/pubs/149886/SexLiesandCybercrimeSurveys.pdf

[14] Fried, D., 'Incentives for Information Production and Disclosure in a Duopolistic Environment', Quarterly Journal of Economics, 99(2), 367, 1984

[15] de Fuentes, J. M., González-Manzano, L., Tapiador, J., & Peris-Lopez, P., 'PRACIS: privacy-preserving and aggregatable cybersecurity information sharing', Computers & Security, 2016

[16] Gal-Or, Esther, and Anindya Ghose., 'The economic incentives for sharing security information', Information Systems Research 16.2, pp. 186-208, 2005

[17] Garrido-Pelaz, R., González-Manzano, L., Pastrana, S., 'Shall we collaborate?: A model to analyse the benefits of information sharing', In: Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security. pp. 15–24. ACM, 2016

[18] Garthwaite, Paul H., Joseph B. Kadane, and Anthony O'Hagan, 'Statistical methods for eliciting probability distributions', Journal of the American Statistical Association 100.470, pp. 680-701, 2005

[19] Gordon, L.A., Loeb, M.P., Lucyshyn, W., Zhou, L., 'The impact of information sharing on cybersecurity underinvestment: a real options perspective', Journal of Accounting and Public Policy 34(5), pp. 509–519, 2015

[20] Julian, 2014, www.infosecurity-magazine.com/opinions/the-history-of-cybersecurity/

[21] LaPointe, Patrick, 'Marketing by the Dashboard Light', New York: MarketingNPV/Association of National Advertisers, 2005

[22] Luiijf, H. A. M., and A. C. Kernkamp, 'Sharing Cyber Security Information: Good Practice Stemming from the Dutch Public-Private-Participation Approach', 2015, www.tno.nl/en/focus-areas/defence-safety-security/cyber-security-resilience/sharing-cyber-security-information/

[23] McAfee & CSIS, 'Net Losses: Estimating the Global Cost of Cybercrime - Economic impact of cybercrime II', 2014, www.mcafee.com/us/resources/reports/rp-economic-impact-cybercrime2.pdf

[24] Myatt, D. P. and Wallace, C., 'Cournot competition and the social value of information', Journal of Economic Theory, 158(B), pp. 466-506, 2015

[25] NCSS (2011), ‘De Nationale Cyber Security Strategie(NCSS) – Slagkracht door samenwerking’, Dutch Ministry of Security and Justice, 2011, www.rijksoverheid.nl/documenten/rapporten/2011/02/22/nationale-cyber-security-strategie-slagkracht-door-samenwerking

[26] B. Petrenj, E. Lettieri, and P. Trucco, ‘Information sharing and collaboration for critical infrastructure resilience{a comprehensive review on barriers and emerging capabilities’, International Journal of Critical Infrastructures, 9(4), pp.304-329, 2013

[27] PriceWaterHouseCoopers, ‘Global state of information security survey 2015’, 2015, www.pwc.com/gx/en/consulting-services/informationsecurity-survey/assets/the-global-state-of-informationsecurity-survey-2015.pdf

[28] Raith, M., ‘A General Model of Information Sharing in Oligopoly’, Journal of Economic Theory, 71(1), pp. 260–288, 1996

[29] Schechter, Stuart E., and Michael D. Smith, ‘How much security is enough to stop a thief?’ International Conf. on Financial Cryptography. Springer Berlin Heidelberg, 2003.

[30] M. K. Sein, O. Henfridsson, M. Rossi, and R. Lindgren, ‘Action Design Research’, MIS Q., vol. 35, no. 1, pp. 37–56, 2011.

[31] Thompson, James D., ‘Organizations in action: Social science bases of administrative theory’, Transaction publishers, 1967.

[32] van Wilsem, J., ‘Cybercrime in de Vitale Infrastructuur en de Relatie met de Politie (CVIP)’, 2016, private communication

[33] C. Zimmerman, ‘Ten Strategies of a World-Class Cybersecurity Operations Center’, Bedford, MA: The MITRE Corporation, 2014

[34] Internet society webpage, Internet governance - why the multistakeholder approach works, April 26, 2016, www.internetsociety.org/doc/internet-governance-why-multistakeholder-approach-works (consulted June 2017)

[35] Kolkman, Olaf, ‘Trust Isn’t Easy: Drawing an Agenda From Friday’s DDoS Attack and the Internet of Things, weblog, www.linkedin.com/pulse/trust-isnt-easy-drawing-agenda-from-fridays-ddos-attack-olaf-kolkman?trk=hp-feed-, Oct. 25, 2016 (consulted June 2017)