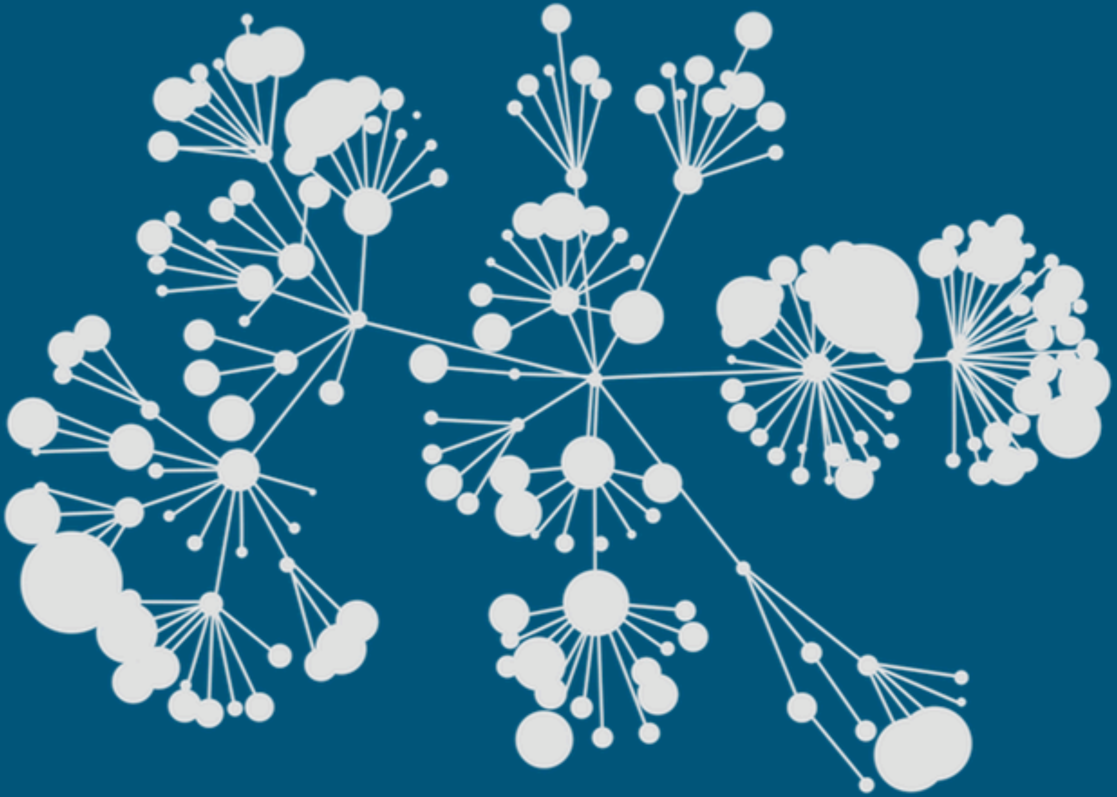
Exploring Hackers Assets: Topics of Interest as Indicators of Compromise

- Mohammad Al-Ramahi
- Izzat Alsmadi
- Joshua Davenport

Texas A&M University-San Antonio



Agenda

1. Introduction
2. Previous/Other Contributions
3. Goals and Approaches
4. Experiment and Analysis
5. Summary

Introduction

- The need to develop actionable intelligence that is proactive is very critical to current security controls and systems.
- Hackers and hacking techniques continue to grow and become more sophisticated.
- As such Security teams start to adopt proactive and offensive approaches within hackers’ territories.
- In this scope, we proposed a systematic approach to automatically extract "topics of interest, ToI" from hackers’ websites.

Introduction

- Those can eventually be used as inputs to actionable security controls or Indicators of Compromise (IOS) collectors.
- As a showcase, we selected the hackers' news website "CrackingFire".
- ToI can be integrated into Indicators of Compromise (IoC) and once correlated with other signs of attacks further cybersecurity offense or defense actions will be triggered.
- We also developed our own dark web crawler and evaluate extracting ToI.
- We observed the types of challenges in both the crawling and the processing stages.

Previous/Other Contributions

- A significant contribution in this area comes from AZSecure team (https://www.azsecure-data.org/home.html) lead by Arizona State University
- Examples of published papers: (e.g. Sagar et al 2015, Park et al 2016, Hansen et al 2017, Sapienza et al 2017, Deliu et al 2017, Abhishek et al 2018, and Pastrana et al 2018).

Previous/Other Contributions

- Similar to our goal, those papers attempted to crawl selected hackers' websites to extract security or cyber intelligence related knowledge.
- Some of those contributions correlated extracted knowledge from hackers' websites with information from cybersecurity experts collected from Online Social Networks (OSN) such as Twitter.

Previous/Other Contributions

- The commonality of the papers surveyed in the relevant literature is related to (1) the goal; extracting useful knowledge and (2) the target; hackers' websites and forums.
- Differences are on what type of knowledge to extract and how to approach the extraction and the analysis activities.

Previous/Other Contributions

- Hackers' profiles include extreme personalities such as those who are looking for high publicity and ego-fulfillment.
- On the other hand, hackers can be much conservative in their publicity and desire to be visible to the public.

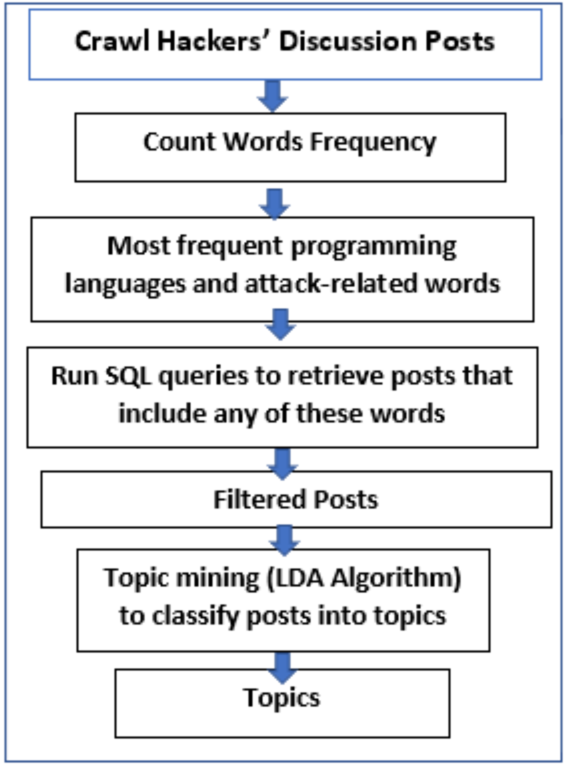
Previous/Other Contributions

- Several challenges related to the overall process of extracting knowledge from hackers’ websites are discussed in those different papers.
- Some of the main challenges are related to hackers’ techniques in information hiding and masquerading in addition to their different techniques to block such content from public accessibility.

Goals and Approaches

- Figure 1 simplifies our model to extract topics of interest from hackers’ websites.
- Our focus in this paper is on the technical aspects of attacks and malware.
- We used a popular topic mining algorithm, (LDA algorithm) to extract the best selection of topics as classes.

Our Topics of Interest Extraction model



Crawl Hackers' Discussion Posts

Count Words Frequency

Most frequent programming languages and attack-related words

Run SQL queries to retrieve posts that include any of these words

Filtered Posts

Topic mining (LDA Algorithm) to classify posts into topics

Topics

Data Collection and Preparation

- In this study, we used CrackingFire forum public dataset that contains 37,572 forum posts ranging from 4/7/2011 – 2/21/2018 (Available http://www.azsecure-data.org/ [Zhang et al 2018).
- This dataset facilitates cybersecurity research concerning with analysis of hacker assets and especially source code analysis of these assets.
- To prepare the dataset, we excluded stop words and used the Term Frequency Inverse Document Frequency (TF-IDF) technique to represent each post.

Topic Modeling: LDA

- Topic models are types of statistical algorithms for extracting the main topics in a collection of documents.
- Latent Dirichlet Allocation (LDA) is one of the common topic modeling algorithms used in literature.
- The algorithm produces a set of topics with a probability distribution over words in each topic.
- The algorithm also generates probability distributions over topics for each document.

Experiment and Analysis
1) Extracting ToI from CrackingFire

- We selected one hackers’ forum or website, CrackingFire, to evaluate our model.
- However, our model is generic and can be expanded to other websites.
- The LDA identified 10 topics and within each topic showed the top-10 words and their relative weight (i.e. probability).
- The naming of topics was based on the logical connection between these 10 most frequent words for a topic.

Topics of Interest “ToI” related to SQL Injection and their most relevant terms.

ToI	Key Terms
Sql injection	Sql, injection, hacking, attack, lesson, mirror, htc, ddos, theme, id
The Structured Query Language injection (SQLI)	Php, crackingfire, Ramadan, analyzer, Brazilian, sqli, com, expiration, army, asp'
SQLi Dumper	Sqli, dumper, MediaFire, pass, hacknho, link, hq, com, private, cyberakline'

Examples of ToI and their most significant terms

ToI	Key Terms
Google dork	Inurl, php, upload, dork, members, usd, com, login, id, proxy'
List of proxies and spoilertarget	Proxies, anonymous, com, http, shell, spoiler, spoilertarget, php, list, www
Bots	Apk, nm, kent, tab, wow, bots, tried, script, wp
Silent Exploit FUD	Com, cyber, https, http, renew, Brazilian, army, fud, exploits
Hackers tactics and websites	Navigon, euq, hackers, vip, nav, com, sites, coins, gmail, igraal
CyberGhost VPN crack	Vpn, http, com, cyberghost, crack, id, php, www, premium, hidden"
Hidden contents	Hidden, file, rar, content, http, net, click, download, data, use

A summary of findings

- We have several incidents of False Positive (FP) terms, where our model suggests they are key terms to the ToI and they were not (either discovered manually or through the prediction algorithms).

- This can be caused by several factors such as:
  - Accuracy in most topic analysis cases is size-dependent where larger datasets from the subject can show more insights and hence better results.
  - Given the large content of the website, our dataset is an early small attempt that needs more collection, analysis and tuning.

A summary of findings

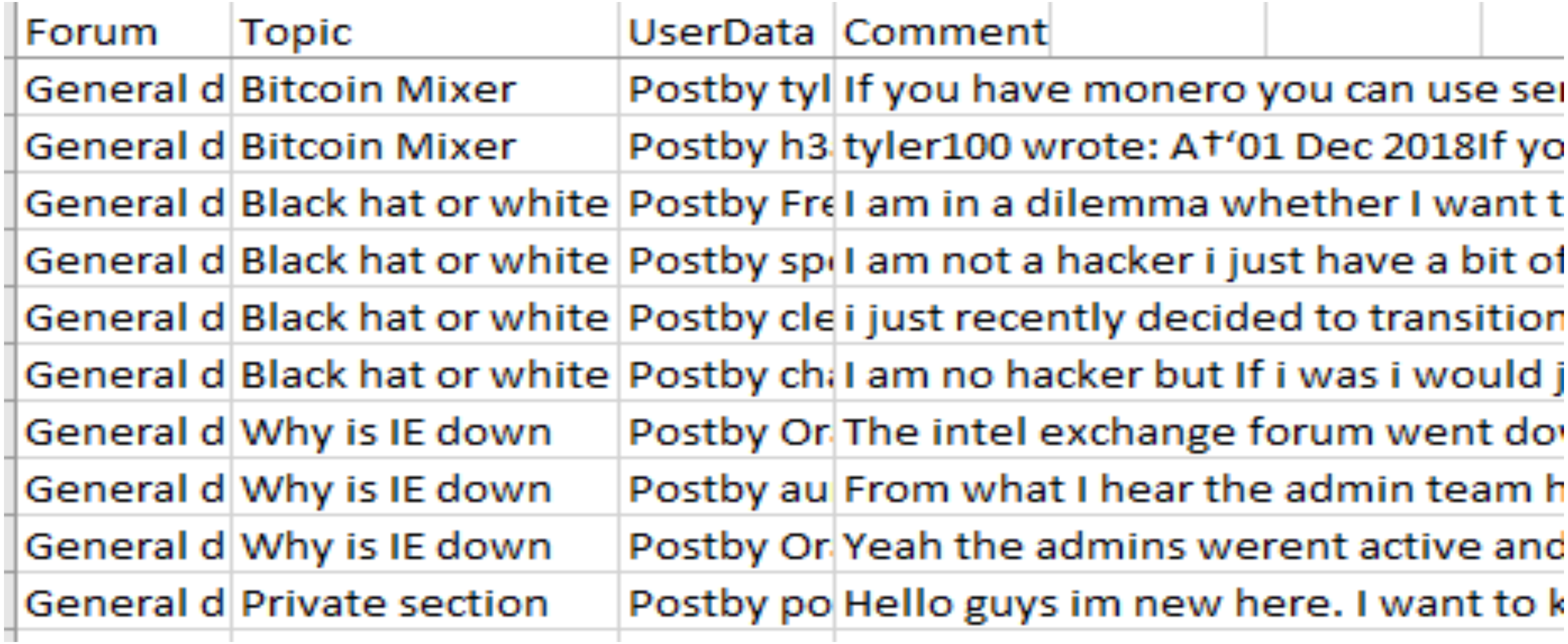
- Attacks usually have contexts and tools that judge content purely based on individual words may miss such context. This may make some irrelevant words, relevant or vice versa.
- Users on OSNs in general and hackers in particular use slang and unstructured language and terms part of their discussions.
- Hackers may also use “hacking conventions” as their own way of encrypting their discussions.
- Spams exist in hacking forums and can impact the accuracy of ToI extractions, without implementing proper spam reduction or elimination methods.

2. Extracting ToI from Torum

- In the second experiment, we selected a website that has its own hacking/malware related categories.
- This can help us compare the algorithmic classification of posts in the different categories with those predefined by the website.
- The data we extracted from Torum included the following main columns (see Figure 2 for a sample of data extracted):

1. The forum and topic under which the post is located.
2. Metadata about the user, who wrote the post, date/time of the post, etc.
3. The content of the post

Extracted data from Torum, a sample



Forum	Topic	UserData	Comment
General d	Bitcoin Mixer	Postby tyl	If you have monero you can use se
General d	Bitcoin Mixer	Postby h3	tyler100 wrote: A†‘01 Dec 2018If yo
General d	Black hat or white	Postby Fre	I am in a dilemma whether I want t
General d	Black hat or white	Postby sp	I am not a hacker i just have a bit o
General d	Black hat or white	Postby cle	i just recently decided to transition
General d	Black hat or white	Postby ch	I am no hacker but If i was i would j
General d	Why is IE down	Postby Or	The intel exchange forum went do
General d	Why is IE down	Postby au	From what I hear the admin team h
General d	Why is IE down	Postby Or	Yeah the admins werent active and
General d	Private section	Postby po	Hello guys im new here. I want to k

Conclusion

- In this paper, our goal is to extract relevant cyber intelligence knowledge from hackers' websites.
- Ultimately, such systems can be deployed as agents throughout the Internet and act as early warning agents for possible security attacks or malwares.
- Using a publicly available dataset collected from CrackingFire website, we utilized topic mining to analyze the contents of the posts and identify "Topics of Interest, ToI".
- The extracted ToI represent known attacks or popular attack tools.

Conclusion

- In the second experiment, we built our own Darkweb crawler.
- We tested the crawler using Torum website as it has predefined malware/hacking related categories.
- We observed several challenges related to the process of crawling from the dark web as well as extracting relevant ToI.
- Our next call is to integrate such ToI as a new category of IoC that can be eventually fed as actionable IOCs to security controls and systems.

Thank you