DISCOVER: Mining Online Chatter for Emerging Cyber Threats

Anna Sapienza*
University of Southern California
Information Sciences Institute
annas@isi.edu

Sindhu Kiranmai Ernala
Georgia Institute of Technology
sernala3@gatech.edu

Alessandro Bessi
University of Southern California
Information Sciences Institute
bessi@isi.edu

Kristina Lerman
University of Southern California
Information Sciences Institute
lerman@isi.edu

Emilio Ferrara
University of Southern California
Information Sciences Institute
ferrarae@isi.edu

ABSTRACT

Widespread adoption of networking technologies has brought about tremendous economic and social growth, but also exposed individuals and organization to new threats from malicious cyber actors. Recent attacks by WannaCry and NotPetya ransomware cryptoworms, infected hundreds of thousands of computer systems world wide, compromising data and critical infrastructure. In order to limit their impact, it is, therefore, critical to detect—and even predict—cyber attacks before they spread. Here, we introduce DISCOVER, an early cyber threat warning system, that mines online chatter from cyber actors on social media, security blogs, and darkweb forums, to identify words that signal potential cyber attacks. We evaluate DISCOVER and find that it can identify terms related to emerging cyber threats with precision above 80%. DISCOVER also generates a time line of related online discussions on different Web sources that can be useful for analyzing emerging cyber threats.

CCS CONCEPTS

• **Security and privacy → Malware and its mitigation**; **Software and application security**; *Intrusion detection systems*; *Vulnerability management*;

KEYWORDS

Web mining; cyber security; cyber threat prediction

ACM Reference Format:
Anna Sapienza, Sindhu Kiranmai Ernala, Alessandro Bessi, Kristina Lerman, and Emilio Ferrara. 2018. DISCOVER: Mining Online Chatter for Emerging Cyber Threats. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3191528

1 INTRODUCTION

The world has become increasingly interconnected, with individuals and organizations linked by networks that people use daily to socialize, receive information and education, buy and sell products and services, manage finances, find work, etc. While this global network brings a wealth of advantages, it also exposes its people to new threats [16] from cyber-attackers who can compromise and misuse their computer systems and data. Cyber attacks are growing in number: in 2016, more than 4000 cyber attacks have occurred daily.[1] Cyber attacks are also growing in diversity, with new phishing attacks, breaches of personal data, malware, trojans, botnets, etc. cropping up daily. The impact from cyber attacks on society is also growing. Recently, for example, individuals as well as organizations all over the world fell victim to *WannaCry* ransomware attack, which targeted computers running Microsoft Windows operating systems. The ransomware encrypted victim's files, demanding ransom payment in return for decryption key. In a similar ransomware campaign, *NotPetya* infected many organizations worldwide in June 2017. To mitigate the risk from cyber attacks and reduce their damage, we need new methods capable of predicting an attack [21], or at least detecting it in its early stages.

The growth of cyber threat has increased the likelihood that signals of impending attacks will be visible in the open public data sources [5]. Cyber attackers exploit vulnerabilities using tools, techniques, and tradecraft. Therefore, to conduct an attack, malicious actors typically have to 1) identify vulnerabilities, 2) acquire the necessary expertise and tools to use them, 3) choose targets, 4) recruit participants, and 5) plan and execute the attack. Other actors—system administrators, security analysts, and even victims—may discuss vulnerabilities, threats, or coordinate defenses against exploits. These discussions are often conducted in online forums, including blogs and social media, thereby creating potential signals to identify an upcoming attack or a new cyber vulnerability [24]. Existing approaches focus on using single Web source as signal for predicting vulnerabilities or exploits [19, 20]. In this paper, we introduce DISCOVER, a method that leverages multiple online data sources as signals to generate warnings indicative of new potential cyber threats, which in the present paper is defined as an unusual word, which could be either related to a cyber attack or be the actual name of the cyber threat (e.g., name of a malware, trojan, exploit, etc.).

DISCOVER monitors, in real time, multiple channels of online chatter related to cyber security, including blogs of cyber security experts and "white hat hackers," as well as social media posts, and checks for the co-occurrences of terms to uncover threats in the discussions of malicious actors on the Dark Web forums and

*A. Sapienza, S.K. Ernala, and A. Bessi contributed equally to this work.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191528

[1] https://www.justice.gov/criminal-ccips/file/872771/download

marketplaces. DISCOVER processes the data from these sources by employing data mining techniques to identify novel terms related to a potential cyber threat, which it returns as a warning. Furthermore, the framework uses signals from multiple data sources to create a time line of discussions of the threat. The threats discovered by the system could alert security experts in a timely manner to take precautionary steps. Such an early warning generation system could help organizations and victims prepare and limit their vulnerability to cyber attacks.

The rest of the paper is organized as follows: in Sec. 2 we describe the data sources used as an input by the algorithm and how they are preprocessed; in Sec. 3, we introduce the DISCOVER framework (which is an extension of the model we presented in [25]) including details on data retrieval infrastructure and warnings generation. We then evaluate the method and present the experimental results in Sec.4, by analyzing several case studies. We review the existing literature and analyze the problem of detecting and predicting cyber threats from online data sources in Sec. 5,. Finally, we conclude in Sec. 6 with a discussion on the uses and implications of the framework and future work in this problem space.

2 DATA PROCESSING

Our two primary data sources for warning generation are social media (Twitter) and blogs of cyber-security experts. We also use data collected from darkweb to find mentions of warnings generated by DISCOVER to create timelines of warnings.

2.1 Data Collection

Social media. Twitter is a popular micro-blogging, social media platform where users post short messages ("tweets"), restricted to 140 characters. We compiled a list of recognized cyber-security experts who post frequently on Twitter about cyber-security issues. This manually curated list includes 69 international researchers and security analysts associated with security firms, as well as widely-followed white hat hackers. We collect tweets posted by these experts on their timeline on an hourly basis. We use the official Twitter API to collect data in real time and store it in an Amazon EC2 instance. This data is then retrieved by DISCOVER using Elastic Search, an open source search engine based on Apache Lucene that provides a distributed, multitenant-capable full-text search with a schema-free JSON documents. Each data point has fields including the author of the tweet, their profile information, location and timestamp of the tweet etc.

Cyber security blogs. The top blogs written and curated by cyber-security experts and white hat hackers form the complementary data source for DISCOVER. These blogs have rich technical information on the latest exploits, software vulnerabilities, popular ransomware, malware and other topics in cyber-security. We begin with a manually-curated list of 290 security blogs. We then crawl the blogs and extract data from them in a unified RDBMS schema (using MongoDB backend). Finally, the algorithm retrieves the related data through the Elastic Search API. Each data entry is characterized by different fields. Here for each post in blogs we focus on: *DatePublished*, the date on which the post was published, its *URL*, and *text*, providing the actual contents of the post.

Darkweb forums. Deepweb refers to unindexed and anonymous sites on the internet. The part of Deepweb that is not accessible through standard browsers or search engines, but only via anonymization protocols such as Tor and i2p is termed as the Darkweb. To crawl the data from discussion forums on the darkweb, we adopt the methods used in [19, 23]. To extract cyber security related data from the darkweb, we started with a manually-compiled list of 263 sites that are forums or marketplaces relating to malicious hacking and/or online financial fraud, including fishing, spear-fishing, data breaches, ransomware etc. Each site is crawled three times per week. The diversity of the sites in the manually compiled list necessitates custom crawlers, instead of common crawling methods based on the protocol and site structure. Analogously to the blogs, once data is crawled and parsed from several sites, it is stored in a unified RDBMS schema (MongoDB backend) to simplify data cleaning process. This also enables us in identifying only cyber-security relevant information from the crawled data, since many forums and marketplaces on the darkweb are known to be involved in other illicit activities such as drug markets and the sale of stolen goods. Finally, the data is retrieved by DISCOVER for warning generation using the Elastic Search API. Each data point is a long form text post containing metadata such as publication date, authors' usernames, authors' reputations, etc. Here, we query the database to monitor mentions about specific warnings that DISCOVER generates.

2.2 Data Preprocessing

These three data sources are very different in nature, each providing a unique type of signal. Content from Twitter and cyber-security blogs is cleaner compared to darkweb forums. Since the former is written by security experts, it is highly topical and rich in technical jargon. The latter, however, is a collection of information from darkweb sources on diverse topics. These posts also include code snippets, tutorials on exploits/vulnerabilities, data dumps of personal information such as email addresses, passwords, etc., among non-cyber topics, such as drug trade. The writing style within the darkweb forums is often intentionally difficult to parse, with words concatenated into new terms and multiple languages used within a single post [19].

Based on the exploratory analysis, we designed DISCOVER to take as input data related to Twitter and cyber-security blogs, while also monitoring the mentions of new potential threats on the darkweb. We apply a two step filtering and data pre-processing procedure on the primary sources Twitter and blogs. The filtering step eliminates terms within text that are not written in English. After filtering, we pre-process the data by removing URLs, symbols, numbers etc., and tokenize the text to obtain a unique list of terms.

3 THE DISCOVER FRAMEWORK

In this section, we present a detailed description of the DISCOVER framework, depicted in Fig. 1. This is divided in two main parts: the *text mining infrastructure*, used to parse the discussion in the different sources, and the *warning generation methodology* in which novel terms are detected as potential cyber threats.

3.1 Text Mining

The data pre-processing stage results in a large list of words that might not be relevant to cyber threats at all. To "discover" novel terms potentially indicative of cyber threats, we filter out "known" terms using a four stage filtering process. At each stage, we exclude terms by filtering them out if they occur in any of the following dictionaries:

(1) **English dictionary** - 236,736 commonly used English terms based on the NLTK English corpus are used to build this dictionary. Terms such as interview, hello, because are removed as they do not represent potential cyber threats.
(2) **Stopwords dictionary** - 3136 stopwords e.g. to, on, a, for, ... that form this dictionary are removed;
(3) **Domain vocabulary** - Domain vocabulary such as technical terms and context-specific terms form the body of the chosen data sources. They are however descriptive in nature, and hence do not represent a potential warning word for cyber threats. Similarly, each data source has a temporally accumulated form and style of writing. To exclude such domain specific lexicon, we build this dictionary based on the past data for each of the data sources. Based on the chosen warning generation period for the experiments, we use data from each source from January 2013 to August 2016. After pre-processing, we tokenize this data to build the domain vocabulary.
(4) **Threat dictionary** - 25 general terms indicating known types of cyber-threats e.g. ddos, phishing, data breach, botnet, etc. for a significant portion of the data; We manually curated this list of words. These words are excluded in the filtering process (but used in the next stages) as they do not stand by themselves as a new cyber threat warning.
(5) **Italian dictionary** - 129, 121 common Italian words, e.g. intervista, attacco, spazio, etc. form this dictionary. We use this dictionary only for the Twitter data, since some of the cyber-security experts tweet in Italian. All of the blogs dataset is primarily written in English. Other non-English dictionaries can be included upon finding their usage among the experts in the dataset.

Using English dictionary, stopwords dictionary, we filter out common words that are unlikely to be related to cyber-threats; whereas by means of the technical dictionary we remove several context-specific words that have been used in the past by the users of the individual data sources that we are monitoring. Note that the threat dictionary can be enlarged to incorporate new terms as they enter cyber-security vernacular.

3.2 Warning Generation

In the final step before warning generation, we impose some constraints to check the words that pass filtering process. Given the viral nature of online chatter, we do not want to generate warnings simply based on words that were previously not seen. Such words could represent misspellings of known words or idiosyncratic names. Hence, we need to exclude terms that have unique occurrence: we exclude words that occur only once in all posts during the given time period ($count > 1$).

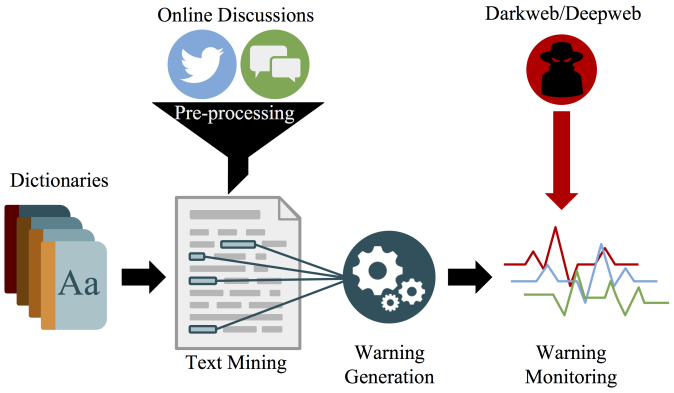

Figure 1: DISCOVER framework, from data pre-processing to warning generation and monitoring.

Additionally, we want to ensure that the detected term is related to a cyber-security topic. To ensure this, we require that the term co-occurs with a term from our threat dictionary, which we call *context*.

Any novel term that meet these requirements ($count > 1$ and $n.contextwords > 0$) from both data sources will be a warning generated by the DISCOVER framework. The warning generation occurs at an hourly rate from the Twitter data source and at a daily rate for cyber-security blogs. Each warning is in the following format.

- The time period (day, hour) during which DISCOVER has generated the term as a warning
- The discovered warning term that is likely to be related to a current of future cyber attack
- The data source that generated the term as a warning
- The frequency of the warning term in the given time period
- The list of associated threat words that co-occur as context for the discovered term

4 RESULTS

4.1 Method Evaluation

To evaluate our framework, we let DISCOVER generate warnings for online chatter over a time period from September 1st, 2016 to January 31st, 2017. We have ground truth data from this time period that was generated by earlier implementations of DISCOVER. This ground truth set consists of 661 warnings generated from Twitter and annotated by five experts, and 103 warnings generated from the blogs data annotated by three experts. The annotators were asked to independently evaluate each warning and mark it as a *true cyber threat* or *false flag* (not a cyber threat). In particular, a word is defined as a true cyber threat if it is related to an actual attack that occurred in the selected time period. To identify whether the attack occurred before, during or after the warning occurred, annotators were asked to leverage Google search for "investigative" purposes. Moreover, a discovered word was marked as a true cyber threat if the majority of the annotators agreed in their evaluations (i.e., at least 3 out of 5 annotators for Twitter warnings, and 2 out of 3

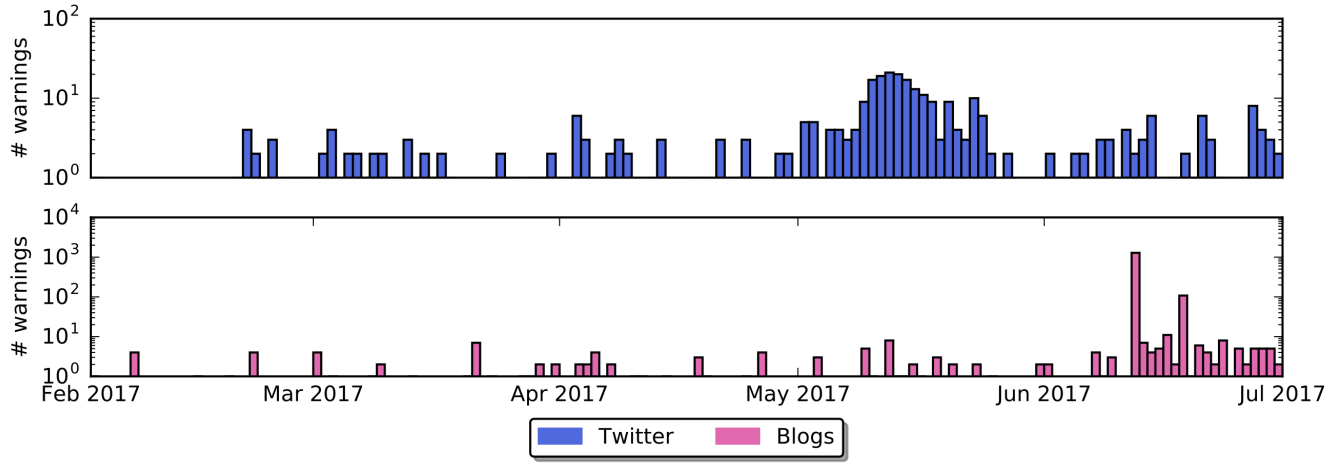

Figure 2: Daily count of warnings generated by DISCOVER from Twitter and blogs.

annotators for blogs). The two sets of annotations combined allow us to evaluated our framework.

In Tab. 1, we reported the evaluations of the generated warnings separately for the two data sources and the final precision of DISCOVER, given by the combination of these results.

Table 1: DISCOVER precision on the different data sources and on the combined data.

Data Source	Num. warnings	Precision
Twitter	661	84%
Blogs	103	59%
Twitter + Blogs	764	81%

As shown in Tab. 1, the 84% of the warnings coming from the Twitter data are true cyber threats, while the 59% of the warnings deriving from blogs data are related to real cyber threat. This lower precision could be improved by changing the algorithm constraints (count and context) on the different data sources. As an example, blogs entries are in general longer texts than Twitter entries, and as we are monitoring cyber-security blogs, they also contain on average more context words than Twitter data. Thus, the majority of the warnings generated by the blogs data source are characterized by more than one context word. By increasing the constraint we have on the context then, such as requiring the presence of 2 or more context words, we could discard some of the generated warnings and increasing the precision on the data source.

However, the overall precision reached is high, i.e., 81%, and as we will discuss in the following section, the use of blogs data as an additional source allows DISCOVER to detect in advance some of the highest-impact recent cyber attacks. Based on these observations, we decided to keep the parameters (word count and context) the same for both Twitter and blogs, thus balancing efficiency and generality in DISCOVER.

4.2 Scenario Analysis

To test the framework in identifying warnings relevant to imminent cyber-threats, we run DISCOVER on data collected from February 2017 to June 2017. DISCOVER generated 344 warnings from Twitter and 1565 warnings from blogs during this time period. The daily number of warnings generated by both the data sources is shown in Fig. 2.

The top warnings, along with their type, the time at which DISCOVER generated the warning, and the source that first produced it, are reported in Table 2. We identify warnings related to a variety of cyber attacks during this time period, including malware, ransomware, data breaches, botnets and other exploits. There were ten warnings that were generated by both the data sources: ‘medoc‘, ‘industroyer’, ‘nayana’, ‘notpetya’, ‘kasperagent’, ‘wannacry’, ‘crashoverride’, ‘dahua’, ‘wannacrypt’, ‘macspy’. Among these, ‘industroyer’, ‘crashoverride’, ‘dahua’, ‘macspy’ were first generated by Twitter and the remaining were first identified as warnings by blogs. This shows the first advantage of leveraging multiple data sources for warning generation.

The second advantage of leveraging multiple data sources for warning generation is that we are able to provide a cyber monitoring platform where, after the first time a new warning regarding a threat has been generated, we can monitor for the warning term in the remaining data sources. This provides a temporal landscape of the evolution of discussions regarding a cyber threat among the data sources. In this regard, we use warnings generated from the primary data sources, Twitter and blogs. We utilize darkweb as a secondary data source to monitor warnings. Of several threats during this time period, we elaborate on the temporal landscape of three types of attacks—ransomware, exploit and data breach.

4.2.1 Ransomware. Wannacry — On April 18, 2017 DISCOVER generated a warning for a new term, ‘wannacry’ from the blogs data source. Although there were mentions of this term before the day of first warning, the term did not pass the constraints imposed in terms of the $count > 1$ and $context > 0$. This means that either the number of mentions of the term was equal to one or that there was

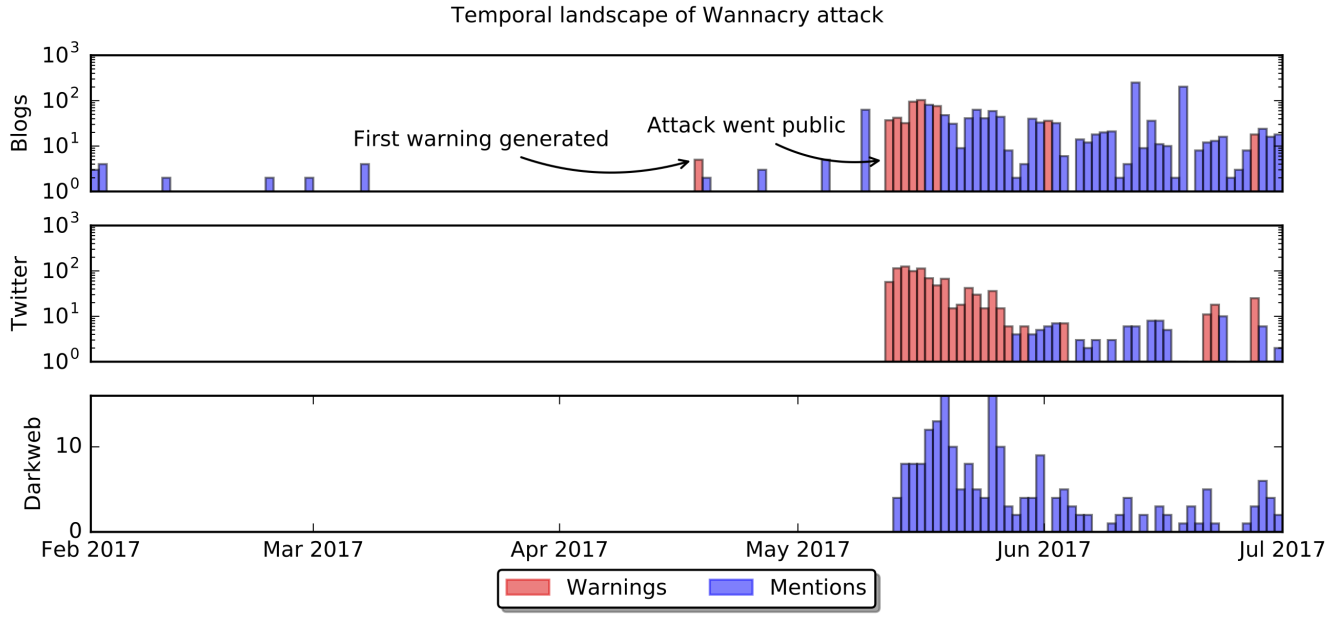

Figure 3: Temporal landscape of warnings and mentions related to the Wannacry attack

Table 2: Top warnings generated from February - June 2017

Discovered term	Warning generation date	Source	Type of threat
cloudpets	2017-02-27	Twitter	data breach
coachella	2017-03-01	Twitter	data breach
stonedrill	2017-03-06	Twitter	malware
petrwrap	2017-03-15	Twitter	ransomware
incapta	2017-03-24	Twitter	botnet
eternalblue	2017-05-12	Twittter	exploit
wannacry	2017-04-18	Blogs	ransomware
notpetya	2017-02-01	Blogs	ransomware
maarten	2017-04-03	Blogs	malware
pwnwiki	2017-06-12	Blogs	malware
lightbulb	2017-06-25	Blogs	iot, ddos
ghosthook	2017-06-23	Blogs	exploit

no overlapping between the text and the threats dictionary we use. From this time onward, apart from a couple of mentions on blogs, the same warning is re-generated on both Twitter and blogs again on the 12th of May, 2017. On this day, the Wannacry ransomware became a worldwide cyber attack targeting computers running on Microsoft Windows Operating system. The Wannacry cryptoworm attacked Microsoft systems by encrypting data on the systems and demanding ransom payments in the form of Bitcoin cryptocurrency. From 12th May onward, there were recurrent warnings generated by DISCOVER for the term 'wannacry'. On the same day, we also observed a warning for the term 'eternalblue' from Twitter data. Later, Eternal Blue was discovered to be an exploit leaked by the Shadow Brokers hacker group on April 14, 2017, and was used as part of the Wannacry ransomware attack. Alongside 'wannacry' and 'Eternalblue', there were warnings generated for terms such as 'wannacrpyt', 'wcry', 'wanacry' which are lexical variations of the original term. This presents an interesting evidence of lexical variations used as a means of discussing an imminent cyber threat in online spaces, to circumvent the usage of the original term.

The case of Wannacry also illustrates the significance of leveraging multiple data sources for the task of warning generation. Despite having a lower precision for generating valid warnings, when compared to Twitter data, the blogs data source provides a unique sensor to capture long form discussions and news on cyber attacks and vulnerabilities, before the content got popularized. Similarly, after 12th May, we observe a rapid increase in mentions of 'wannacry' on the darkweb forums and marketplaces. As a supplementary source, darkweb data provides an automatic verification step on whether the warning word is a one-time occurrence / new

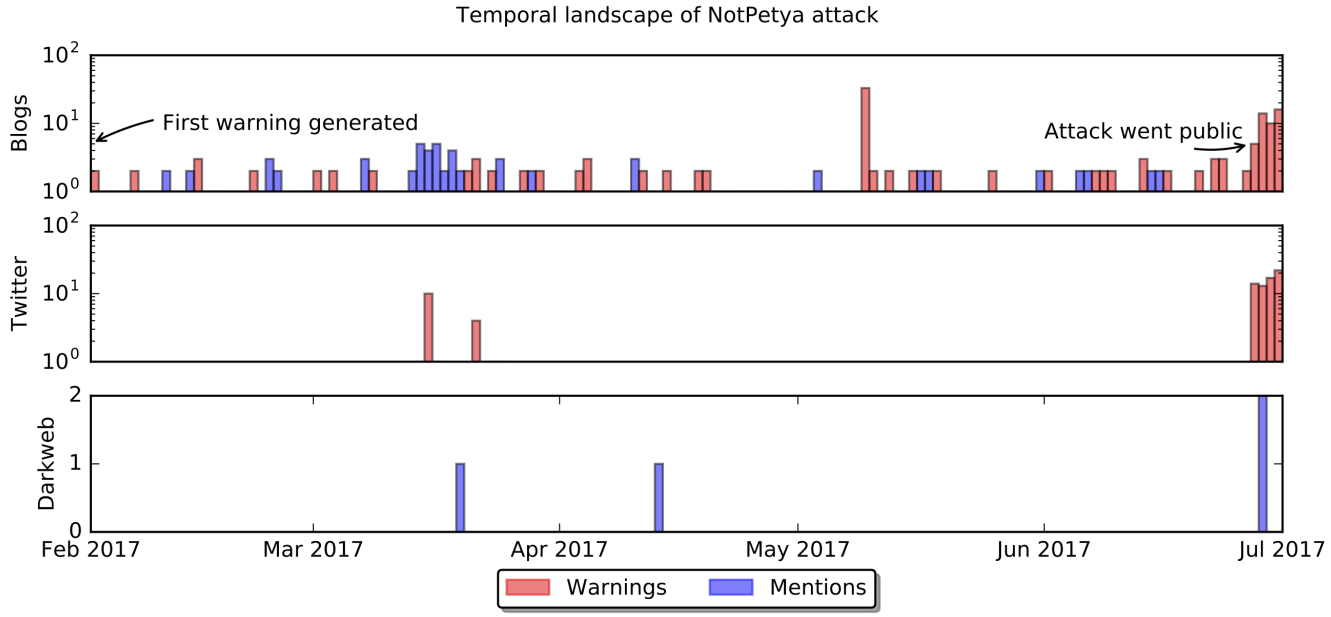

Figure 4: Temporal landscape of warnings and mentions related to the NotPetya malware attack

vocabulary or has on-going mentions on other data sources.

NotPetya — News regarding the Petya malware that swamped websites of Ukrainian organizations, including banks, ministries, newspapers and electricity firms hit popular media on 27 June 2017. The cyber attack affected multiple nations including France, Germany, Italy, Poland, Russia, United Kingdom, the United States and Australia. Similar to Wannacry, Petya used the EternalBlue exploit previously discovered in older versions of the Microsoft Windows operating system. The malware encrypted files on the system and demanded US$300 in bitcoin to receive instructions to decrypt their computer. At the same time, the malware also exploited the Server Message Block protocol in Windows to infect local computers on the same network, and any remote computers it can find. The attack got popularly named as NotPetya, to distinguish the version used in the Ukraine cyberattacks which was a modified version of the original Petya malware.

DISCOVER generated a warning for NotPetya almost a month before the attack went public. The first warning was generated on Feb 1st 2017, followed by couple of mentions and warnings until early March 2017. While monitoring for NotPetya on the other data sources, we observed no activity until mid March (15 March 2017), when data from Twitter generated a warning for NotPetya. This was followed by another warning from Twitter on 21 March 2017. Additionally, during the same period of increased activity around the usage of NotPetya, darkweb showed similar signals with increase in mentions for the term. Finally, after over 2 months, the attack was public on 27 June 2017 when both the primary data sources generate multiple warnings for the term. Such a temporal landscape with recurrence in discussions on an existing malware like NotPetya, becomes a strong indicator of an imminent cyber threat. It can be highly beneficial for security experts to use such a monitoring tool. Finally, similar to 'wannacry', the warning term 'NotPetya' had multiple lexical variations such as 'petrwrap' and 'petyawrap'.

4.2.2 Malware. Kasperagent — Kasperagent is a Microsoft Windows malware targeting users in the United States, Israel, Palestinian Territories, and Egypt since July 2015. It was discovered by Palo Alto Networks Unit 42 and ClearSky Cyber Security, and publicized in April 2017 in the targeted attacks in the middle east leveraging decoy Palestinian Authority documents. The threat actors used shortened URLs in spear phishing messages and fake news websites to direct targets to download the malware. These malware samples then dropped various decoy documents associated with the Palestinian Authority, the governing body of the emerging Palestinian autonomous regions of the West Bank and Gaza Strip. DISCOVER generated the first warning for Kasperagent from the blogs data source on 12 June 2017, before it got popularized around the 14th June, 2017. This demonstrates the possibility of using DISCOVER as an early warning generation tool yielding actionable insights to analysts and decision makers.

4.2.3 Exploit. Ghosthook — During the week of June 22 to June 27, 2017 security researchers discovered the Ghosthook attack technique, which uses features of the Intel CPUs (central processing units) to take over 64-bit Windows systems. It was reported that "Windows has traditionally been safe from most cybercriminals trying to install rootkits, but the GhostHook attack can bypass PatchGuard, which was specifically developed to protect its operating system at the kernel level" [2]. Although hooking rootkits

[2]https://securityintelligence.com/news/ghosthook-attack-reveals-kernel-level-threat-in-64-bit-windows-systems/

is not always used for malicious purposes, researchers note that hackers would require a malware present on the system to exploit a rootkit. During this time period, DISCOVER generated a warning for Ghosthook early on the 23rd June 2017. The warning was generated by the blogs data source with associated context as "exploit, rootkit, malware". The warning was only generated by blogs and not Twitter. This demonstrates the novelty of each individual data source in generating a particular kind of warnings that might not be found in the others. Early identification of such rootkits and malwares could be highly beneficial for companies in mitigating and fixing the threat.

4.2.4 Data breach. Cloudpets — On Feb 28 2017, news broke that personal information of more than half a million people who bought internet-connected teddy bear toys from Cloudpets has been compromised. The leaked information included email addresses, passwords as well as profile pictures and more than 2 million voice recordings of children and adults who had used the CloudPets stuffed toys. The companyâĂŹs toys could connect over Bluetooth to an app, allowing parents to upload or download audio messages for their child. According to online news sources, the parent company Spiral Toys left customer data of its CloudPets brand on a database that wasn't password-protected. "In fact, at the beginning of January, during the time several cybercriminals were actively scanning the internet for exposed MongoDB's databases to delete their data and hold it for ransom, CloudPets' data was overwritten twice, according to researchers" [3] . DISCOVER generated the first warning for Cloudpets on Feb 27 2017 with 2 mentions from the experts feed on Twitter data source. Using contextual information DISCOVER related the warning to the threat words - accounts, breach. Subsequently, multiple warnings were generated from the Twitter data stream on the 28th Feb 2017, until the 1st of March, 2017. Cloudpets breach is an example of warning that was generated only by a single data source in the DISCOVER framework. There were no mentions of cloudpets in the blogs sources. This demonstrates the unique nature of each data source; reporting nature of Twitter rather than long-form description, counter-measures against cyber threats as on blogs.

5 RELATED WORK

In this paper, we leverage signals from multiple online data sources such as the activity of cyber-security experts on social media (Twitter) and blogs towards building an early warning generation system for cyber threats. Prior work has explored these data sources with similar motivations.

The activity of hacker groups on darkweb forums has been identified as a rich data source in detecting threats posing risk to individuals, corporates, and the government. Previous research has studied the landscape of this online space in terms of the individual participants and information disseminated. They found that individuals on these forums advertise tools such as malware samples, source codes and also sell on open black markets operating on-line [1, 14, 26]. Information on such cyber vulnerabilities is disseminated among the hacker community commonly in the form of tutorials (both text and video), directly enabling readers to launch criminal cyber attacks such as denial of service, SQL injections etc. [5]. Alongside the advertising of vulnerabilities, stolen personal data such as credit card information [4], accounts information such as during the Ashley Madison hack are put up on sale on these forums. Research has also studied individuals on these forums from a demographic, sociological perspective, as a hacker community [10, 15]. The presence of such communities was identified to be common across several geo-political regions where information technologies are either ubiquitous or rapidly growing, including the US, China, Russia, the Middle-East etc. [4, 18].

In the recent past, social media (such as Twitter) has also emerged as a rich data source for variety of prediction tasks ranging from stock market [7], elections [29], epidemiology [2, 8], health and well-being [9] etc. Specifically, in the domain of cyber-security previous work has focused on the study of manipulation and abuse [11], detection and effects of spam [3, 30], social bots [13, 27], malicious campaigns [6, 12, 22, 28], etc., on Twitter. However, there is an untapped wealth of information based on the activity of security experts and white hat hacker groups on Twitter as well as grievances and complains on softwares by regular users. In this direction, most recent work by Sabottke et al. [24] used Twitter for identification of cyber vulnerabilities. In this paper, we leverage the experts' activity on Twitter as a novel signal for cyber threat warning generation.

Finally, alongside the rich data sources, computational methods for the identification and prediction of cyber threats has been explored. Okutan et al. used Bayesian networks to predict cyber attacks using unconventional signals from Twitter, the GDELT project and cyber-security blogs [20]. Similarly, towards the task of forecasting zero-day vulnerability discovery rates, David Last presented ongoing research on Vulnerability Discovery Models [17] for both global and software specific categories for example, Browser, Operating system, Video vulnerabilities.

Despite the rich body of work utilizing unconventional data sources for threat detection, they have only been analyzed as individual signals. In this paper, we provide a robust framework for cyber-threat warning generation using multiple data sources to extract unique knowledge from each data source as well as a temporal landscape of warnings prior to a cyber-attack.

6 CONCLUSION

In this paper, we presented DISCOVER, an early warning generation algorithm, whose aim is to predict cyber threats by mining online discussions.

Our framework takes as an input unconventional and public data sources related to cyber security topics. Here, we focus on the analysis of two main data sources: Twitter accounts of cyber security experts, and cyber security related blogs. The system monitors tweets and blog posts published online daily and, by mining their text, detects unusual words that can be related to a cyber threat. Then, it produces alerts for each of the discovered words, along with a context that helps to identify the type of cyber threat, e.g., ransomware, malware, phishing attack, data breach etc. Finally, it looks for mentions of the generated warning on the darkweb. This last step combined with the previous ones allows DISCOVER

[3] https://motherboard.vice.com/en_us/article/pgwean/internet-of-things-teddy-bear-leaked-2-million-parent-and-kids-message-recordings

[4] https://www.theguardian.com/technology/2015/oct/30/stolen-credit-card-details-available-1-pound-each-online

to build a temporal landscape of online discussions related to the specific warning.

We evaluated the method over the period going from Sept 1, 2016 to Jan 31, 2017, for which we have a ground truth for the warnings generated from Twitter posts. Moreover, for the same time period, we asked cyber security experts to evaluate the outcome of DISCOVER on blogs data. The evaluation shows that DISCOVER reaches a warning average precision above 81%, respectively of 84% for Twitter warnings and above 59% for blogs.

Despite the lower precision derived from blogs in combination with Twitter, we proved how this additional data source has a key role in the warning generation procedure. Running DISCOVER on both the sources indeed allowed to detect in advance two major cyber attacks: *Wannacry* and *NotPetya*.

We could tune the constraints of the algorithm to better fit the different data sources. However, by increasing the thresholds some of the true cyber threats that we found could be discarded. Moreover, in the present version of DISCOVER we reach a good balance between having a general algorithm that can be run on several data sources, a high precision, and a wide spectrum of detected cyber threats.

Future work will be devoted to enhance DISCOVER by extending the list of cyber security experts upon which we rely when monitoring online discussions. One possible direction would be to use Natural Language Processing (NLP) techniques as well as topic analysis to automatically detect cyber security related forums, blogs, and Twitter authors. Moreover, these techniques could help in the extraction of further details from darkweb forums. We plan to detect contextual information about a warning, such as the source of the attack (hackers) and the targets, and try to identify when the attack will occur.

Other directions include extending DISCOVER to identify lexical variations related to the same cyber threat, and to generate dynamic warnings. We could adapt the framework to keep track of a certain word after its generation, and make the relevance of warnings decay if that word is not mentioned for a certain time period.

7 ACKNOWLEDGEMENTS

This work was supported by the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C-0112. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, or the U.S. Government.

REFERENCES

[1] Luca Allodi. 2017. Economic Factors of Vulnerability Trade and Exploitation. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1483–1499.

[2] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1568–1576.

[3] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. 2010. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, Vol. 6. 12.

[4] Victor Benjamin and Hsinchun Chen. 2014. Time-to-event modeling for predicting hacker IRC community participant trajectory. In *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*. IEEE, 25–32.

[5] Victor Benjamin, Weifeng Li, Thomas Holt, and Hsinchun Chen. 2015. Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*. IEEE, 85–90.

[6] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. (2016).

[7] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.

[8] David A Broniatowski, Michael J Paul, and Mark Dredze. 2013. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one* 8, 12 (2013), e83672.

[9] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. *ICWSM* 13 (2013), 1–10.

[10] Hanno Fallmann, Gilbert Wondracek, and Christian Platzer. 2010. Covertly Probing Underground Economy Marketplaces.. In *DIMVA*, Vol. 10. Springer, 101–110.

[11] Emilio Ferrara. 2015. Manipulation and abuse on social media. *ACM SIGWEB Newsletter* Spring (2015), 4.

[12] Emilio Ferrara. 2017. Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. *First Monday* 22, 8 (2017).

[13] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[14] Thomas J Holt and Eric Lampke. 2010. Exploring stolen data markets online: products and market forces. *Criminal Justice Studies* 23, 1 (2010), 33–50.

[15] Tim Jordan and Paul Taylor. 1998. A sociology of hackers. *The Sociological Review* 46, 4 (1998), 757–780.

[16] Vipin Kumar, Jaideep Srivastava, and Aleksandar Lazarevic. 2006. *Managing cyber threats: issues, approaches, and challenges*. Vol. 5. Springer Science & Business Media.

[17] David Last. 2016. Forecasting Zero-Day Vulnerabilities. In *Proceedings of the 11th Annual Cyber and Information Security Research Conference*. ACM, 13.

[18] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M Voelker. 2011. An analysis of underground forums. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 71–80.

[19] Eric Nunes, Ahmad Diab, Andrew Gunn, Ericsson Marin, Vineet Mishra, Vivin Paliath, John Robertson, Jana Shakarian, Amanda Thart, and Paulo Shakarian. 2016. Darknet and deepnet mining for proactive cybersecurity threat intelligence. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*. IEEE, 7–12.

[20] Ahmet Okutan, Shanchieh Jay Yang, and Katie McConky. 2017. Predicting cyber attacks with bayesian networks using unconventional signals. In *Proceedings of the 12th Annual Conference on Cyber and Information Security Research*. ACM, 13.

[21] Jamal Raiyn et al. 2014. A survey of cyber attack detection strategies. *International Journal of Security and Its Applications* 8, 1 (2014), 247–256.

[22] Jacob Ratkiewicz, Michael Conover, Mark R Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and Tracking Political Abuse in Social Media. *ICWSM* 11 (2011), 297–304.

[23] John Robertson, Ahmad Diab, Ericsson Marin, Eric Nunes, Vivin Paliath, Jana Shakarian, and Paulo Shakarian. 2017. *Darkweb Cyber Threat Intelligence Mining*. Cambridge University Press.

[24] Carl Sabottke, Octavian Suciu, and Tudor Dumitras. 2015. Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits.. In *USENIX Security Symposium*. 1041–1056.

[25] Anna Sapienza, Alessandro Bessi, Saranya Damodaran, Paulo Shakarian, Kristina Lerman, and Emilio Ferrara. 2017. Early Warnings of Cyber Threats in Online Discussions. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*.

[26] Lance Spitzner. 2003. The honeynet project: Trapping the hackers. *IEEE Security & Privacy* 99, 2 (2003), 15–23.

[27] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. The DARPA Twitter bot challenge. *Computer* 49, 6 (2016), 38–46.

[28] Kurt Thomas, Chris Grier, and Vern Paxson. 2012. Adapting Social Spam Infrastructure for Political Censorship.. In *LEET*.

[29] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2011. Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social science computer review* 29, 4 (2011), 402–418.

[30] Alex Hai Wang. 2010. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*. IEEE, 1–10.