Incremental Hacker Asset Collection and Classification for Proactive Cyber Threat Intelligence

By

Ryan Williams

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

2018

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: Ryan Williams

APPROVAL BY MASTER'S 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

ABSTRACT

Cyber-threats consistently prove a significant threat to organizational security and consumer privacy. Security breaches cost companies billions of dollars a year and compromise the personal information of millions of individuals. These cyber-attacks are conducted in a variety of methods including the use of malicious tools. Cyber Threat Intelligence (CTI) aims to learn from and combat these attacks, but current CTI efforts rely heavily on internal data that leads to reactive mitigation efforts. Hacker forums, where users can share malicious tools, provide a source of external intelligence that can be utilized in proactively defending against possible threats. This research proposes an incremental hacker forum crawler for the collection and classification of file attachments shared in hacker forums. Specifically, a web crawler has been developed to incrementally collect new attachments posted in hacker forums. Once an attachment is found, a state-of-the-art recurrent neural network classifies it into a number of possible exploit types. The results of this study indicate, among other findings, that system and network exploits are shared significantly more than other exploit types.

1 INTRODUCTION

Cyber threats continue to evolve, becoming more sophisticated in hopes to subvert security measures employed by individuals and organizations. Organizations, in particular, are high value targets for cyber-attacks due to the valuable data they manage. The average organizational cost of data breach for U.S. companies in fiscal year 2017 was $7.35 million (Ponemon, 2017). One contributing factor to this high cost is the time it takes to recover from a cyber-attack. On average, it takes 191 days to identify a data breach and another 66 days to contain it (Ponemon, 2017).

Breaches can be caused by any number of different vulnerabilities. While system glitches and human negligence can be contributing factors, attacks from malicious insiders or criminals are more prominent, contributing to 47% of breaches (Ponemon, 2017). Many organizations operate internal systems such as security information and event management systems (SIEMs), to provide insights into various threats. Internal systems can provide valuable information after a breach, but are typically considered reactive forms of CTI.

Open source intelligence (OSINT), on the other hand, can provide companies with insights into possible threats. OSINT is intelligence collected from publicly available sources, and can offer significant value to proactive CTI by alerting organizations to threats they were not previously aware of (Bromiley, 2016). One form of OSINT is hacker community data. The hacker community is comprised of hacker forums, darknet markets, carding shops, and internet relay chat (IRC) channels. Compared to other hacker community platforms, hacker forums provide data richness (metadata), data permanence, freely available Tools, Techniques, and Procedures (TTP), and usually less vetting.

Users on hacker forums discuss potential attacks and share cyber-attack assets such as attachments, source code, tutorials, and more (Samtani et al., 2015). The BlackPOS malware, used in breaches at Target and Home Depot, was being shared in hacker forums months after the breaches occurred (Samtani et al., 2016). While it is unknown whether the exploit was shared before the breaches, the fact that it was still circulating on hacker forums means it was still a potential threat to other organizations managing insecure systems. Information found in hacker forums can directly inform proactive cyber threat intelligence and mitigation. Figure 1 provides a typical example of a hacker forum post.

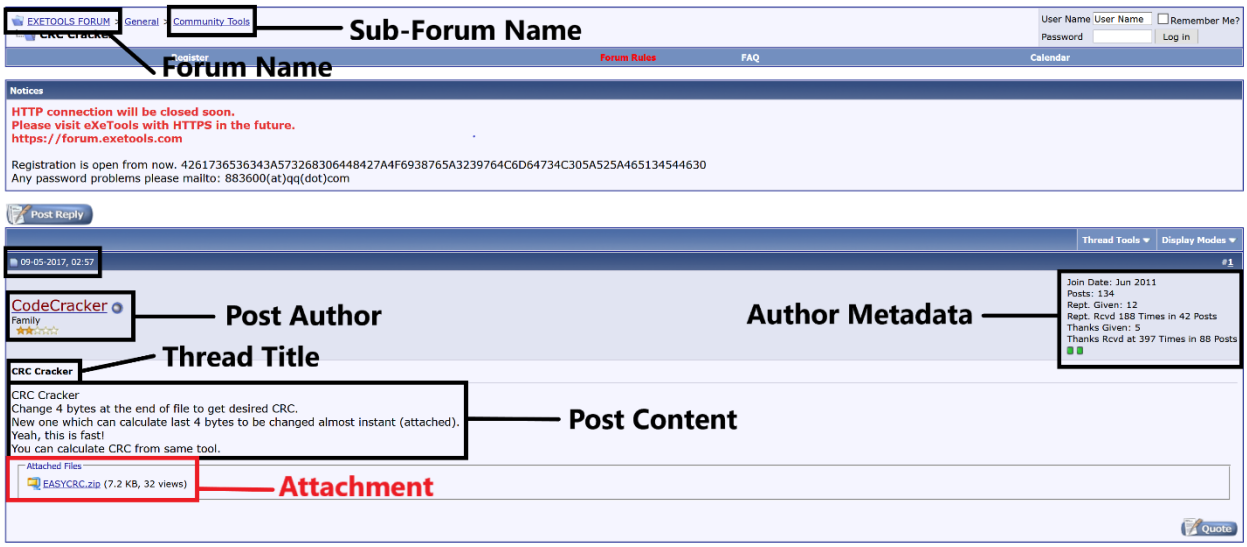

Figure 1. Example Hacker Forum Posting

Altogether, hundreds of hacker forums exist that accommodate hundreds of thousands of members who generate millions of posts containing tens of thousands of malicious assets. Ashiyane, a popular Arabic hacker forum, boasts 101,000 threads containing 712,000 posts generated by 412,000 users. Given the actionable intelligence found in hacker forum data, this study aims to provide an incremental approach to collecting and classifying hacker forum attachments for proactive CTI.

2 LITERATURE REVIEW

To form the basis of this research, literature on cyber threat intelligence is reviewed to understand current approaches in generating useful threat intelligence. Web forum crawling is also reviewed to learn what techniques are used to navigate and index web-based forums. Finally, previous hacker forum collection efforts are studied to examine what techniques and findings previous efforts have found when collecting hacker forums.

2.1 Cyber Threat Intelligence

Threat intelligence is the process of understanding the threats to an organization based on available data points (Bromiley, 2016). As an extension, CTI is threat intelligence related to computers, networks, and information technology (Farnham, 2013). CTI can help victims identify delivery mechanisms, indicators of compromise, actors, and specific motivators (Shackleford, 2015).

While CTI can provide valuable insights, it is often criticized for relying too heavily on the use of internal information including log data, honeypots, IDS and IPS output, and SIEM output. This information is highly relevant to the organization collecting it, but often leads to a reactive approach to CTI. Ultimately, internal threat intelligence can help companies give context to what they know about the attack, how they have been attacked, and what they are currently protecting (Bromiley, 2016).

Relevant external CTI data can be actionable and help give context to internal data (Bromiley, 2016). External threat intelligence can help companies understand what they do not know, how they may been attacked, and what they should be focused on protecting (Bromiley, 2016). There are many different sources of external CTI, including external threat feeds (Settanni et al., 2017), social media (Mittal et al., 2015), and darknet community data (Samtani et al., 2015; Bou-Harb, 2016;Grisham et al., 2017). Hacker forums, a form of darknet community data, provide insights into what malicious users are discussing, planning, and sharing (Samtani et al., 2015).

Proactive CTI involves combining internal and external threat intelligence. Timely and comprehensive collection of this data is vital in providing relevant and actionable CTI.

2.2 Web Forum Crawling

Internet forums are platforms where users can interact with others through discussion of various topics. Typically, users can share information through text, images, videos, links, and attachments. Forums are HTML based, making it possible to navigate and collect their contents using a software tool called a web crawler. Web crawlers are software programs that traverse the internet by following hyperlinks and collect web pages using the HTTP protocol (Fu et al., 2010). Their usual intention is to create a local index of web pages (Fu et al., 2010). Traditional web crawlers adopt a Breadth-First Search (BFS) strategy to navigate hyperlinks (Jiang et al., 2013). Figure 2 provides a traversal graph based on the BFS strategy.

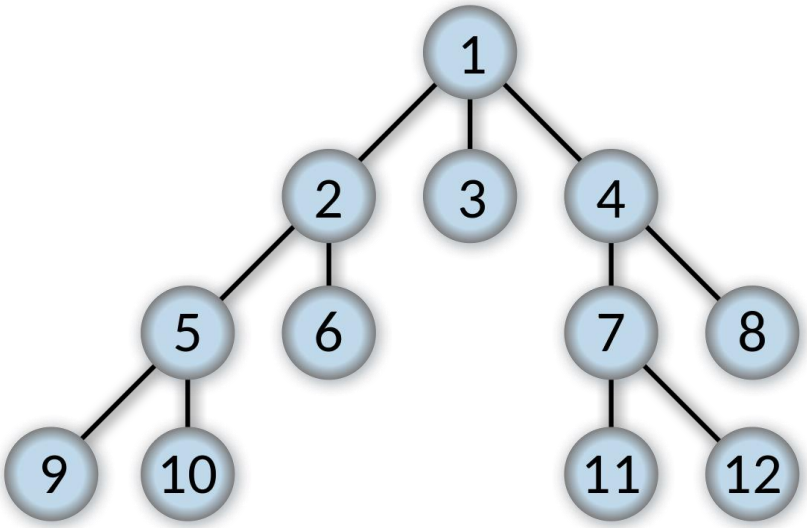

Figure 2. Breadth-First Search Traversal Graph

Characteristics of internet forums make generic web crawlers inefficient for data collection (Jiang, et al., 2013). These characteristics include:

- Duplicate links – links with different URLs that point to the same page

- Uninformative pages – pages with no relevance to the crawler's purpose (e.g. login page, FAQ, etc.)
- Page-flipping links – links that connect multiple pages within one thread

Hacker forums are built on similar frameworks (e.g. vBulletin) that other forums use. Beyond normal forum limitations, hacker forums tend to employ certain anti-crawling measures (Baravalle et al., 2016). These include:

- Authentication – requiring credentials to access content
- Turing tests – differentiating real users from software robots
- Throttling – limiting the amount of page requests a user can make in a period of time
- Obfuscation – generating text with Javascript
- Network traffic analysis – determining abnormalities in network packets and requests

Efforts have been made to subvert anti-crawling measures. Human intervention can help crawlers pass authentication tests such as CAPTCHA, artificially limiting the speed at which the crawler requests new web pages can prevent the targeted site from becoming suspicious, and manually creating a new session by restarting the crawling process frequently can allow continuous collection (Baravalle et al., 2016).

2.3 Hacker Forum Collection Efforts

Previous efforts have been made to collect and analyze information found on hacker forums. Table 1 provides a breakdown of past hacker forum collection efforts.

Authors	Forums Crawled	Data Collected	Collection Procedure
Fu et al., 2010	**109** Middle Eastern, Latin American, U.S.	**~ 4,000,000** files (Text-based, HTML, images, etc.)	Incremental
Macdonald et al., 2015	**1** Unspecified	**Unspecified** forum posts	Batch
Samtani et al., 2015	**5** English, Russian	**3,251** attachments; **14,944** source code files; **671,633** forum posts	Batch
Benjamin et al., 2015	**10** English, Russian	**99,353** forum posts	Batch
Nunes et al., 2016	**21** English	**162,872** forum posts	Batch
Grisham et al., 2017	**4** English, Russian, Arabic	**43,462** attachments; **481,922** forum posts	Batch

Table 1. Previous Hacker Forum Collection Efforts

These collection efforts have focused primarily on assets (Fu et al., 2010; Samtani et al., 2015; Grisham et al., 2017) and user posts (Benjamin et al., 2015; Macdonald et al., 2015; Nunes et al., 2016). Assets can take many forms including attachments, source code, and tutorials shared on hacker forums (Samtani et al., 2015). Many of these efforts utilize batch collection methods to crawl and gather data. This means the forum is collected all at once, often without any intention of re-crawling the forum later on to collect newly generated posts.

3 RESEARCH GAPS AND QUESTIONS

Based on prior studies found in the literature review, a number of research gaps have been identified. First, current CTI efforts focus heavily on the use of internal information to generate threat intelligence. This means current security measures are often handled reactively instead of

proactively. Second, web forum characteristics make traditional web crawling techniques ineffective for efficient navigation and indexing. Finally, previous hacker forum collection efforts have been primarily focused on batch collection of hacker assets. As threats evolve over time, these static collections become less insightful. With these research gaps in mind, the following research questions have been proposed to guide the study:

- Can incremental crawling be applied effectively to hacker forum collection?
- What kinds of attachments are being shared on hacker forums?
- What value can an up-to-date hacker forum collection provide CTI?

4 RESEARCH DESIGN

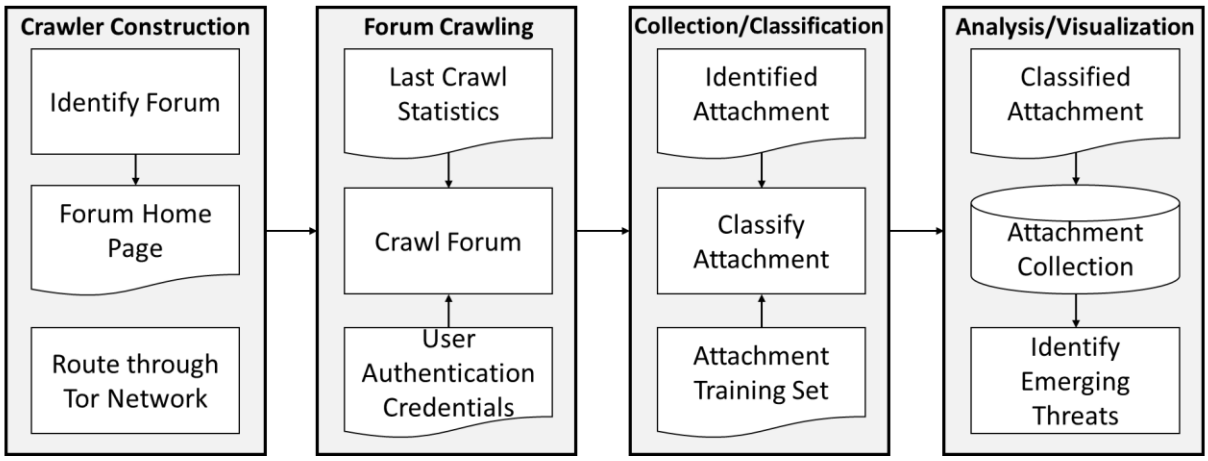

Figure 3. Research Design Diagram

In order to answer the proposed research questions, a four-phase research design has been developed (Figure 3). The four phases include crawler construction, forum crawling, collection and classification, and analysis and visualization. Each phase is described in greater detail below.

4.1 Crawler Construction

This phase focuses on the initial crawler configuration. The crawler itself is written in the Python programming language. Python provides a great number of libraries dedicated to web crawling, HTML parsing, data science, and other important functionalities needed for this study. The bulk of the crawler's functionality is provided by the "requests_html" library for HTML requests, the "BeautifulSoup" library for HTML parsing, and the "keras" library for classification.

From the literature review, it was determined that a traditional web crawling approach would not be effective for crawling web forums. On top of this shortcoming, an incremental approach to asset collection provided a number of challenges. It is not enough that the crawler be able to collect every attachment on a hacker forum, it must also revisit that hacker forum and only collect new attachments that it has not seen before. To accomplish this, the crawler utilized the underlying forum framework (i.e. HTML structure) and metadata (e.g. post date) to target specific areas of the forum that need to be crawled.

Framework	# of Forums	% of Forums
vBulletin	21	28.38%
XenForo	15	20.27%
Invision	9	12.16%
MyBB	6	8.1%
phpBB	5	6.76%

Table 2. Distribution of Forum Frameworks from 74 Hacker Forums

Unfortunately, there are many different forum frameworks, each with their own unique structure and naming schemes. During an initial analysis of different forum frameworks, it was discovered

that vBulletin was the most widely used framework for hacker forums. 74 hacker forums were categorized based on their underlying framework, with Table 2 displaying the results.

Based on this finding, the crawler was tailored to work with hacker forums utilizing the vBulletin framework. Since forums generally follow a similar structure, this crawler could be modified to crawl other frameworks by changing the specific HTML tags it searches for on each page. Once a vBulletin hacker forum has been identified by the user, the home page is passed to the crawler. Since many of these forums exist on the Dark Web, all traffic is routed through Tor, a network of servers running specialized software that provide anonymity for the user. This ensures that the crawler remains anonymous while also allowing it to access content hidden from the surface web. After the connection to Tor has successfully been established, the crawler can begin searching for attachments.

4.2 Forum Crawling

This phase of the research design focuses on the incremental approach to forum crawling. The crawler uses a Depth-First Search (DFS) approach to collecting hacker forums. As opposed to breadth-first search, DFS aims to explore each branch as far as possible before moving on. Figure 4 provides a traversal graph based on the DFS strategy. In the context of hacker forums, the crawler starts with one sub-forum and crawls each topic and post within that sub-forum before moving on to the next sub-forum. Based on the structure and naming schemes of the vBulletin framework's HTML code, the crawler can target specific links on the page and follow a systematic approach to attachment collection. If the forum requires an account to access its content, the user can create an account and provide the credentials to the crawler. The crawler will sign into the forum with the credentials and begin its collection efforts.

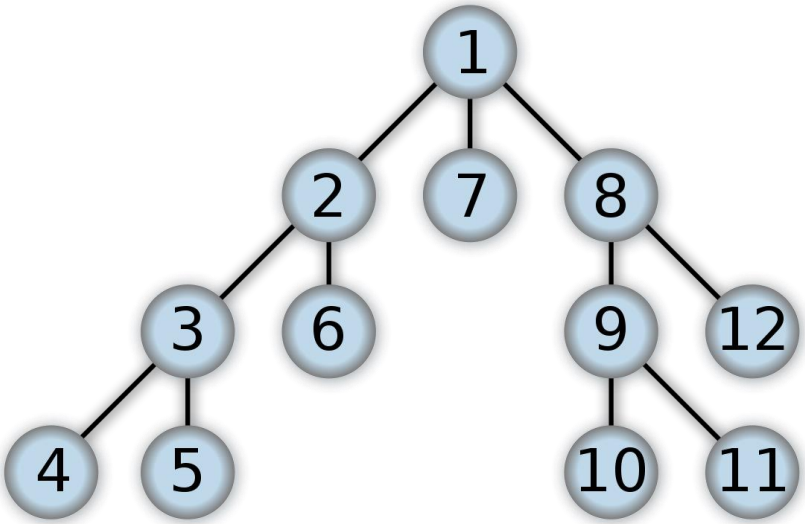

Figure 4. Depth-First Search Traversal Graph

Starting at the home page, the crawler will identify all of the forum's sub-forums and begin crawling each sub-forum one by one. If a sub-forum is contained within another sub-forum, which is common with vBulletin forums, the crawler will also collect those sub-forums. As forums are characterized by their data richness, there is a great deal of information that can be parsed from the HTML code and utilized to make the crawler more efficient. For instance, within each sub-forum, user-made topics are already listed in chronological order by most recent activity. With this information, the crawler can quickly determine when the sub-forum was last active and compare that date to the date that sub-forum was last crawled. If there is any new activity, the crawler only has to parse through the new user topics and then move on to the next sub-forum. Within each topic, posts are also ordered chronologically, with the oldest posts listed first. In order to maximize efficiency, the crawler starts crawling at the very last post within a thread. With this strategy, the crawler can quickly identify if any new posts have been made in topics it has already crawled.

Another important bit of information that forums usually provide is whether an attachment exists within a user topic. As the focus of this research is collecting attachments in hacker forums, knowing which user topics contain attachments significantly increases the efficiency of the crawler. Once a new topic is found and it is discovered to contain an attachment, the crawler will enter that topic and begin parsing through user posts to find and collect the attachment's information.

4.3 Collection and Classification

This phase is composed of two major sections. The first section focuses on training a Recurrent Neural Network (RNN) to be used for classifying hacker forum attachments. An RNN was chosen for classification because it can be used in text mining by representing words as a vector. Forum posts, and the majority of the information that can be collected about them, are largely text-based. With this in mind, a collected attachment, and all of its contextual information, can be represented as a vector and input into an RNN for classification.

To determine the best performing RNN for hacker attachment classification, three different RNNs were benchmarked on a gold standard set. These include a basic RNN, a Gated Recurrent Unit (GRU) RNN, and a Long-Short Term Memory (LSTM) RNN. The data used to train and test these RNNs was provided by previous research focused on collecting and classifying attachments on hacker forums (Samtani et al., 2015). This gold standard set is a collection of roughly 15,000 hacker forum attachments and their contextual information (e.g. sub-forum name, author name, post content, etc.) retrieved from multiple hacker forums. Latent Dirichlet Allocation (LDA) was used to identify prevalent themes for each attachment post and, based on the LDA results, experts manually labeled topics for each. With these topics, attachments were labeled by exploit type, including (Samtani et al., 2015; Grisham et al., 2017):

- System – Exploit vulnerabilities within a given system
- Web – Exploit web specific applications and technologies
- Network – Exploit TCP/IP vulnerabilities to damage a network
- Database – Exploit vulnerabilities with database software and technologies
- Mobile – Exploit mobile operating systems

Table 3 displays the results of the RNN benchmark with the gold standard set. Each model was trained on 80% of the data and tested on the remaining 20%. Based on these results, the trained LSTM RNN was used for classifying new attachments collected by the forum crawler.

Model	Precision	Recall	F-Measure
Basic RNN	90%	90%	90%
GRU	96%	96%	96%
LSTM	**97%**	**98%**	**98%**

Table 3. RNN Models Benchmark

The second section of this phase occurs in conjunction with the previous phase of the research design, forum crawling. As the crawler collects new attachments and their contextual information, this information is immediately passed to the trained LSTM RNN for classification. Once the LSTM RNN has classified the attachment, the information is parsed into a database for storage and further analysis.

4.4 Analysis and Visualization

The last phase of the research design focuses on pulling out the valuable information that an up-to-date hacker forum collection can provide CTI. Once attachments have been collected, classified, and stored in a database, this information can be taken and analyzed to gain further

insights. This study focuses on two main areas for analysis, author activity and exploit postings. To accomplish this, Tableau is used to visualize characteristics within the data and help identify emerging threats.

5 RESEARCH TESTBED

A total of 10 hacker forums were collected, including OpenSC, Garage4hackers, Hacksden, AntiOnline, Crackingzilla, WebCracking, SafeSkyHacks, Ashiyane, Hack, and Haker. These forums were chosen for a number of reasons. First, they are all built using the vBulletin forum framework. They also allow users to embed attachments directly into their posts. Finally, they can be accessed without monetary payment. Table 4 displays the 11 features that were collected for each attachment. Each attribute that in bold was passed to the LSTM RNN for classification.

Attribute	**Description**
Forum Name	Title of the forum currently being crawled
Author Name	Name of the user that made the post
Sub Forum ID	Unique identifier for the sub forum containing the attachment
Sub Forum Name	Title of the sub forum containing the attachment
Thread ID	Unique identifier for the thread containing the attachment
Thread Name	Title of the thread containing the attachment
Post ID	Unique identifier for the post containing the attachment
Post Date	Date that the post containing the attachment was made
Attachment URL	Direct web address to the hosted attachment
Attachment Name	Name of the attachment file
Exploit Type	Predicted exploit type of the attachment file

Table 4. Collected Attributes for Each Attachment Found

Table 5 breaks down the statistics of the 10 forums that were crawled and collected.

Forum	Language	# of Sub Forums	# of Threads	# of Posts	# of Assets
Hacksden	English	70	10,359	61,534	77
Crackingzilla	English	61	11,451	167,206	1
Garage4Hackers	English	47	1,544	8,620	51
OpenSC	English	56	22,897	184,211	1,179
AntiOnline	English	39	14,771	160,897	77
WebCracking	English	109	7,832	92,025	7
SafeSkyHacks	English	100	12,780	31,733	89
Ashiyane	Arabic	49	65,251	538,708	1,388
Hack	Polish	52	10,452	63,515	52
Haker	Polish	34	924	9,374	9

Table 5. Forum Collection Statistics

Ashiyane contained significantly more threads and posts than other hacker forums collected. Unsurprisingly, it also contained the most attachments. OpenSC also contained significantly more attachments than most of the other forums, but was not significantly larger when considering number of threads and posts. A few forums contained a surprising lack of post attachments. Crackingzilla, WebCracking, and Haker all hosted less than 10 attachments while containing a large amount of threads and posts. This could point to different policies on sharing attachments that could prevent attachments being directly embedded within forum posts.

6 RESULTS AND DISCUSSION

6.1 Attachment Collection

In total, 2,930 attachments were collected and classified from the 10 hacker forums. Table 6 displays the distribution of exploit types. For this study, attachments were only collected if they

were directly embedded in the user's post. More attachments exist in hacker forums but they are not provided directly in the post itself. In this situation, the provider of the exploit will require users who want to access the attachment to follow a hyperlink leading to a third party file sharing site. Occasionally, the provider will also lock the attachment behind a set of credentials that they provide in the post. For this study, these external attachments were ignored because contextual information such as file name that the LSTM RNN uses during classification is not always readily available. It is also not guaranteed that the hyperlinks provided actually lead to the supposed exploit.

Exploit Type	Number of Attachments	% of Total Attachments
System	1738	59.32%
Network	910	31.06%
Web	147	5.0%
Database	121	4.1%
Mobile	14	0.5%

Table 6. Exploit Type Distribution

System exploits make up a majority of the attachments shared on hacker forums at 59.32%. These attachments cover a wide range of different exploits including crypters, keyloggers, and remote access trojans (RATs). Network exploits also make up a significant chunk of the shared attachments at 31.06%. This category includes exploits such as botnets and distributed denial of service (DDoS) attacks. Web (i.e. XSS, SQL injection), database (i.e. MySQL attack), and mobile (i.e. Android attack) exploits comprise a relatively small number of attachments being shared at 9.6% combined.

6.2 Threat Identification

With the attachments collected and classified, a wide variety of insights can be extracted from the data. For this study, visualizations created in Tableau were utilized to identify trends and emerging threats in hacker forums by perform an analysis on two primary areas:

- Exploit postings – What exploits types are being shared, when are exploits being shared, and what individual exploits have been shared recently
- Author activity – Which users are the most active, which forums are the most active, and what forums share the most of each exploit type

6.2.1 Individual Exploit Postings

Proactive CTI depends on timely information. Figure 5 provides an example of a dashboard that lists individual exploits that have been collected by the forum crawler.

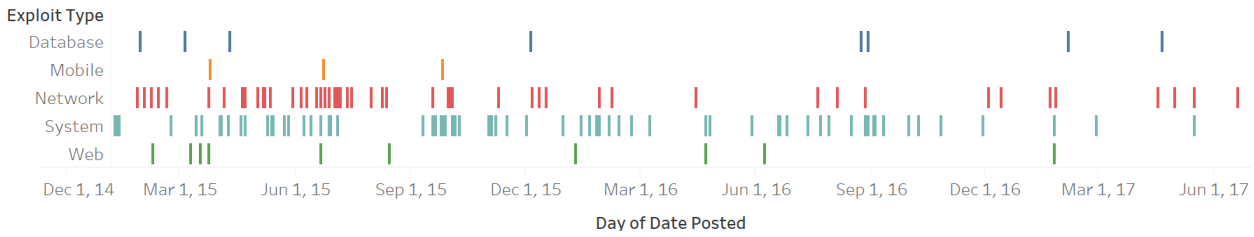

Figure 5. Individual Exploit Postings

The exploits are displayed on a timeline that can be adjusted to display information from the last month or show trends over multiple years. The figure above shows all of the exploits that were shared from the beginning of 2015 to the middle of 2017. The figure shows that exploit postings are not as frequent now as they were in 2015. When using the dashboard, the user can also hover over each exploit to reveal more information including exploit name, author, forum/sub-forum/thread, and the URL of the attachment download. This information becomes significantly more important for recently shared exploits, where users can see exactly what new threats are

being shared in hacker forums. As the incremental crawler continues to collect new exploits, analysis would be immediately available for proactive threat intelligence.

6.2.2 Exploit Count Timeline by Month and Year

Other visualizations can be utilized to better identify exploit trends over time. Figure 6 reinforces that exploits are not being shared as much on hacker forums as they were a few years ago.

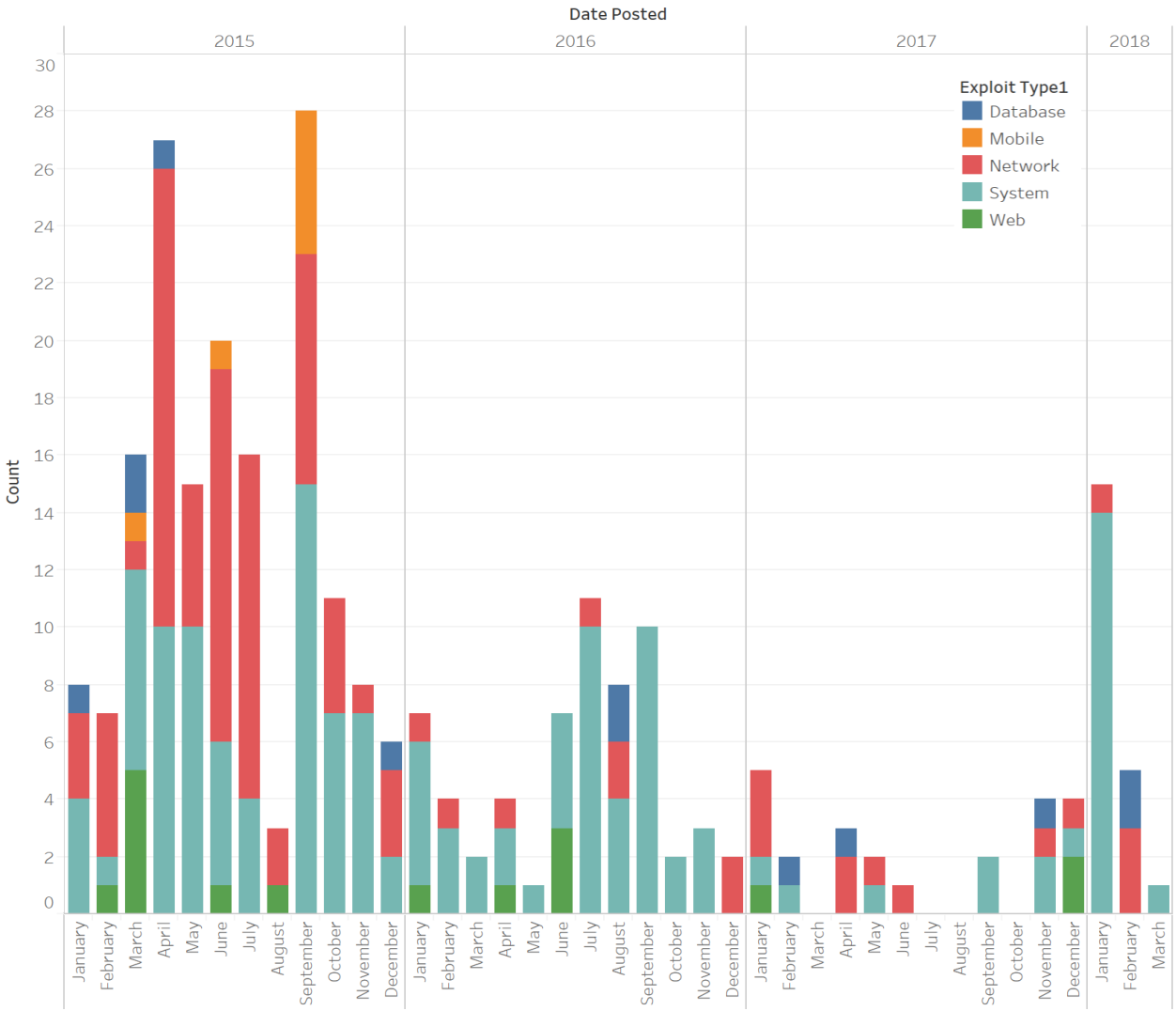

Figure 6. Exploit Count by Month/Year, Bar Chart

2017 appeared to be a particularly slow year for exploit postings. For four of the twelve months that year (March, July, August, and October) the crawler found zero exploits that were shared.

Over the past three years, these were the only months where no exploits were found. Surprisingly though, January of this year saw the highest number of exploit postings (15) since September of 2015. The majority of these recent postings have been system exploits, providing valuable direction for where to focus CTI efforts for this upcoming year.

6.2.3 Author Activity All-Time

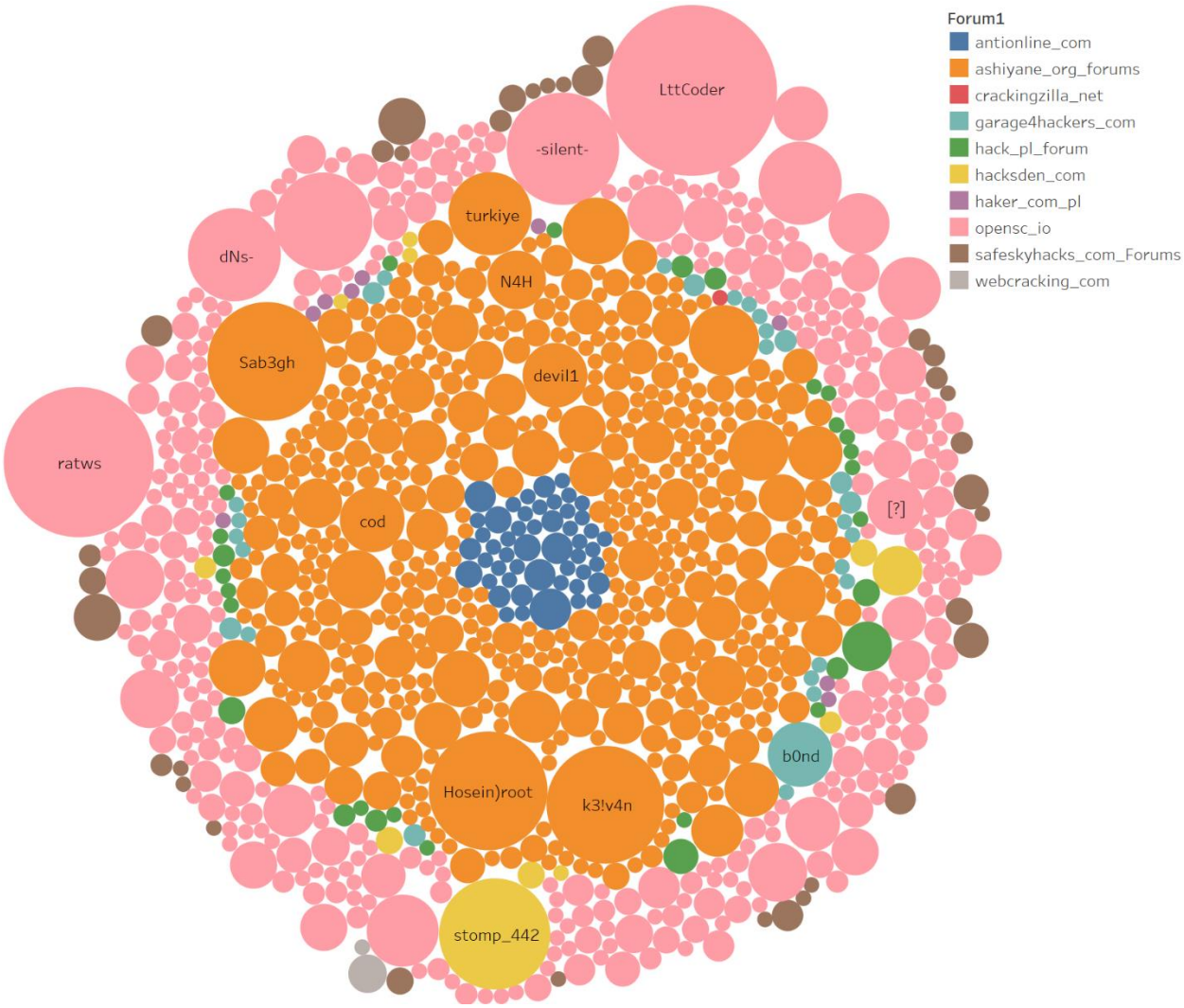

Figure 7. Author Activity All-Time, Packed Bubbles

While it is important to consider the specific exploits being shared in hacker forums, it can also be valuable to look at which communities are the most active and how many authors are

contributing to exploit dissemination. This information can point to what types of exploits each forum shares the most. While forums like Ashiyane contain postings of many different exploit types, OpenSC and Garage4Hackers focus primarily on system and network exploits. While system and network exploits are the most common postings, this type of analysis becomes more valuable when forums are identified that focus primarily on the lesser shared exploit types such as database and web. These unique forums can be closely monitored for lesser utilized but potentially overlooked threats. Figure 7 displays the all-time most active authors, based on number of attachments posted.

The size of the author's bubble indicates how many exploits they have shared and the color of the bubble shows what forum that author was posting in. As we previously saw in Table 5, OpenSC and Ashiyane are by far the most active hacker forums in both number of active authors and number of exploits shared. Both of these forums have many significant contributors, but this is not the case for all forums. Hacksden appears to have one primary contributor for exploit postings, while AntiOnline has more active authors but none of them stand out.

6.2.4 Author Activity by Year and Exploit

By combining both exploit data and author activity we can gain further insights into which exploits, authors, and forums should warrant more attention. Figure 8 provides a breakdown of author activity by year and exploit type.

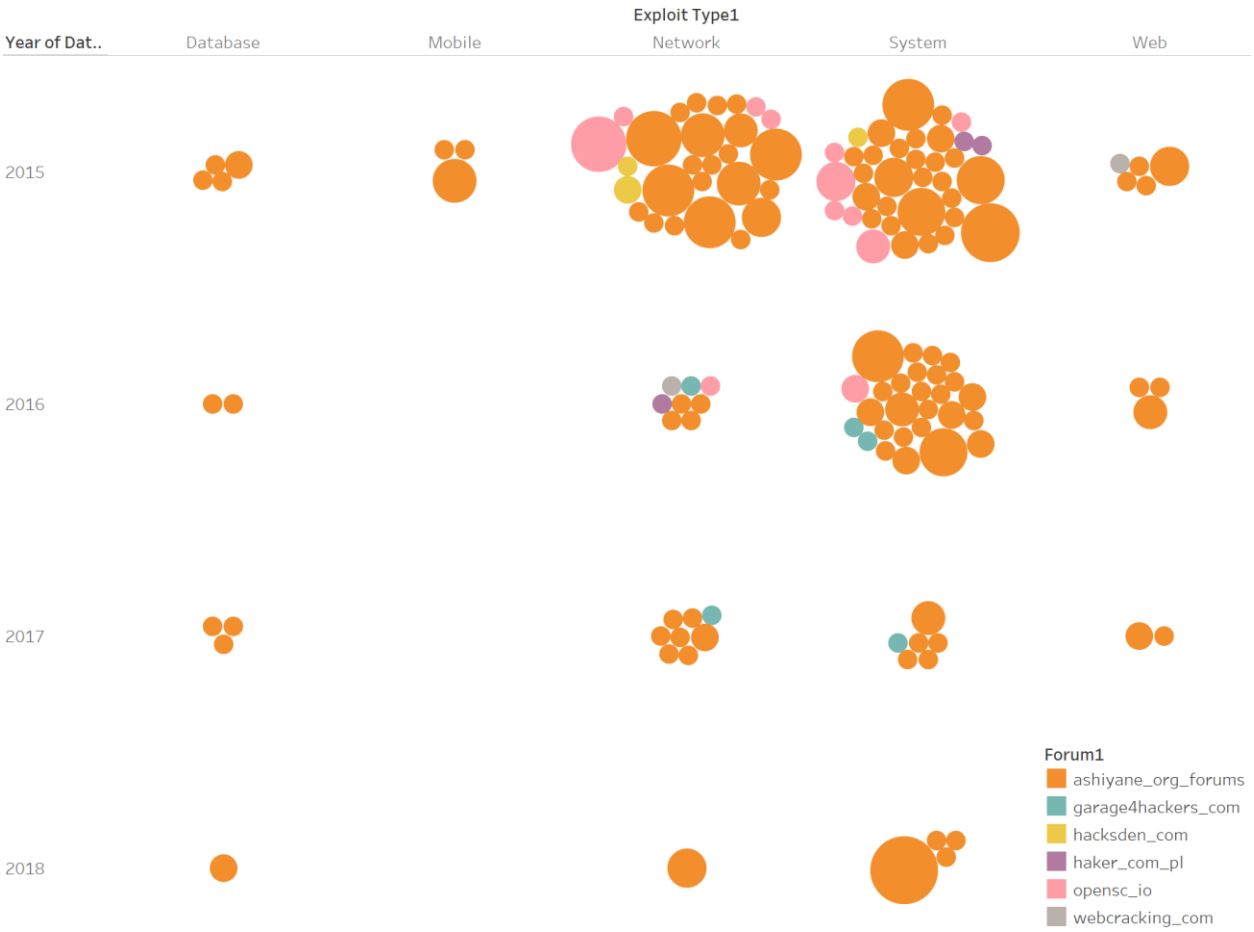

Figure 8. Author Activity by Year and Exploit, Packed Bubbles

This visualization reinforces previous findings while also providing new insights not readily apparent from the other visualizations. System and network exploits continue to remain the most posted attachments within hacker forums. On top of this, there were many more exploit postings in 2015 than in the years after. Surprisingly, while Ashiyane has been expectedly active, OpenSC has been relatively dormant for the past few years. This visualization, in particular, provides valuable insights into what exploit types are being shared the most and which forums are currently the most active in providing those exploits.

7 CONCLUSION AND FUTURE DIRECTIONS

Cyber threats are an ever growing problem. Organizations continue to amass large amounts of personal information on users and customers. This information can be extremely valuable to hackers looking to sell it on the dark web. In order to protect user information, organizations must rely on more than internal intelligence to mitigate threats. The internally focused reactive approach to CTI is not efficient or effective in identifying and seeking to proactively prevent future threats as they evolve to subvert security measures. External CTI can prove invaluable for proactive threat mitigation. One approach to external CTI is hacker forum collection. These forums contain a great deal of insights into what exploits are being shared and discussed by hackers within the community.

In this study, an incremental crawler combined with an LSTM RNN is proposed to provide proactive CTI through the collection and classification of hacker forum attachments. Specifically, a Python-built web forum crawler is utilized to gather attachments from vBulletin based hacker forums and classify them into a number of different exploit categories. OSINT becomes less valuable over time, and for that reason, this crawler is built to continually crawl available forums for new attachments. In this way, the exploit collection continues to grow and provide value to CTI.

The results of this study indicate a large number of exploits shared on hacker forums target system and network vulnerabilities, as opposed to web, database, and mobile vulnerabilities. 90.38% of the 2,930 collected attachments were classified as either system or network exploits. The findings also imply that exploit sharing has become less popular over the past few years, with 2017 being a particularly slow year. In 2013, over 450 attachments were shared on the hacker forums crawled for this study. In comparison, only 23 attachments were collected for the

year of 2017. This study also proposes a number of different approaches to extracting value from the exploit collection. Specifically, exploit postings and author activity are analyzed to gain insights into which exploit types, authors, and forums should be more heavily observed for proactive CTI.

The approach to incremental hacker forum collection proposed in this study can be expanded in a number of different directions. As of now, only 10 forums have been crawled and collected. This is primarily a limitation with the focus on the vBulletin framework. With a bit of work, the crawler could be repurposed to become compatible with other forum frameworks. This would allow the collection of more forums and greatly enhance the value of the hacker exploit collection. The crawler could also collect new assets beyond attachments. Users in hacker forums also share source code as well as tutorials on how to carry out certain attacks. These different assets could be collected and classified with a similar approach this study took to attachments. Finally, more in-depth analysis and visualization of the data can be performed to provide greater insights.

8 REFERENCES

Benjamin, V., Li, W., Holt, T., & Chen, H. (2015). Exploring Threats and Vulnerabilities in Hacker Web: Forums, IRC and Carding Shops. *IEEE International Conference on Intelligence and Security Informatics: Securing the World through an Alignment of Technology, Intelligence, Humans and Organizations, ISI*. https://doi.org/10.1109/ISI.2015.7165944

Bou-harb, E. (2016). A Probabilistic Model to Preprocess Darknet Data for Cyber Threat Intelligence Generation. *IEEE ICC Communication and Information Systems Security Symposium*. https://doi.org/10.1109/ICC.2016.7510881

Bromiley, M. (2016). Threat Intelligence: What It Is, and How to Use It Effectively. *SANS Institute InfoSec Reading Room*, 15.

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

Fu, T., Abbasi, A., & Chen, H. (2010). A Focused Crawler for Dark Web Forums. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*. https://doi.org/10.1002/asi

Grisham, J., Samtani, S., Patton, M., & Chen, H. (2017). Identifying Mobile Malware and Key Threat Actors in Online Hacker Forums for Proactive Cyber Threat Intelligence. *IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI*. https://doi.org/10.1109/ISI.2017.8004867

Jiang, J., Song, X., Yu, N., & Lin, C.-Y. (2014). FoCUS : Learning to Crawl Web Forums. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *25*(6), 1293–1306.

Macdonald, M., Frank, R., Mei, J., & Monk, B. (2015). Identifying Digital Threats in a Hacker Web Forum. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 926–933. https://doi.org/10.1145/2808797.2808878

Mittal, S., Das, P. K., Mulwad, V., Joshi, A., & Finin, T. (2016). CyberTwitter: Using Twitter to generate alerts for cybersecurity threats and vulnerabilities. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 860–867. https://doi.org/10.1109/ASONAM.2016.7752338

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 International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI*. https://doi.org/10.1109/ISI.2016.7745435

Ponemon Institute LLC. (2017). 2017 Cost of Data Breach Study, (June), 1–34. Retrieved from https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=SEL03130WWEN&

Samtani, S., Chinn, K., Larson, C., & Chen, H. (2016). AZSecure Hacker Assets Portal: Cyber Threat Intelligence and Malware Analysis. *IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI*. https://doi.org/10.1109/ISI.2016.7745437

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

Settanni, G., Shovgenya, Y., Skopik, F., Graf, R., Wurzenberger, M., & Fiedler, R. (2017). Acquiring Cyber Threat Intelligence Through Security Information Correlation. *IEEE International Conference on Cybernetics, CYBCONF 2017 - Proceedings*. https://doi.org/10.1109/CYBConf.2017.7985754

Shackleford, D. (2015). Who's Using Cyberthreat Intelligence and How? *SANS Institute InfoSec Reading Room*, 25.