Strategic Cyber Threat Intelligence Sharing: A Case Study of IDS Logs

Spike E. Dog, Alex Tweed, LeRoy Rouse, Bill Chu, Duan Qi, Yueqi Hu, Jing Yang, Ehab Al-Shaer
College of Computing and Informatics
University of North Carolina at Charlotte
Charlotte, NC 28223, USA
{spike.e.dog, atweed87}@gmail.com,{lrouse6, billchu, qduan, yhu12, Jing.Yang, ealshaer}@uncc.edu

***Abstract*—Cyber threat intelligence sharing is emerging as an important tool for network security as it can identify evolving threat patterns and prevent attackers from replicating their early success across the Internet. However the types of information sharing being practiced today are at the tactical level focusing on specific attacks, e.g. characteristics of a piece of malware, and black listed IP addresses and domains. In this paper we argue sharing cyber intelligence at a more strategic level is needed. By strategic information we mean information about salient common features of groups of attacks and attackers. Strategic information allows us to take actions that are much closer to the source of the attacks. For example instead of block an IP address as opposed to shutting down the botnet. We propose at set of strategic cyber threat indicators and show how they can be derived using an IDS log from a large commercial enterprise.**

Keywords— security analytics; information visualization; cyber threat intelligence; machine learning; intrusion detection

I. INTRODUCTION

With increasing sophistication and ever expanding scale of cyber-attacks cyber threat information (CTI) sharing is emerging as a promising mechanism to help prevent large scale cyber-attacks. The idea is for companies and government agencies to timely share cyber-attack information found on their networks. Similar to crimes in the physical world, a cyber-attack is usually successful after repeated attempts and the attacker uses the same or similar techniques against multiple targets. Like crime watch, this community oriented defense strategy can alert all concerned as soon as an attack attempt is discovered so defensive postures can be configured to prevent similar attacks. Mitre Corp, with support from the U.S. Department of Homeland Security (DHS) established Structured Threat Information Expression (STIX), a standard data format for sharing cyber-attack information [9].

Most cyber threat information sharing platforms, e.g. Threat Connect, Hailataxii, X-force, contain threat information at the tactical level. By strategic cyber threat information we refer to information on salient common features of groups of attacks and attackers. Strategic cyber threat information can lead to actions aiming closer to the mastermind of cyber-attacks. Table 1 summarizes some of the key differences between strategic and tactic cyber threat information.

TABLE I. TACTIC VS. STRATEGIC THREAT INFORMATION

	Tactic CTI	**Strategic CTI**
Info. Source	Honey pots, incident reports, logs	Analysis based on substantial amount of tactic info.
Shared Attack source	IP address, domains	Hosting networks, botnets
Action	Firewall blocking	Eliminate botnets, improve hosting policy

The security community needs to identify a framework for sharing strategic CTI leveraging existing definition of STIX and TAXII. First, we must identify a set of specific strategic cyber threat indicators that can be timely shared. Second, we must define analysis methods used to obtain each strategic cyber threat indicator so that (a) effort of data analysis is not duplicated, and (b) strategic CTI from different sources can be compared. In this paper we propose three strategic cyber threat indicators. The first indicator is *network neighborhood*, which groups attacking IP addresses based on the organization that owns the network. The second indicator is *attack type distribution* showing distribution of types of attacks across network neighborhoods. The third indicator is *attack coordination* showing evidence of coordinated activities between attackers. We will demonstrate how to obtain such information using an IDS log from a typical commercial IDS system deployed at a Fortune 100 enterprise.

In Section II of the paper we give a brief overview of the data set we used to illustrate the value of these strategic threat indicators. Section III describes in detail each of the three indicators and how they can be derived from an IDS log. Section IV compares our efforts with related efforts. Section V summarizes our contributions.

II. DATA SET DESCRIPTION AND DATA ENRICHMENT

The data set consists of cyber attack events captured over a six-month period by HP Tipping Point Intrusion Prevention System (IPS) instances placed within the network of a Fortune 100 enterprise. There were over 122 million alerts in the IDS log. Event data was appropriately anonymized to prevent exposure of sensitive network configurations, then incrementally exported to CSV files, which we later used for

978-1-5090-2279-3/16/$31.00 ©2016 IEEE

analysis. Each event contained the following attributes: event time, attack type, source IP, source port, and destination port. Note that as a limitation of our information protection agreement, exported events did not contain destination IP addresses, which prevented any identification of sequential or simultaneous attacks on specific hosts. As attribution of an event to an IP address was fundamental to our analysis, we removed from the data set any events that did not contain a valid source IP address. We considered two reasons for such events to occur. One, that many denial-of-service attacks require or allow an attacker to spoof the source IP address with an invalid one. Two, that each exported CSV file was constrained to a maximum size, which resulted in the final event of some files being truncated and therefore incomplete. After the data set was cleaned of recondite events, the remainder was enriched using the IP2Location database, which provided the Internet Service Provider (ISP) and geolocation data associated with the source IP address. Further enrichment of the data set using the Collective Intelligence Framework (CIF) and Shodan were explored, and are discussed in later sections.

III. Strategic Cyber Threat Indicators

A. Internet Neighborhood Analysis

Previous research reported that blacklisted IP addresses tend to cluster into Internet neighborhoods [7,8], or Internet service providers. Two metrics were suggested to measure the uncleanliness of an Internet neighborhood. Spatial uncleanliness refers to the tendency for compromised hosts to cluster in unclean networks. Temporal uncleanliness refers to the tendency for compromised hosts to repeatedly appear in unclean networks. The suggestion was that tracking and blocking traffic from unclean Internet neighborhoods may be a viable strategy to improve network security.

We first analyzed the IDS log by grouping number of reported events from each IP address. We used IP2Location to map IP addresses to their registered owners. We found that the top 10 IP addresses accounted for 30% of the events, the top 28 IPs accounted for 45% of events, and top 40 IPs accounted for 50% of events. We further broke down the top forty IP addresses into the following categories:

- General threat: 31% of alerts came from hosts that were down, web servers, or game servers.
- Researchers: 17% of alerts came from known reputable security research organizations. These include Shadow Server, the Open Resolver Project, and German universities.
- Hosting providers: 2% of total events came from web hosting providers.

The top 7 countries of origin for detected events, excluding those identified as research traffic, are listed in Table 2.

TABLE II. Countries of orignin for top 7 attack IPs

Rank	Country	Percentage of events
1	China	43%
2	United States	24%
3	Russia	17%
4	Germany	7%
5	Netherlands	4%
6	Romania	4%
7	Israel	2%

We next grouped IP addresses into their assigned owners, referred to as the Internet Service Provider or ISP for the rest of the paper. Of the 131,361 ISPs in our data set, we noted approximately 66.31%, are allocated less than a full class C block. We specifically looked at two metrics in detail: total number of IPs involved in alerts and percentage of IP addresses involved for a given ISP.

We found 207 ISPs where 100% of its IP addresses triggered at least one alert. However, the average events per IP address was only 1.04. The vast majority of ISPs, 99.84%, have less than 1% of their allotted IP addresses involved in events. Table 3 lists the top 10 ISPs based on the percentage of IP addresses that triggered events. We also list the average number of events triggered by each IP address.

TABLE III. Top 10 ISPs based on percentage of IP addresses

Rank	ISP	Average alerts/IP address	Percent IP address
1	Comcast Cable Inc.	2.46	0.058
2	Verizon Wireless	3.11	0.047
3	Comcast Cable Holding	3.32	0.046
4	AT&T Internet Service	2.28	0.037
5	Cox Communications	2.69	0.036
6	Charter Communications	2.33	0.030
7	Verizon Online	3.66	0.023
8	Qwest Communications	3.29	0.023
9	Time Warner Cable	2.78	0.011
10	Amazon Technologies	10.15	<0.001

As one of the primary objectives of this level of analysis is to help cyber security engineers make risk-based decisions, an attempt is made to establish a projected measure of maliciousness for a given Internet neighborhood. Using ISPs as the basis, we calculated the average of events that would be triggered by an arbitrary IP address within its assigned range. Table 4 lists the top 10 ISPs with the highest average events

per allotted IP addresses, these were ISPs where almost all of its IPs are reported in alerts.

TABLE IV. TOP 10 ISPS WITH HIGHEST AVERAGE EVENTS PER ALLOCATED IP ADDRESSES

Rank	ISP	Average events/IP	Total IPs
1	Business Static Service	6.86	14
2	Bidwell Title & Escrow CHI	4.50	2
3	Widomaker	4.50	2
4	Direct Bassett	2.22	5
5	Sila Heating & AC	2.22	5
6	SCI TEC INC	2.20	2
7	BRSTL BO Police Hall	2.20	5
8	PRU Homesale YWGC	1.67	3
9	Barry Swenson Builder	1.50	4
10	Karmart Auto Nissan	1.20	5

A very small number of IP addresses generate a large percentage of all events. The top 10 ISPs in terms of event count contributed nearly half (49.89%) of all events, and originated from 1433 distinct IP addresses, approximately 0.05% of those recorded. This suggests those IP addresses could be easily blacklisted and blocked at the network perimeter, rather than blocking traffic originating from anywhere within the ISP. However, attackers can bypass blacklisting with ease by changing hosts periodically, and may use public blacklists to identify those to abandon. A relatively small percentage of all recorded IP addresses, 2.98%, triggered only a single event, but they account for 26.35% of all recorded ISPs. The vast majority of IP addresses reported are involved in only a few attacks and may otherwise engage in legitimate business. Blocking by Internet neighborhood must depend on the type of neighborhood. It may be difficult to block large Internet service providers serving a significant percentage of one's customer base. However, it is feasible to identify hosting providers with a history of hosting compromised machines. ISP providing cellular services may present challenges, as attacks launched from mobile devices over a cellular network could be forwarded from any in a large pool of the carrier's allotted IP addresses. We noticed that prominent cellular ISP providers are among ISPs hosting attack IPs.

B. Attack type distribution

We performed analysis on types of attacks detected by the IDS for any patterns of interest. A majority of the alerts were either reconnaissance scans or potentially denial of service attacks, e.g. DNS version requests and half-open TCP packets. Given that large amount of attack traffic originate from land based ISP, cloud computing platforms, and cellular ISP, we looked at attacks from these source in more detail and revealed some interesting patterns. The number in parentheses besides each ISP represents its rank in terms of the number of events which originated from it.

For cellular carriers, we focused on the dominant attack types originating from the ISP with the highest event count controlled by three major US carriers: Verizon Wireless (51), AT&T Mobility (62), and T-Mobile USA Inc. (93). Combined, they account for approximately .56% of all events. Only seven event types constitute greater than 99% of all events originating from these ISPs, some of which originated from only one of these ISPs, as shown in Table 5.

TABLE V. DOMINANT ATTACK TYPES FOR THREE MAJOR US CELULAR ISPS

Event Type	Percent	ISPs
TCP: Invalid TCP Traffic: Possible Recon Scan (FIN no ACK)	53.7	All
HTTP: Server 500 Error	20.3	All
Invalid TCP Traffic: Possible Recon Scan (SYN FIN)	10.0	T-Mobile
SSL: Handshake Failure	7.3	All
HTTP: Curl Web Page Retrieval Attempt	5.3	AT&T
IP: Fragment Bad MF Bits	2.0	Verizon
UDP: Length Invalid	1.3	Verizon

For attacks originating from land-based ISPs, we selected five under US-based providers with the highest contribution to events: Time-Warner Inc. (27), AT&T Internet Services Inc. (33), AT&T Services Inc. (173), Comcast Cable Communications Inc. (38), and Comcast Cable Communication Holdings Inc. (45). Combined, they contributed approximately 1.89% of all events, over 99.3% of which were of six different event types, as shown in Table 6. Notably, with the exception of SIPVicious Brute Force SIP Tool, all event types in this list also appear in the list for cellular carriers, but with different probabilities. In particular, HTTP Server 500 Errors were roughly four times more likely to originate from one of these land-base ISPs than from one of the cellular ones.

TABLE VI. DOMINANT ATTACK TYPES FOR THREE MAJOR US LAND-BASED ISPS

Event Type	Percent	ISPs
HTTP: Server 500 Error	79.39	All
SIP: SipVicious Brute Force SIP Tool	9.00	All
HTTP: Curl Web Page Retrieval Attempt	5.81	All
SSL: Handshake Failure	4.10	All
TCP: Invalid TCP Traffic: Possible Recon Scan (FIN no ACK)	0.62	AT&T, Time-Warner
HTTP: Curl Web Page Retrieval Attempt	0.41	Time-Warner

The three cloud platform ISPs we chose to examine exhibited very different behavior than either the cellular or land-based ISPs. We examined event types from Amazon Inc. (28), Amazon Technologies Inc. (42), and Google Inc. (35), which combined contributed approximately 1.37% of all events. Of these, 99.75% were of one of seven types, as shown in Table 7. Interestingly, only event type to also appear in either the cellular or cloud service providers' lists is HTTP Curl Web Page Retrieval Attempt, though it appears here with a significantly higher probability.

TABLE VII. DOMINANT ATTACK TYPES FOR THREE MAJOR US CLOUD COMPUTING PROVIDERS

Event Type	Percent	ISP
HTTP: Curl Web Page Retrieval Attempt	62.13	Both
HTTP: Windows Executable Download	30.24	Google
HTTP: Apache HTTP Server-X-Forwarded-For-Denial-Of-Service	3.32	Both
HTTP: Soft Hyphen Obfuscated URL	1.73	Both
SIP: SipVicious Brute Force SIP Tool	0.86	Amazon
TCP: Header Length Invalid e.g. Fragroute	0.76	Amazon
TCP: Header Incomplete	0.70	Amazon

We found the variation in the top event types sourced from ISPs of the three identified categories to be surprising. It suggests that attackers are tailoring their attacks to the type of resources each tends to have available. For instance, mobile devices have limited computing power, and are therefore not well suited to launching DDoS attacks. However, they have high utility when used to launch distributed reconnaissance scans, such as fingerprinting target servers using invalid TCP packets.

C. Coordinated attacks

Cyber attacks can be divided into two main categories: unintelligent and intelligent. The former are frequently executed by malicious programs or inexperienced attackers, both of which launch exploits simplistically against targets that are chosen opportunistically and at random. In those instances, little valuable cyber threat intelligence can be derived from the attacker. The latter, however, are executed by more skilled attackers who employ strategies designed to overcome the matured cyber defenses of specific targets. One such strategy, intended both to circumvent threat detection and to overwhelm certain defense mechanisms, such as IDS and IPS, is to launch attacks from hundreds or thousands of hosts in a coordinated fashion. As this requires an attacker (which may be a group of individuals) to marshal a large number of computing resources, it is reasonable to assume at least some of those resource may be used in multiple attacks over time - against single or multiple targets. Therefore, detecting, identifying, and characterizing the hosts participating in coordinated attacks can help security engineers to prioritize their efforts and contribute to valuable strategic cyber threat intelligence.

We explored two overall strategies for coordinated attack detection: a statistics-based, bottom-up approach, and an exploratory visualization-based approach, each of which contributed to a case study.

Statistics-based approach

Our statistics-based approach begins with the simple assertion: that two hosts may be engaged in a coordinated attack if they each trigger the same exploit at approximately the same time. Relative to our data set, this meant finding the sets of all IP addresses identified as the source of any event of a given type within the same recorded minute. We explored aggregations of smaller and larger duration, and chose to settle on one minute; that option appeared to provide adequate granularity while also maintaining manageable storage and computational requirements. We acknowledge the potential for members of the set of hosts or any given event type and minute to be characterized by coincidence; uncoordinated attackers happened to trigger the same event type in that same time. Therefore, we used this aggregated data set as a starting point for discriminating more compelling indicators of coordination, of which we explored two general views: temporal and behavioral.

Under the temporal view, hosts controlled by the same attacker may exhibit similar attack behavior over time, particularly if centrally controlled, such as in a botnet. Thus, we looked for similarities in the first observed, last observed, and event counts for hosts executing the same event type. This analysis resulted in discovery of a large number of incidents where a given host triggered a given event type only during a simple recorded minute in the entire data set, which we coined a one-off. These are noteworthy because this behavior would be atypical of unintelligent attackers such as hosts infected by self-propagating malware or those controlled by spammers or script-kiddies. Further analysis of such events revealed that often, a large number of one-offs were detected for the same event type in a short period of time, which would not likely result from software errors. This activity could be explained if those hosts only executed the exploit because they were carefully controlled by a botnet or similar mechanism. Findings of this type contributed to our Gnu Bash Remote Code Execution case study described below. A second interesting discovery were a large number of one-offs for certain event types that curiously occurred at similar minutes within the hour, but at different periods of the day. This suggests these attacks were programmed to occur at regular intervals and may have been executed by multiple instances of the same software. This analysis contributed to the Wget Web Page Retrieval Attempt case study described below.

Under the behavioral view, we consider that successful compromise of a system may involve a variety of exploits

being executed in combination. Therefore, we identified groups of event types that were each triggered by the same IP address during the same minute, and then further identified the sets of IP addresses that triggered the same combinations of events at the same times. An additional benefit of this approach was identifying the correlation between certain attack types, which may allow cyber threat analysts to create tuned IDS/IPS rules that more readily detect coordinated hosts that are individually contributing a minimal combination of complementary exploits, which would otherwise go unnoticed amid the constant 'background noise' of events.

Machine learning /visualization-based approach

This approach was holistic and exploratory in nature. Our initial step was generating histograms by event count over time for individual event types to identify interesting features in the data, such as spikes or repeating patterns. As a result, we hoped to recognize related and possibly coordinated attack behaviors that could not be directly correlated using our statistics-based approach. We focused on two scenarios: groups of hosts triggering a given attack in a similar manner but at different times, and groups of hosts triggering different attacks in a similar manner and at the same time. The first scenario would be evidence of one or more attackers using the same tools, such as shared attack scripts. This information assists a cyber threat analyst to prioritize research efforts and tailor defensive strategies to the tools being used.

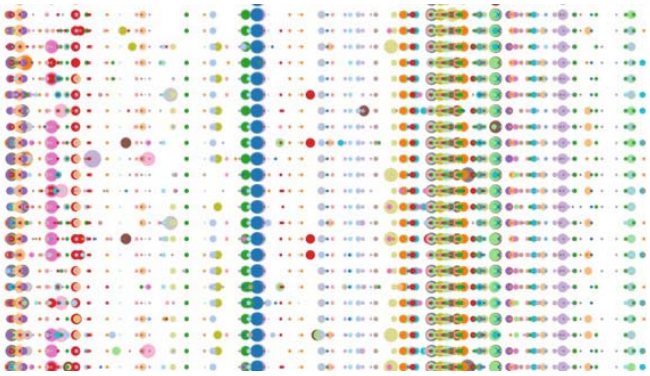
Each alert from the IDS log is treated as a tuple (IP, Attack type, time). We divide time into 5-minute intervals. All events with the same 5-minute window share the same time value in the tuple. We then applied an unsupervised learning method [10] to cluster these tuples. Our goal is the find groups of IPs performing the same attack during the same 5-minute window. Such groups of IPs may represent coordinated attacks from a botnet. Figure 1 shows a typical attack pattern where each line represents activity of an IP address over time. Columns in the figure represent different time windows and the color/shape of the dot represent different attack types.

In vast majority of cases, attacks are not synchronized, as illustrated by the random distribution of dots in Figure 1. This method is promising to quickly discover clusters of attacks that may represent coordinated (synchronized) attacks. Figure 2 shows a cluster we found that is likely the result of small botnet where all members of this group is performing the same action at approximately the same time.

This example could reveal more complicated strategies in which coordinated attackers can use an effective combination of exploits while avoiding detection by keeping network traffic volume per exploit and per host below common alert thresholds for DDOS. Motivations for this example include a better awareness of the skill of your threat actors and fine-tuning automated detection and response capabilities.

Fig. 1. Typical attack distribution over time

Fig. 2. A possible coordinated attack distribution over time



One must be mindful that with any analysis involving large amounts of data, a pattern like shown in Figure 2 could be due to statistical coincident. Other plausible security explanations must be found to validate these findings. Analysis following this example was responsible for initial discovery of events detailed in the Gnu Bash Remote Code Execution case study. It is a good example of using the exploratory approach, the statistical approach, and security analysis to identified coordinated attacks.

Gnu Bash Remote code execution case study

Gnu bash remote code execution vulnerability was a first discovered in Sept 2014 and often referred to as “Shellshock”. TippingPoint was informed on Sept. 25 and many systems were not patched until Oct. 2014. The IDS log we analyzed showed high volumes of shellshock attacks on Sept 26 for about 1.5 hours. After this day no shellshock attacks were reported in the IDS log. This attack came from 4,867 IP addresses. These addresses belonged to 1,562 registered ISPs. The largest block, 549 addresses, belonged to Akamai Technologies Inc. The most reasonable explanation for such an attack is a coordinated botnet.

Wget Webpage retrieval attempt case study

Wget and Curl are often used maliciously to download files. We discovered a large number of both Wget and Curl attempts

throughout the IDS log. We found 2,531 distinct IP addresses issued Wget at four different time windows: 0:46, 0:47, 1:46, 1:47. Alerts for each address range from 1 to 303 with an average of 33 events per IP address. Due to these Wget instances appearing in coordination consistently at these specific time windows throughout the IDS log data, we may reasonably attribute this attack to a botnet that were programmed to recur at these times.

Other coordinated reconnaissance events

We found 15 IP addresses from Hurricane Electric Inc. issued "NTP Monitor List Request Command" together for 2,048 minutes. We found 3 IP addresses from Prolexic Technologies Inc. issued "IP: Length Invalid e.g. Whisker" together for 15,057 minutes. Some other coordinated attacks of interest include "SSL: Handshake Failure" (event count: 414,109), and "SSH Login Attempt on Non Standard Ports" (event count: 37,508).

A common characteristic of all the examples we discussed above is that they have low traffic volume. So a botnet detection mechanism on an ISP network that is looking for DDOS attackers may easily miss them.

IV. Comparison with Related Works

Our analysis of spatial and temporal distributions of attacks from internet neighborhoods suggest blocking ISP with lots of attack IPs as suggested by [8] is not a viable option as it can deny access of legitimate users. However, tracking internet neighborhoods where concentrated attacks are coming from can be a good strategy to foster constructive dialogs between enterprises that are recipients of large cyber-attacks and internet providers to collaboratively reduce attack traffic.

Previous work on detecting coordinated attacks is mainly associated botnet detection. Their focus is on detecting computers that may be part of a botnet on a given network [11], such as an ISP or a large enterprise. Such research has looked for coordinated attacks based data collected from the network where attacks originate. We explored the approach to look for coordinated attacks based on incoming attack traffic to a given network. Future research is needed to combine the two approaches where coordinated attacks detected using techniques discussed in this paper can be shared with ISPs where the attacks originated to find botnets.

There is a significant body of research on using log analysis to improve security, such as detecting web application flaws [1,2], DOS attacks [3], and APT activities [4]. More recently researchers have started to use big data analysis techniques to detect early-state infections, e.g. [5,6]. We are not aware of any previous work to detect coordinated attacks based on IDS logs.

We attempted correlate our findings on attack IP addresses to publically available information sources in the same time window. We took a random sample of 10,000 IP addresses and queried Shodan. We found very few matches. We looked up a sample of IP addresses in the IDS log on other public blacklisted IPs (CyberDNA Internet Watch , Collective Intelligence Framework and Spamhause) to see if others have reported attacks originating from these IP addresses during the same time frame. Again we found no matches. This suggests publicly available threat intelligence has limitations and we need to encourage more threat information sharing by private companies.

V. Conclusion

State-of-the-art threat intelligence sharing could be enhanced by sharing more strategic threat information. Through a case study of analyzing an IDS log from a private company, we demonstrated how this could yield useful strategic threat information without disclosing sensitive information. We identified the following strategic threat information: internet neighborhood, attack type distribution, and attack coordination.

A common challenge for finding patterns in large data sets is how to estimate the statistical significance of the finding. We were able to validate some of the findings based on additional security and application contexts.

Acknowledgment

We acknowledge the contributions of Andrew Mannikus in comparing our results with CIF.

References

[1] Detecting Attacks on Web Applications from Log files, 2008, SANS Institute Reading Room site, https://www.sans.org/reading-room/whitepapers/logging/detecting-attacks-web-applications-log-files-2074.

[2] P. Dange, and D. Shah "Web Log Analysis for Security Compliance Using Big Data" *International Jouranal of Advanced Research in Computer Science and Software Engineering*, 5(3), March 2015.

[3] G. Munz, and G. Carle, "Real-time analysis of Flow Data for Netowrk Attack Detection" in *Proceedings of the 10th IFIP/IEEE International Integrated Network Management*, pp. 100-108, 2007.

[4] N. Virvilis, O. Serrano, and L. Dandurand, "Big Data Analytics for Sophisticated Attack Detection", *ISACA Journal*, Volume 3, 2014

[5] B. Nafziger, Practical Attack Detection, Analsysi, and Response Using Big Data, Semantics, and Kill Chians within the OODA Loop, 2015. SANS Institute Reading Room Site, https://www.sans.org/reading-room/whitepapers/warfare/practical-attack-detection-analysis-response-big-data-semantics-kill-chainswithi-35990.

[6] A. Oprea, Z. Li, T. Yen, S. Chin, and Sumayah Alrwais, Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data, 2014, http://arxiv.org/abs/1411.5005.

[7] M. G. Cesar, *Internet Bad Neighborhoods.* No. 12. Giovane Cesar Moreira Moura, 2013.

[8] M. Collins, T.J. Shimeall, S. Faber, J. Janies, R. Weaver, and M. De Shon, "Predicting future botnet addresses with uncleanliness" in *Proceedings of the IMC. CERT network Situational Awareness Group*, 2007.

[9] MITRE. STIX-Structured Threat Information Expression. Available: stix.mitre.org/. [Accessed: Jan-24-2016].

[10] A. McCallum, "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu 2002.

[11] G. Gu, R. Perdisci, J. Zhang, and W. Lee. " BotMiner: Clusting Analysis of Network Traffic for Protocol- and Structure- Independent Botnet Detection" USENIX Security Symposium 2008.

This work is sponsored in part by a grant from the National Science Foundation 1129190.