Identifying Mobile Malware and Key Threat Actors in Online Hacker Forums for Proactive Cyber Threat Intelligence

By

John Grisham

A Master's Paper Submitted to the Faculty of the

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS

ELLER COLLEGE OF MANAGEMENT

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2017

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona.

Brief quotations from this thesis are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part must be obtained from the author.

SIGNED: John Grisham

APPROVAL BY MASTERS PAPER ADVISOR

This paper has been approved on the date shown below:

_____ _____

Dr. Mark Patton Date

Lecturer of Management Information Systems

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. ABSTRACT

Cyber-attacks are constantly increasing and can prove difficult to mitigate, even with proper cybersecurity controls. Currently, cyber threat intelligence (CTI) efforts focus on internal threat feeds such as antivirus and system logs. While this approach is valuable, it is reactive in nature as it relies on activity which has already occurred. CTI experts have argued that an actionable CTI program should also provide external, open information relevant to the organization. By finding information about malicious hackers prior to an attack, organizations can provide enhanced CTI and better protect their infrastructure. Hacker forums can provide a rich data source in this regard. This research aims to proactively identify mobile malware and associated key authors. Specifically, the usage of a state-of-the-art neural network architecture, recurrent neural networks, to identify mobile malware attachments followed by social network analysis techniques to determine key hackers disseminating the mobile malware. Results of this study indicate that many identified attachments are zipped Android apps made by threat actors holding administrative positions in hacker forums. The identified mobile malware attachments are consistent with some of the emerging mobile malware concerns as highlighted by industry leaders.

2. INTRODUCTION

With an ever-evolving cyber threat environment, organizations need to take a more proactive approach to cybersecurity (EY, 2014). The average total cost of a data breach to a company is $4 million with 48% of breaches occurring due to hackers and criminal insiders (Ponemon Institute, 2016). Hackers are constantly inventing tools to obtain confidential information and are becoming better at identifying gaps and vulnerabilities in an organization's security (EY, 2014).

One avenue hackers can use to attack organizations is through mobile malware. For example, using a company's "Bring-Your-Own-Device" policy against them to attack employee phones outside of the company network and using them as a foothold to enter the internal company network. Also, due to valuable personally identifiable information (PII) being stored on mobile devices with services such as Android Pay becoming commonplace, mobile devices are increasingly targeted by criminals. With 1.4 billion smartphones in 2015, and five out of six phones running Android, hackers have a large attack surface to work with (Symantec, 2016). In 2015, 430 million new pieces of malware were found (Symantec, 2016). To combat the mobile malware risk, organizations need to look at ways to mitigate attacks to their employees' phones, both work and personal, and internal network infrastructure.

In order to prevent potential cyber-attacks, organizations rely on cyber threat intelligence (CTI) to provide insight on the malware and threat landscape they face. However, current CTI processes reactively look at malware in cyberspace, focusing on studying attacks after they have occurred. According to Bromiley, "CTI cannot be dated information that fails to help an organization protect itself or better understand their threats, such as their attackers and their related techniques" (Bromiley, 2016). To be more proactive, it is essential to utilize an approach combining both external sources of intelligence, which can help identify previously unknown threats, and internal knowledge, which identifies current threats based on the organizations critical assets. This more "holistic" approach is shown in Figure 1, which combines both sources and shows how they complement one another. Used in tandem with internal intelligence, external sources can provide valuable information, such as attackers' Tactics, Techniques, and Procedures (TTPs).

Figure 1 - A Combined Threat Intelligence Approach (modified from Bromiley, 2016)

Malicious threat actors will often use online hacker forums to share their TTPs used to compromise systems (Abbasi, Li, Benjamin, Hu, & Chen, 2014; Benjamin & Chen, 2012; Samtani & Chen, 2016). For example, hackers will often share mobile malware variants in the form of forum attachments such as in Figure 2. Identifying and studying such posts and the threat actors who make the posts can contribute to a novel and proactive form of CTI. As such, hacker forums can provide a rich data source of malware and threat actors.

Figure 2 - Mobile Malware Attachment in the Arabic Forum, Ashiyane

Overall, there are hundreds of hacker forums, with millions of posts, tens of thousands of members, and tens of thousands of malicious tools. Given the presence of attack vectors and threat actors in forums, this research aims to: develop proactive CTI by collecting large, international hacker forums containing these attack vectors; use deep learning text classification to identify emerging mobile malware trends; and leverage social network analysis (SNA) to identify key threat actors disseminating these assets.

The remainder of this paper is organized as follows: First, literature related to mobile malware, CTI, hacker forums, recurrent neural networks, and social network analysis is reviewed. Second, the collection process and resulting research testbed is detailed. Subsequently, the key findings and results are summarized. Finally, several promising directions for future work are highlighted and a conclusion to this research is provided.

3. LITERATURE REVIEW

For the purposes of this research, five areas of literature were reviewed and purposes are detailed below:

- **Mobile malware**: identify trends and evolution of relevant malware threats
- **CTI**: identify current data sources and approaches to create effective CTI
- **Hacker communities**: insight to the types of threats, information sharing, and threat actors on such communities
- **Recurrent Neural Networks (RNN's)**: better understand state-of-the-art techniques to classify and identify relevant text
- **Social Network Analysis (SNA)**: identify methods to determine hacker social structures and key threat actors.

This review gives a comprehensive picture of the types of studies and industry standards necessary for relevant and proactive CTI.

3.1 Mobile Malware Research

In general, mobile malware is built for Android devices. Also, the open-source nature of the Android operating system makes it easier to create malware. Because Android devices are very popular, they make for a lucrative target for hackers to distribute malware. To propagate mobile malware, code obfuscation and drive-by downloads are the most common trends to install malware on unsuspecting phone users (Symantec, 2016; Zhou & Jiang, 2012). In addition, of available in-the-wild mobile malware, 86% are repackaged versions of legitimate applications with malicious payloads (Zhou & Jiang, 2012). These "new" applications can be found on third-party Android app markets, but also on the official Google Play Store itself (Vidas, Votipka, & Christin, 2011). This is in part due to Google having less stringent app screening processes (Symantec, 2016). This means that many mobile malwares go undetected for long periods of time since users believe they are downloading a legitimate and useful app that would otherwise infect their phone without their knowledge. Once successfully downloaded, malware will use privilege escalation attacks to exploit the Android Operating System (OS) and gain root access to the device (Zhou & Jiang, 2012).

Because mobile malware can easily acquire sensitive information without user knowledge once installed, it is imperative to devise ways to detect and stop it from running. To determine if an app is malware prior to install, researchers have measured the malware behavior potential of a particular app's permissions at install (Chakradeo, Reaves, & Enck, 2013). However, finding a malware once installed can be difficult. To find undiscovered malware and determine its capabilities, static or dynamic analysis is required. Prior studies have seen malware remaining hidden in emulated lab environments (Dilger, 2014) and even hidden in volatile memory (Tung, 2014). As such, understanding malware characteristics and behavior prior to an infection or attack is a key component in providing proactive CTI.

3.2 Cyber Threat Intelligence

CTI is defined as threat intelligence related to computers, networks, and IT (Farnham, 2013). Proper implementation of CTI can provide a valuable tool for an organization to understand their threat landscape

(Bromiley, 2016). By utilizing CTI, organizations can see improvements in their detection of and response to internal attacks (Shackleford, 2015).

Traditionally, CTI focuses on identifying threats and threat actors using a combination of internal feeds such as Intrusion Detection Systems (IDS), antivirus, and system logs. In order to sift through the large amounts of data to find and stop threats, organizations can implement a combination of human and machine intelligence, such as through a Security Information and Event Management (SIEM) (Shackleford, 2015). This type of machine automation where the SIEM presents potential incidents in an easily human-digestible format, such as dashboards, provides an essential view of a company's infrastructure and the internal threat landscape that they face. For example, an organization will typically devise approaches to enhance their mobile malware cyber-defenses after they have seen incidents of mobile malware identified in their networks by their SIEM or IDS. While this approach is valuable and can help mitigate attacks as they happen, it is reactive in nature as it relies on activity which has already occurred.

CTI experts have argued that an actionable CTI program should involve not only traditional, internal approaches, but external, open information relevant to the organization (Bromiley, 2016). This type of information provides organizations a view into their external threat landscape to find threats they may have been previously unware of. Also, it gives context of attacks against the organization, which can shorten times from detection to remediation (Bromiley, 2016). There are many external feeds that can be utilized in this fashion, such as Twitter, Facebook, and other types of forum data. However, many of these feeds would be impossible to view all at once and would be a waste of computing resources if it could be automated. So, to provide threats that are relevant to an organization, analysts must combine data with contextual information (i.e., internal incidents with external knowledge) (Bromiley, 2016). This helps CTI be more proactive by finding threats before they occur, helping to understand attackers, and identifying hacker TTP's (Bromiley, 2016). One potentially rich, external data source that can offer significant value in developing proactive CTI is online hacker forums, a grounds for hackers to share new TTPs and targets of interest.

3.3 Hacker Forum Research

Given the large amounts of hacker forum literature available and their differing research focus, the literature review presented in this section references selected works in Appendix A, a hacker forum literature table related to research focusing on threats, key hacker identification, and usage of forums for CTI.

In order to readily and effectively share with peers, hackers will utilize various communication and information sharing mediums such as Internet-Relay-Chat (IRC), carding shops, DarkNet Marketplaces, and hacker forums (Benjamin, Li, Holt, & Chen, 2015). These communication channels are typically found in a snowball-styled approach where researchers will use search engines to search hacking-related keywords (e.g., "hacker", "malware", and "forum") (Chu, Holt, & Ahn, 2010; T. J. Holt, 2012; Li & Chen, 2014; Samtani, Chinn, & Chen, 2015). Once channels are identified, the researcher can use them as pivot points to find additional links to other relevant channels or even different communication mediums. Typically, only a small subset of channels are required to test research questions. As such, researchers will use testbeds of channels containing similar topics of interest (T. J. Holt, 2012; Li & Chen, 2014).

Among these channels, forums offer hackers the ability to freely share malicious tools with each other through forum attachments (Samtani, Chinn, Larson, & Chen, 2016; Samtani et al., 2015). Much of the research in hacker forums focuses on what hacker tools are available. Other studies focus on understanding the characteristics of key hackers and how they interact amongst one another (Samtani et al., 2016, 2015). Such studies have discovered that key hackers are major contributors within their community (e.g., forum moderators or senior members) (Benjamin & Chen, 2012; Samtani & Chen, 2016). Other literature has found that hackers cluster into groups (Abbasi et al., 2014) with hackers typically belonging to multiple hacking groups (Thomas J Holt, Strumsky, Smirnova, & Kilger, 2012).

The remainder of hacker forum research highlighted in Appendix A utilizes the forums to determine emerging threats and related intelligence. By using a combination of information retrieval and machine learning, past scholars have extracted actionable, proactive CTI from hacker forums (Benjamin et al., 2015; Nunes et al., 2016; Samtani & Chen, 2016). These past works have focused on identifying and classifying hacker malware toolsets and

emerging threats based on features such as the name of the attachment, associated post text, sub forum name, and thread title. To better understand the textual nature of the forums, state-of-the-art text classification algorithms such as Recurrent Neural Networks can enable efficient and effective processing of hacker forum data to generate valuable CTI.

3.4 Recurrent Neural Networks

A Neural Network is a learning algorithm that performs computational problem solving using neural nodes. It is useful in computationally difficult classification tasks (e.g., speech recognition and photo recognition). For these complex functions where generalization is required beyond immediate training sets, neural networks are better than traditional architecture such as Support Vector Machines (Bengio & Lecun, 2007). Specialized neural network architectures, such as a RNN, can be utilized in text mining classification by representing words as vectors (i.e., word embedding). A RNN connects prior words and allow for variable lengths in word embedding sequences. In order for the RNN to properly learn sentence structure and the meaning of words, it produces probabilities to find which words follow each other. In addition, a way for the network to learn over time is through backpropagation, which allows the network to learn through error comparisons in prior iterations. This error handling is common across all types of neural networks. However, one issue with RNNs is that they cannot handle error well over time, leading to improper learning (Hochreiter & Schmidhuber, 1997). To solve this, Long Short-Term Memory (LSTM) RNN architectures are commonly utilized.

LSTM's add nodes that enforce constant error (Hochreiter & Schmidhuber, 1997). LSTM was further improved with a "forget" node, resetting the network state to prevent indefinite networks (Gers, Schmidhuber, & Cummins, 2000). LSTMs have been successful in many different applications such as machine translation (Sutskever, Vinyals, & Le, 2014) and parsing (Dyer, Ballesteros, Ling, Matthews, & Smith, 2015; Vinyals et al., 2015). Despite the value RNN's and LSTM's can play in identifying malicious content in hacker forums by generalizing sentence meaning, they cannot provide the mechanisms to identify who the key threat actors are for these malicious tools. However, social network analysis can provide significant value in such a task.

3.5 Social Network Analysis

Current work in understanding hacker community relationships utilize some form of social network analysis (SNA) (Thomas J Holt et al., 2012; Motoyama, McCoy, Levchenko, Savage, & Voelker, 2011; Samtani & Chen, 2016). Social networks consist of nodes, or actors, that interact with other nodes through edges, or relationships. Another type of social network, a two-mode social network comprises two separate types of nodes, typically actor nodes affiliated with event nodes (Faust, 1997). In a forum context, two-mode networks can be converted to one-mode with actors tied to each other through posts in a shared thread (Samtani & Chen, 2016; Stewart & Abidi, 2012; Zhang et al., 2009).

Hacker and dark network forums typically convert to one-mode networks to understand threat actors' social groups and capabilities (Lu, Luo, Polgar, & Cao, 2010; Samtani & Chen, 2016; Zhang et al., 2009). Converting to a one-mode network also allows researchers to calculate all of the potential centrality measures (e.g., degree, betweenness, closeness, and eigenvector) for dark (Thomas J Holt et al., 2012; Isah, Neagu, & Trundle, 2015) and hacker networks (Lu et al., 2010; Sarvari, Abozinadah, Mbaziira, & Mccoy, 2014).

3.6 Research Gaps and Questions

Several research gaps were identified from the literature review. First, current CTI is reactive in nature, as it relies primarily on internal network data with limited contextual analysis. Second, malware in a CTI context is typically found and analyzed after an attack. Consequently, threat actors are identified after an attack. Finally, it is unclear if there exists any study identifying mobile malware in hacker forums, a rich data source that can provide proactive CTI. To help address these gaps, the following research questions are posed for study:

- What are the trends of mobile malware within hacker forums?
- What are the emerging mobile malware threats within hacker forums?
- Who are the key threat actors for mobile malware in hacker forums?

4. RESEARCH TESTBED AND DESIGN

4.1 Testbed Selection

The first stage of the research design (Figure 3) focuses on data collection and pre-processing. During the collection phase of this research, a total of 62 forums were identified. 28 of which were not collected because they either went offline during collection or were unavailable without paying for access. However, 34 total forums were collected and are detailed in Appendix B. Of those collected, there were multiple languages collected: 17 English, 13 Russian, and 4 Arabic forums.

For the purposes of this study, four hacker forums are selected: 1 English, 1 Russian, and 2 Arabic forums. Table 1 summarizes the collection. These forums were selected for several reasons. First, these forums are known in the hacker community to contain mobile malware. Second, these forums could be accessed without payment or invitation. Finally, these forums represent multiple geo-political regions, thus resulting in a diverse, international dataset. Following forum identification, all web pages were collected through Tor-obfuscated crawlers with relevant information parsed into a relational database. Overall, the collection has 481,922 posts made by 43,272 authors in 46,292 threads with dates ranging from May 2003 to October 2016. There are also have 43,462 attachments.

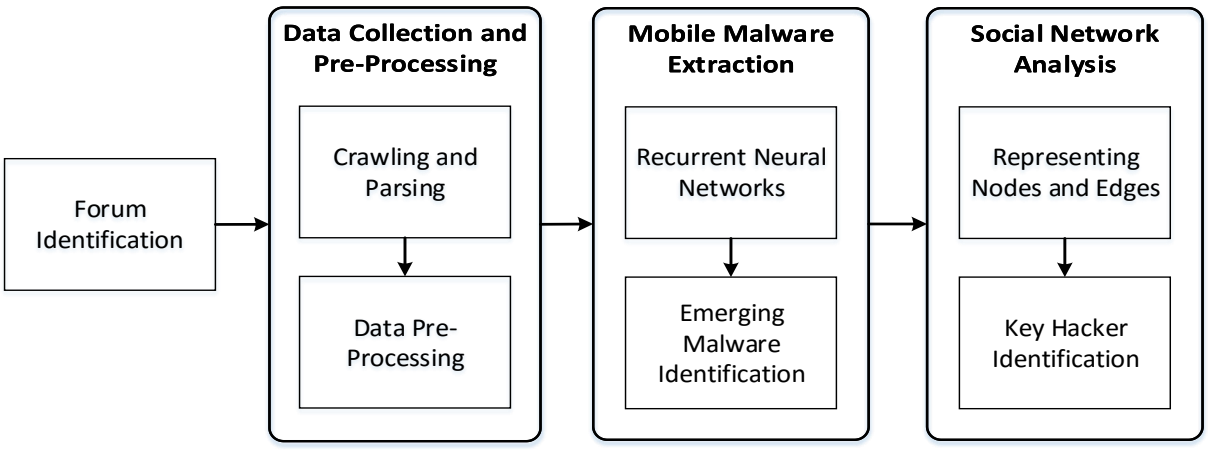

Figure 3 - Research Design

Forum	Language	# of Posts	# of Threads	# of Authors	# Attach.
Ashiyane	Arabic	32,247	8,575	6,406	41,191
Hackhound	English	19,880	2,683	1,832	650
VBSpiders	Arabic	199,978	33,423	33,423	661
Zloy	Russian	229,817	1,611	1,611	960
Total:	**-**	**481,922**	**46,292**	**43,272**	**43,462**

Table 1 - Research Testbed

4.2 Mobile Malware Extraction

The second stage of the framework (Figure 3) focuses on identifying the mobile malware within the collection. To do so, the dataset is limited to only posts with attachments. Attachments signify that the author is actively sharing a piece of malware, compared to having to infer from textual cues when a poster is talking about a particular piece of malware. Doing so also cuts neural network processing time and improves the model's training phase since it does not need to find malware based on context. An LSTM RNN binary classifier using the Python Deep Learning Library, Keras (Chollet, 2015), was trained to determine mobile malware against other types of attachments. Textual forum characteristics such as sub-forum name, thread title, post content, and attachment name were used to classify mobile malware.

To improve the neural network, post content data was split on sentences using a sentence tokenizer. These sentences were then put into separate records that duplicated all other content (i.e., sub-forum, thread title, attachment). This was done because when the LSTM reads input data, it creates a maximum record length. The length is then applied to every record, with shorter records padded with blank inputs. As a result, the LSTM architecture is able to handle variable length sentences. 8,437 total records were used with hold-out validation of 6,000 training inputs and 2,437 test records. The model was benchmarked against another RNN architecture also capable of handling error over time with similar performance to LSTM, Gated Recurrent Unit (GRU). GRU is identical to LSTM except for missing a forget gate. This forget gate decreases processing time and cycles of the processing unit. However, current research is inconclusive as to which RNN architecture performs the best and

depends entirely on the dataset (Chung, Gulcehre, Cho, & Bengio, 2014). As such, both are used on the dataset to determine best fit.

Both architectures were evaluated using standard information retrieval measures of precision, recall, and F-Measure. Following identification, amounts of mobile malware are plotted over time to identify the trends and popularity of mobile malware in forums. Recently shared pieces of mobile malware are given more consideration to provide timely CTI.

4.3 Social Network Analysis

The final component of the framework (Figure 3) uses mobile malware author information to create a social network and identify key threat actors. Observing mobile malware authors in a social network provides insight on key hackers that can be easily acted on (Bromiley, 2016). Authors found in threads that contain mobile malware attachments are gathered for creating these hacker networks in mobile malware. Consistent with prior literature, bipartite networks are constructed connecting actors to mobile malware related threads they post in. Then, the two-mode network is projected into a one-mode network, resulting in hacker co-occurrences with one another (Samtani & Chen, 2016).

A one-mode network allows for ease of calculations for centrality measures, providing information on key threat actors. Prior literature has used one-mode hacker networks to understand centrality and connections between specific actors (Thomas J Holt et al., 2012; Isah et al., 2015). For this research, co-occurrences do not provide a definite interaction amongst hackers, but rather provide information on key members. These key members found with or around trending mobile malware can inform organizations of the mobile malware: threat landscape, threat trends, and threat actors.

5. RESULTS AND DISCUSSIONS

5.1 RNN Evaluation

To determine the effectiveness of the LSTM RNN architecture, it was evaluated against a GRU RNN to find which performed better on the data. A representative training set of the data was used to train the networks. Both were evaluated on precision, recall, and F-Measure as shown in Table 2. Consequently, both architectures had near perfect scores on non-mobile malware attachments (designated as Non-MM) due to the large amount of support for non-mobile malware posts, as represented in the data. For the goal of this research, the LSTM classified mobile malware is the more important value (as designated with asterisks). Overall, the systems achieved similar performances, as seen in prior research (Chung et al., 2014). The model was then applied to the dataset.

	Precision	**Recall**	**F-Measure**
LSTM Non-MM	0.99	1.00	1.00
LSTM MM	**0.95****	**0.81****	**0.87****
LSTM Average	0.99	0.99	0.99
GRU Non-MM	0.99	1.00	1.00
GRU MM	0.87	0.86	0.86
GRU Average	0.99	0.99	0.99

Table 2 - Benchmark Metrics for LSTM and GRU

5.2 Mobile Malware Extraction

For the purposes of this research, the Arabic hacker forum, Ashiyane, was chosen as a case study. Ashiyane is a well-known Arabic forum with the majority of attachments available in the testbed. In addition, it is a long-standing and active forum containing mobile malware related posts and topics from its inception in May 2003 to when it was last crawled in October 2016.

The resulting classified mobile malware from the LSTM model was then plotted over time based on their associated postdate. Generally, sharing new mobile malware attachments amongst hackers allows them to

continuously improve their knowledge, toolsets, and tactics. As such, instances of shared malware in Ashiyane indicate the popularity of mobile malware being used in hacker toolkits over time. Looking at Figure 4, over the life of the forum there is a multitude of mobile malware in certain months. One reason could be due to "mega-threads", where many posts occur in a single thread, resulting in above average mobile malware attachment counts. In addition, these mega-threads see a large amount of zipped Google Play Store apps, consistent with prior literature of repackaged apps as mobile malware (Zhou & Jiang, 2012). Some of the attached apps are popular, but most are otherwise obscure apps in the Play Store. In addition, the shared apps are typically free on the Play Store, but some forum apps are paid Play Store apps, such as the one in Figure 5. This could be a way to share mobile malware amongst the forum users, who can then use it as part of their tools to infect their victims even if the forum user has no knowledge of mobile app creation. Hackers could place the app on a third party app store or repackage the code as a payload in other apps to entice a victim to download it, since it would otherwise cost them $1.99. Those who download and install the app to their phone would then be subject to mobile malware and be a pivot point for a hacker to get into an organization if the victim brought and connected their phone to their work Wi-Fi.

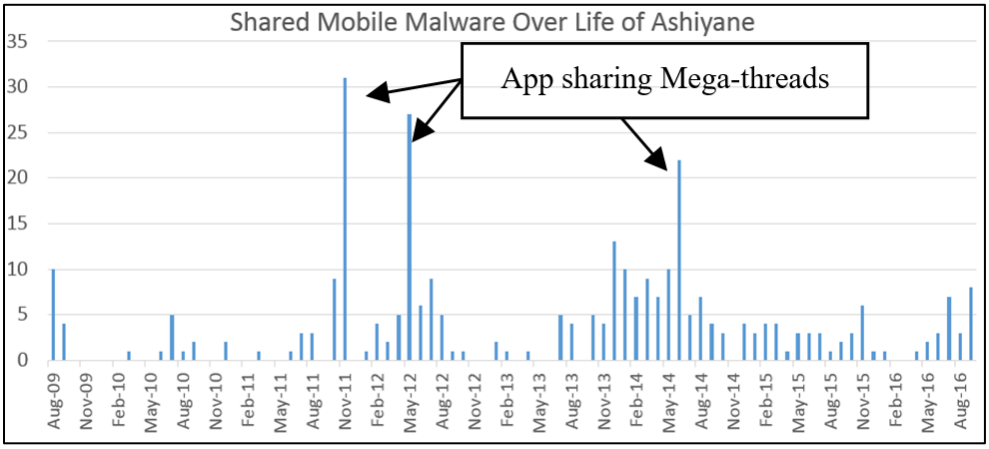

Figure 4 - Mobile Malware Attachments Shared in Ashiyane

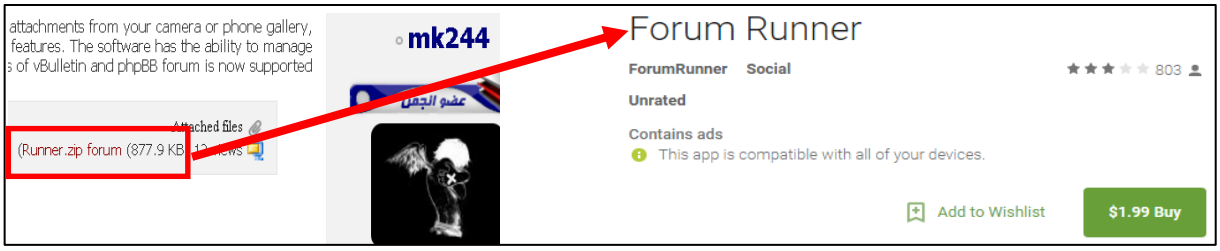

Figure 5 - Attached Hacker Forum App for Sale in Play Store

To provide more proactive CTI, a more recent mobile malware that was shared in September, 2016, is showcased in Figure 6. Again, much of wild mobile malware are repackaged versions of legitimate apps (Zhou & Jiang, 2012). However, in Figure 6, it can be seen that the attachment consists of a custom RAT creator shared on a mobile-related sub-forum that uses a Metasploit module, MSFVenom. This module is capable of using documented exploits built into the Metasploit framework and automatically creating reverse TCP payloads allowing unfettered access to hackers. This particular MSFVenom script is available for many different types of systems such as Windows, Mac, and Linux. It also includes the Android OS. Such an exploit is consistent with some of the emerging mobile malware concerns of PC-like exploit kits for phones as highlighted by Symantec (Symantec, 2016). Also, this piece of malware is documented, updated frequently, and available for immediate download to anyone on GitHub. The poster sharing this malware in the forum could be the original author. However, most likely they are sharing a tool that they utilize and find effective for hacking mobile phones. This also indicates the popularity of the tool amongst hackers on Ashiyane and could be a potential avenue for assessing the threat landscape and relevant hacker TTPs.

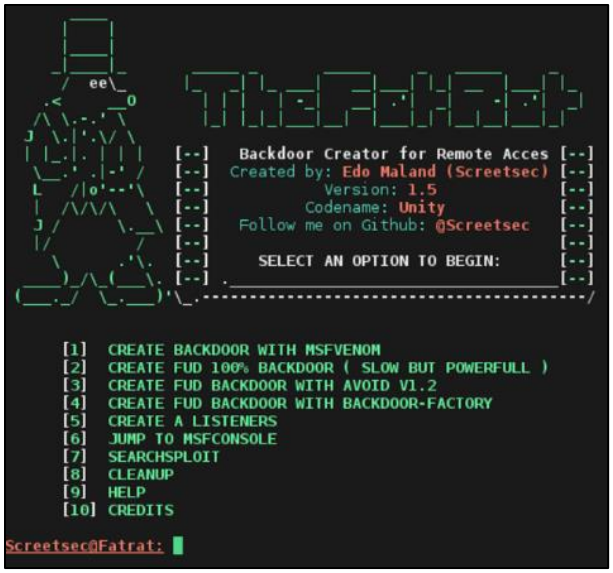

Figure 6 - RAT Creation Script posted in Sep, 2016

5.3 Social Network Analysis

Given the goal of identifying key threat actors, discussion focuses on the mono-partite network rather than a bipartite network modeling hackers and threads. The mono-partite network's topological statistics are summarized in Table 3.

Metric	**Value**
Number of Nodes	100
Number of Edges	562
Network Diameter	5
Connected Components	58
Average Degree	11.24

Table 3 - Mono-partite Network Metrics

Overall, there are 100 nodes and 562 edges in the resultant mono-partite network. A network diameter of 5 indicates a more compact network, meaning hackers have a greater chance of interacting with one another, even if they do not post in the same thread. This is especially true when considering weak ties, which suggest

relationships between groups can occur, even if members do not have a direct tie to one another (Granovetter, 1973). However, a majority of authors have low degree centrality, existing outside of the main network depicted in Figure 7. Typically, these "smaller" hackers prefer to create their own posts and do not engage in shared posts, perhaps in an attempt to gain reputation. Also, they typically do not share more than one or two attachments.

Conversely, the top five authors (summarized in Table 4 and colored in red in Figure 7) have the highest degree and eigenvector centrality compared to the rest of the network. Respectively, these scores indicate higher information dissemination and influence on the network. Further information about the seniority of these key mobile malware authors shown in Table 4 demonstrate that they also hold administrative positions in the forums. This is consistent with prior key hacker identification literature (Benjamin & Chen, 2012; Samtani & Chen, 2016).

Ranking	Author	Degree Centrality	Eigenvector Centrality	Forum Role	Join Date
1	LinX64	58	1.000	Admin	7/26/2013
2	AsAs	48	0.983	Executive	10/26/2012
3	reza20112	46	0.976	Executive	10/23/2013
4	GNU-Linux	46	0.973	Executive	7/15/2012
5	HosseinCactus	44	0.969	Executive	12/3/2011

Table 4 - Top Mobile Malware Authors

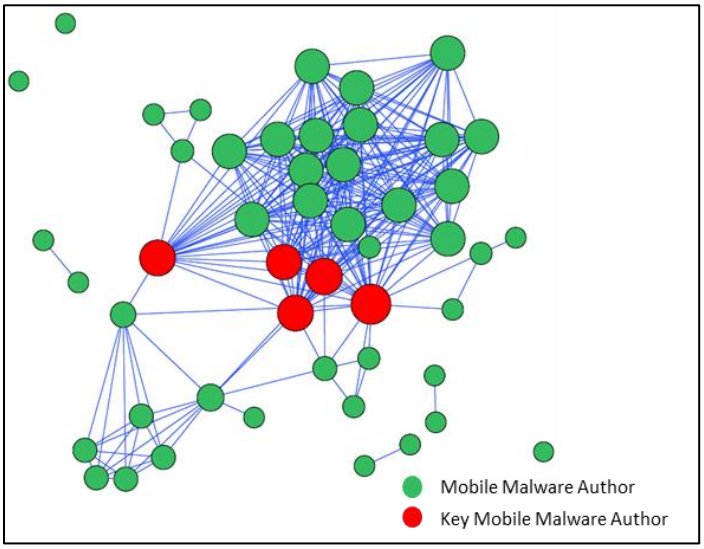

Figure 7 - Mono-partite Social Network

6. CONCLUSION

6.1 Summary

Current industry CTI practices have been reactive to external threats. Organizations are not aware of many threats and only find out about them once their vulnerabilities have been exploited and the damage has been done. Because of this, companies pay millions of dollars to control and fix issues as they arise. Organizations will continue to have devastating cyber-attacks that cripple their business if they are not aware of their current threat landscape or do not participate in proactively finding emerging threats and mitigating against them.

To solve this, many organizations are beginning to push management to spend funds to collect and utilize external CTI for proactive purposes. As referenced in Figure 1, combining external and internal data sources allows an organization to better protect against current and future attacks by creating more effective CTI. One external data source organizations can use are attachments from hacker forums. This hacker forum data can be used to synthesize proactive CTI and better understand the threat landscape.

This research aims to identify mobile malware and key threat actors within hacker forums for external, proactive CTI. Specifically, attachments in hacker forums were analyzed by applying state-of-the-art classification and network analysis techniques. Results of the framework on a particular forum, Ashiyane, containing mobile malware-related posts and topics determined hundreds of mobile malware attachments. More recently shared mobile exploits on the forum were consistent with future trends forecasted by anti-virus expert, Symantec (Symantec, 2016). In addition, much of the mobile-related attachments in the forum were apps, also consistent with prior literature stating most mobile malware are repackaged Google Play Store apps (Zhou & Jiang, 2012). Finally, determining key mobile malware threat actors using social network measurements enforced prior research findings that many key hackers hold forum administration roles.

This framework can be applied to other hacker forum assets to determine trends and key disseminators in areas relevant to the user (e.g. banking Trojans for financial corporations, ransomware for hospitals, or phishing emails for educational institutions). By finding more information about their relevant threat landscape, users can

proactively determine key threats to their organization and accordingly respond. Furthermore, it can help discover previously unknown threat actors and their associated TTPs. This again allows the user to better inform their threat landscape, making for a more iterative and effective CTI process.

6.2 Future Directions

There are several promising future directions to expand upon this work. For example, the social network component of the study could be expanded to examine “followers”, hackers who post in threads with malware but do not implicitly share malware. Finding who followers are connected to and measuring their abilities through centrality or closeness to key hackers and later become a key hacker could help determine hacker social structures in addition to proactively identifying future key threat actors.

Another avenue for future work is using the identified mobile malware binaries for traditional malware analysis and attribution. By downloading, analyzing, and ingesting these malware characteristics for usage in CTI protocols that specialize in malware research and classification, such as Malware Attribute Enumeration and Characterization (MAEC) or YARA, organizations can be more proactive, better informed, and ultimately more effective in their defense against these types of threats.

Overall, each of the areas of expansion presented here can offer significant value to this research. In addition, they both help in the creation of proactive and holistic CTI measures. However, neither extension provides superior CTI when selecting one over the other, as both impact future proactive CTI directions and depends significantly upon an organization’s relevant threat landscape.

6.3 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. NSF DUE-1303362 (SFS).

7. REFERENCES

Abbasi, A., Li, W., Benjamin, V., Hu, S., & Chen, H. (2014). Descriptive analytics: Examining expert hackers in web forums. *Proceedings - 2014 IEEE Joint Intelligence and Security Informatics Conference, JISIC 2014*, 56–63. https://doi.org/10.1109/JISIC.2014.18

Bengio, Y., & Lecun, Y. (2007). Scaling Learning Algorithms towards AI. *MIT Press*, (1), 1–41.

Benjamin, V., & Chen, H. (2012). Securing cyberspace Identifying key actors in hacker communities. *IEEE Xplore*.

Benjamin, V., Li, W., Holt, T., & Chen, H. (2015). Exploring Threats and Vulnerabilities in Hacker Web : Forums , IRC and Carding Shops. *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*, 85–90.

Bromiley, M. (2016). Threat Intelligence : What It Is , and How to Use It Effectively. *SANS Institute*.

Chakradeo, S., Reaves, B., & Enck, W. (2013). MAST: Triage for Market-scale Mobile Malware Analysis. *ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)*, 13–24. https://doi.org/10.1145/2462096.2462100

Chollet, F. (2015). Keras. Github. Retrieved from https://github.com/fchollet/keras

Chu, B., Holt, T., & Ahn, G. (2010). Examining the Creation, Distribution, and Function of Malware On Line. *Department of Justice Abstract*, 1–183.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv*, *1*, 1–9.

Dilger, D. (2014). New Android "RAT" infects Google Play Apps, turning phones into spyware zombies. Retrieved from http://appleinsider.com/articles/14/03/07/new-android-rat-infects-google-play-apps-turning-phones-into-spyware-zombies

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., & Smith, N. a. (2015). Transition-Based Dependency Parsing with Stack Long Short-Term Memory. *Acl*, 334–343. https://doi.org/10.3115/v1/P15-1033

EY. (2014). Cyber Threat Intelligence - How To Get Ahead Of Cybercrime. *EY*, (November).

Farnham, G. (2013). Tools and Standards for Cyber Threat Intelligence Projects. *SANS Institute*, (October), 27.

Faust, K. (1997). Centrality in affiliation networks. *Social Networks*, *19*(2), 157–191. https://doi.org/10.1016/S0378-8733(96)00300-0

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: continual prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471. https://doi.org/10.1162/089976600300015015

Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

Holt, T. J. (2012). Examining the Forces Shaping Cybercrime Markets Online. *Social Science Computer Review*, *31*(2), 165–177. https://doi.org/10.1177/0894439312452998

Holt, T. J., Strumsky, D., Smirnova, O., & Kilger, M. (2012). Examining the Social Networks of Malware Writers and Hackers. *International Journal of Cyber Criminology*, *6*(1), 891–903.

Isah, H., Neagu, D., & Trundle, P. (2015). Bipartite network model for inferring hidden ties in crime data. *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, (i), 994–1001. https://doi.org/10.1145/2808797.2808842

Li, W., & Chen, H. (2014). Identifying top sellers in underground economy using deep learning-based sentiment analysis. *Proceedings - 2014 IEEE Joint Intelligence and Security Informatics Conference, JISIC 2014*, 64–67. https://doi.org/10.1109/JISIC.2014.19

Lu, Y., Luo, X., Polgar, M., & Cao, Y. (2010). Social Network Analysis of a Criminal Hacker Community. *The Journal of Computer Information Systems*, *51*(2), 31. https://doi.org/10.1108/13685201211238016

Motoyama, M., McCoy, D., Levchenko, K., Savage, S., & Voelker, G. M. (2011). An Analysis Of Underground Forums. *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference - IMC '11*, 71. https://doi.org/10.1145/2068816.2068824

Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., … Shakarian, P. (2016). Darknet and Deepnet Mining for Proactive Cybersecurity Threat Intelligence. *IEEE*, 1–6. Retrieved from http://arxiv.org/abs/1607.08583

Ponemon Institute. (2016). 2016 Cost of Data Breach Study : Global Analysis. *Ponemon Institute Research Report*, (May).

Samtani, S., & Chen, H. (2016). Using Social Network Analysis to Identify Key Hackers for Keylogging Tools in Hacker Forums. *IEEE*.

Samtani, S., Chinn, K., Larson, C., & Chen, H. (2016). AZSecure Hacker Assets Portal : Cyber Threat Intelligence and Malware Analysis. *IEEE*.

Samtani, S., Chinn, R., & Chen, H. (2015). Exploring Hacker Assets in Underground Forums. *IEEE*.

Sarvari, H., Abozinadah, E., Mbaziira, A., & Mccoy, D. (2014). Constructing and Analyzing Criminal Networks. *2014 IEEE Security and Privacy Workshops*, 84–91. https://doi.org/10.1109/SPW.2014.22

Shackleford, D. (2015). Who's Using Cyberthreat Intelligence and How? *SANS Institute*.

Stewart, S. A., & Abidi, S. S. R. (2012). Applying social network analysis to understand the knowledge sharing behaviour of practitioners in a clinical online discussion forum. *Journal of Medical Internet Research*, *14*(6). https://doi.org/10.2196/jmir.1982

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Nips*, 3104–3112. https://doi.org/10.1007/s10107-014-0839-0

Symantec. (2016). Internet Security Threat Report. *Symantec 2016 Trends*. https://doi.org/10.1016/S1353-4858(05)00194-7

Tung, L. (2014). Modded firmware may harbour world's first Android bootkit. Retrieved from http://www.zdnet.com/article/modded-firmware-may-harbour-worlds-first-android-bootkit/

Vidas, T., Votipka, D., & Christin, N. (2011). All Your Droid Are Belong To Us : A Survey of Current Android Attacks. *USENIX Conference on Offensive Technologies (WOOT)*.

Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I., & Hinton, G. (2015). Grammar as a Foreign Language. *arXiv*, 1–10. https://doi.org/10.1146/annurev.neuro.26.041002.131047

Zhang, Y., Zeng, S., Fan, L., Dang, Y., Larson, C. A., & Chen, H. (2009). Dark web forums portal: Searching and analyzing Jihadist forums. *2009 IEEE International Conference on Intelligence and Security Informatics, ISI 2009*, 71–76. https://doi.org/10.1109/ISI.2009.5137274

Zhou, Y., & Jiang, X. (2012). Dissecting Android malware: Characterization and evolution. *Proceedings - IEEE Symposium on Security and Privacy*, (4), 95–109. https://doi.org/10.1109/SP.2012.16

Appendix A: Hacker Forum Literature Review

Year	Authors	Research Objective	Methods	Data Used	Key Findings
2016	Samtani & Chen	Proactive identification of key hackers	SNA	1 English forum	Key hackers are senior members
2014	Li & Chen	Identify key malware/carders based on customer feedback	Thread classifier, Deep learning	1 forum, Zloy	Framework for identifying key sellers
2014	Abbasi et al.	Automated ID/characterization of expert hackers	IMF, X-means	1 forum, ic0de	Forum members cluster into groups
2012	Benjamin & Chen	Identifying how hackers become key actors	OLS	2 forums; 1 English, 1 Chinese	Contributing and active hackers had highest reputations
2012	Holt et al.	Understand social networks of hackers	SNA, Risk assessment	1 Russian social networking site	Skilled hackers belong to multiple groups
2016	Nunes et al.	Provide proactive CTI using darknet and deepnet data	SVM, SNA	17 darknet markets, 21 forums	Hacker web data is effective for proactive CTI
2016	Samtani et al.	Provide novel CTI and a malware portal	SVM, LDA	2 forums; 1 English, 1 Russian	Current CTI is reactive
2015	Benjamin et al.	Automated ID of potential threats in hacker web	TF/IDF , Topic Clustering	10 forums; 8 IRC channels; 4 carding shops	The Hacker Web provides actionable intelligence
2015	Samtani et al.	Reuse hacker assets for educational purposes	SVM; LDA	5 forums; 3 English, 2 Russian	10-20% of topics are key hacker tools.

Appendix B: Hacker Forum Collection

Forum	Language	Date Range	# of Posts	# of Threads	# of Members
Antichat	Russian	2002 - 10/2016	1,138,339	35,936	59,292
Ashiyane	Arabic/Persian	2003 - 10/2016	34,247	8,575	6,406
Brutezone	Russian	2011 - 10/2016	19,908	8,478	1,080
Carding Forum	English/Russian	2013 - 10/2016	18,342	6,839	1,634
Carding Masters	English		4,379	3,321	1,143
Ccc	Russian	2012 - 10/2016	760	152	393
Crimes	English	2013 - 10/2016	3,958	2,132	2,514
Cclub	English/Russian	2009 - 10/2016	150,502	8,854	9,188
Darkmoney	Russian	2012 - 10/2016	187,435	23,950	21,014
Darknetforums	English/Russian	2015 - 10/2016	14,850	3,942	3,300
Delfcode	Russian	2009 - 10/2016	459	61	38
Devil-group	English	2012 - 10/2016	24,692	5,548	3,301
Ethical Hacker	English	2005 - 10/2016	16,530	2,368	1,861
Exelab	Russian	2004 - 10/2016	328,477	20,751	13,289
Grabberz	Russian	2006 - 10/2016	175,249	13,295	6,782
Greysec	English	2015 - 10/2016	4,538	848	228
Hackthissite	English	2008 - 10/2016	35,336	4,060	6,166
Opensc	English	2005 - 10/2016	145,681	20,486	7,173
Prologic	Russian	2006 - 10/2016	27,908	5,795	3,088
Reverse4you	Russian	2009 - 10/2016	8,454	1,229	399
Sky-Fraud	English/Russian	2013 - 10/2016	31,425	7,859	5
Soqor	Arabic	2004 - 10/2016	32,773	6,965	7,238
Tuts4you	English	2004 - 10/2016	40,666	6,376	2,539
V4-Team	Arabic	2008 - 10/2016	570,213	111,101	30,309
Waraxe	English	2004 - 10/2016	25,279	7,911	4,737
Webcriminal	Russian	2007 - 10/2016	7,502	1,704	621
xakepok	Russian	2009 - 10/2016	48,351	4,529	4,107
Xeksec	Russian	2009 - 10/2016	72,082	49,467	18,832
Garage4hackers	English	2010 - 10/2016	4,012	1,072	558
ISAHackers	English	2012 - 10/2016	21,194	3,209	3,080
Hackhound	English	2012 - 10/2016	19,880	2,683	1,832
VBSpiders	Arabic	2007 - 10/2016	199,978	33,423	21,891
Zloy	Russian	2004 - 10/2016	229,817	1,611	13,145
Total:	**-**	**2004 – 10/2016**	**3,643,216**	**414,530**	**257,183**