2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

BlackWidow: Monitoring the Dark Web for Cyber Security Information

Matthias Schäfer
Department of Computer Science
University of Kaiserslautern
Kaiserslautern, Germany
schaefer@cs.uni-kl.de

Markus Fuchs
SeRo Systems
Kaiserslautern, Germany
fuchs@sero-systems.de

Martin Strohmeier
Cyber-Defence Campus
armasuisse
Thun, Switzerland
martin.strohmeier@armasuisse.ch

Markus Engel
SeRo Systems
Kaiserslautern, Germany
engel@sero-systems.de

Marc Liechti
Trivo Systems
Bern, Switzerland
marc.liechti@trivo.ch

Vincent Lenders
Cyber-Defence Campus
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Abstract: The Dark Web, a conglomerate of services hidden from search engines and regular users, is used by cyber criminals to offer all kinds of illegal services and goods. Multiple Dark Web offerings are highly relevant for the cyber security domain in anticipating and preventing attacks, such as information about zero-day exploits, stolen datasets with login information, or botnets available for hire.

In this work, we analyze and discuss the challenges related to information gathering in the Dark Web for cyber security intelligence purposes. To facilitate information collection and the analysis of large amounts of unstructured data, we present BlackWidow, a highly automated modular system that monitors Dark Web services and fuses the collected data in a single analytics framework. BlackWidow relies on a Docker-based micro service architecture which permits the combination of both pre-existing and customized machine learning tools. BlackWidow represents all extracted

data and the corresponding relationships extracted from posts in a large knowledge graph, which is made available to its security analyst users for search and interactive visual exploration.

Using BlackWidow, we conduct a study of seven popular services on the Deep and Dark Web across three different languages with almost 100,000 users. Within less than two days of monitoring time, BlackWidow managed to collect years of relevant information in the areas of cyber security and fraud monitoring. We show that BlackWidow can infer relationships between authors and forums and detect trends for cybersecurity-related topics. Finally, we discuss exemplary case studies surrounding leaked data and preparation for malicious activity.

Keywords: *Dark Web analysis, open source intelligence, cyber intelligence*

1. INTRODUCTION

The Dark Web is a conglomerate of services hidden from search engines and regular Internet users. Anecdotally, it seems to the uneducated observer that anything that is illegal to sell (or discuss) is widely available in this corner of the Internet. Several studies have shown that its main content ranges from illegal pornography to drugs and weapons [1], [2]. Further work has revealed that there are many Dark Web offerings which are highly relevant for the cyber security domain. Sensitive information about zero-day exploits, stolen datasets with login information, or botnets available for hire [2], [3] can be used to anticipate, discover, or ideally prevent attacks on a wide range of targets.

It is difficult to truly measure the size and activity of the Dark Web, as many websites are under pressure from law enforcement, service providers, or their competitors. Despite this, several web intelligence services have attempted to map the reachable part of the Dark Web in recent studies. One crawled the home pages of more than 6,600 sites (before any possible login requirement), finding clusters of Bitcoin scams and bank card fraud [4]. Another study found that more than 87% of the sites measured did not link to other sites [5]. This is very different from the open Internet, both conceptually and in spirit: in contrast, we can view the Dark Web as a collection of individual sites or separated islands.

In the present work, we introduce BlackWidow, a technical framework that is able to automatically find information that is useful for cyber intelligence, such as the early

detection of exploits used in the wild, or leaked information. Naturally, analyzing a part of the Internet frequented by individuals who are trying to stay out of the spotlight is a more difficult task than traditional measurement campaigns conducted on the Surface Web.

Thus, a system that seeks to present meaningful information on the Dark Web needs to overcome several technical challenges – a large amount of unstructured and inaccessible data needs to be processed in a scalable way that enables humans to collect useful intelligence quickly and reliably. These challenges range from scalability and efficient use of resources over the acquisition of fitting targets to the processing of different languages, a key capability in a globalized underground marketplace.

Yet, contrary to what is sometimes implied in media reports, few underground forums and marketplaces use a sophisticated trust system to control access outright, although some protect certain parts of their forums, requiring a certain reputation [6]. We successfully exploit this fact to develop an automated system that can gather and process data from these forums and make them available to human users.

In this work, we make the following contributions:

- We present and describe the architecture of BlackWidow, a highly automated modular system that monitors Dark Web services in a real-time and continuous fashion and fuses the collected data in a single analytics framework.
- We overcome challenges of information extraction in a globalized world of cyber crime. Using machine translation techniques, BlackWidow can investigate relationships between forums and users across language barriers. We show that there is significant overlap across forums, even across different languages.
- We illustrate the power of real-time intelligence extraction by conducting a study on seven forums on the Dark Web and the open Internet. In this study, we show that BlackWidow is able to extract threads, authors and content from Dark Web forums and process them further in order to create intelligence relevant to the cyber security domain.

The remainder of this work is organized as follows. Section 2 provides the background on the concepts used throughout, while Section 3 discusses the challenges faced during the creation of BlackWidow. Section 4 describes BlackWidow's architecture before Sections 5 and 6 respectively present the design and the results of a Dark Web measurement campaign. Section 7 discusses some case studies, Section 8 examines the related work and finally Section 9 concludes this paper.

2. BACKGROUND

In this section, we introduce the necessary background for understanding the BlackWidow concept. In particular, we provide the definitions and also explain the underlying technological concepts relating to the so-called Dark Web and to Tor Hidden Services.

A. The Deep Web and Dark Web

The media and academic literature are full of discussions about two concepts, the Dark Web and the Deep Web. As there are no clear official technical definitions, the use of these terms can easily become blurred. Consequently, these terms are often used interchangeably and at various levels of hysterics. We provide the most commonly accepted definitions, which can also be used to distinguish both concepts.

1) The Deep Web

The term 'Deep Web' is used in this work to describe any type of content on the Internet that, for various deliberate or non-deliberate technical reasons, is not indexed by search engines. This is often contrasted with the 'Surface Web', which is easily found and thus accessible via common search engine providers.

Deep Web content may, for example, be password-protected behind logins; encrypted; its indexing might be disallowed by the owner; or it may simply not be hyperlinked anywhere else. Naturally, much of this content could be considered underground activity, e.g., several of the hacker forums that we came across for this work were also accessible without special anonymizing means.

However, the Deep Web also comprises many sites and servers that serve more noble enterprises and information, ranging, for example, from government web pages through traditional non-open academic papers to databases where the owner might not even realize that they are accessible over the Internet. By definition, private social media profiles on Facebook or Twitter would be considered part of the Deep Web, too.

2) The Dark Web

In contrast, the Dark Web is a subset of the Deep Web which cannot be accessed using standard web browsers, but instead requires the use of special software providing access to anonymity networks. Thus, deliberate steps need to be taken to access the Dark Web, which operates strictly anonymously both for the user and the service provider (e.g., underground forums).

There are several services enabling *de facto* access to anonymity networks, for example the Invisible Internet Project (IIP) or JonDonym [7]. However, the so-called

‘Hidden Services’ provided by the Tor project remain the most popular *de facto* manifestation of the Dark Web. In the next section we provide a detailed technical explanation of Tor’s Hidden Service feature, which formed the basis of the analysis done by BlackWidow.

B. Tor Hidden Services

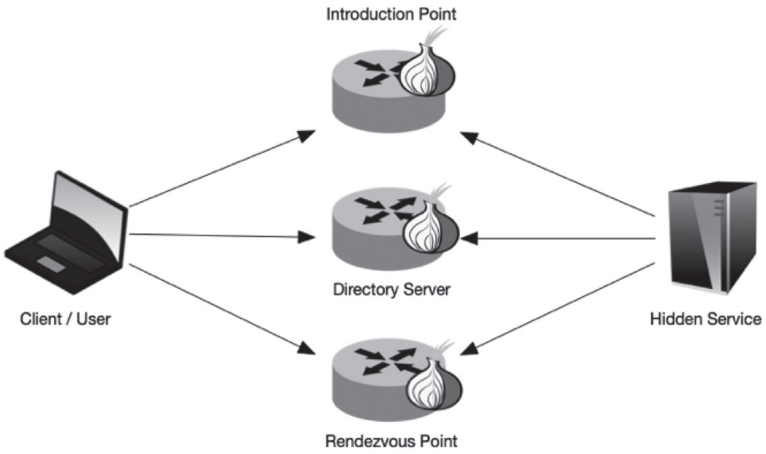
Tor, originally short for The Onion Router, is a project that seeks to enable low-latency anonymous communication through an encrypted network of relays. Applying the concepts of onion routing and telescoping, users obtain anonymity by sending their communication through a so-called *Circuit* of at least three relay nodes.

As Tor is effectively a crowdsourced network, these relays are largely run by volunteers. The network has been an important tool for many Internet users who depend on anonymity, from dissidents to citizens in countries with restricted Internet access. However, there have been many vulnerabilities found and discussed in the literature which could lead to deanonymization of Tor users. As it is not desired to authenticate the identity of every Tor relay, it is widely considered possible that state actors such as intelligence agencies run their own relay nodes, by which they may exploit some of these vulnerabilities in order to deanonymize users of interest [8]. Despite these potential threats, Tor is the best-known and most popular way to hide one’s identity on the Internet.

Besides enabling users to connect to websites anonymously, Tor offers a feature called *Hidden Services*. Introduced in 2004, it adds anonymity not only to the client but also to the server, also known as responder anonymity. More concretely, by using such Hidden Services, the operator of any Internet service (such as an ordinary web page, including forums or message boards, which we are interested in for this work) can hide their IP address from the clients perusing the service. When a client connects to the Hidden Service, all data is routed through a so-called *Rendezvous Point*. This point connects the separate anonymous Tor circuits from both the client and the true server [9].

Figure 1 illustrates the concept: overall, there are five main components that are part of a Hidden Service connection. Besides the Hidden Service itself, the client and the Rendezvous Point, it requires an *Introduction Point* and a *Directory Server*.

FIGURE 1. GENERAL ILLUSTRATION OF THE TOR HIDDEN SERVICE CONCEPT.



The former are Tor relays, which forward management information necessary to establish the connection via the Rendezvous point and are selected by the Hidden Service itself, which is necessary to connect the client and the Hidden Service at the Rendezvous point. The latter are Tor relay nodes, where Hidden Services publish their information and which are then communicated to clients in order to learn the addresses of the Hidden Service's introduction points. These directories are often published in static lists and are in principle used to find the addresses for the web forums used in BlackWidow.

It is unsurprising that Tor Hidden Services are a very attractive concept for all sorts of underground websites, such the infamous Silk Road or AlphaBay and due to their popularity form in effect the underlying architecture of the Dark Web.

3. CHALLENGES IN DARK WEB MONITORING

The overarching main issues in analyzing the Dark Web for cyber security intelligence relate to the fact that a vast amount of unstructured and inaccessible information needs first to be found and then processed. This processing also needs to be done in a scalable way that enables humans to collect useful intelligence quickly and reliably. In the following, we outline the concrete challenges that needed to be overcome in developing BlackWidow.

A. Acquisition of Relevant Target Forums

The first challenge is the identification of target forums that are relevant to our operation, i.e. those that contain users and content relating to cyber security intelligence. Due to the underground nature of the intended targets, there is no curated list available that could be used as input to BlackWidow. Intelliagg, a cyber threat intelligence company, recently attempted to map the Dark Web by crawling reachable sites over Tor. They found almost 30,000 websites; however, over half of them disappeared during the course of their research [1], illustrating the difficulty of keeping the information about target forums up to date.

Combined with the mentioned previously fact that 87% of Dark Web sites do not link to any other sites, we can deduce that the Dark Web is more a set of isolated short-lived silos than the classical Web, which has a clear and stable graph structure. Instead, only loose and often outdated collections of URLs (both from the surface Internet as well as Hidden Services) exist on the Dark Web. Consequently, a fully automated approach to overcome this issue is infeasible and a semi-manual approach must initially be employed.

B. Resource Requirements and Scalability

Several technical characteristics of the acquired target forums require the use of more significant resource inputs. As is typical in analyzing large datasets obtained from the Dark Web, it is necessary to manage techniques which limit the speed and the method of access to the relevant data [10].

Such techniques include the deliberate (e.g., artificial limiting of the number of requests to a web page) and the non-deliberate (e.g., using active web technologies such as NodeJS, which break the use of faster conventional data collection tools). Typically, these issues can be mitigated by expending additional resources. Using additional virtual machines, bandwidth, memory, virtual connections or computational power, we can improve the trade-off with the time required for efficient data collection. For example, by using several virtual private networks (VPNs) or Tor circuits, it is possible to parallelize the data collection in case there is a rate limit employed by the target.

Surprisingly, a factor not challenging our resources was the habit of extensively vetting the credentials or ‘bona fides’ of forum participants before allowing access. A sufficient number of the largest online forums are available without this practice, which enabled data collection and analysis without having to manually circumvent such protection measures. However, since we did encounter at least some such forums (or parts of forums), our approach could naturally be extended to them, although this would require significant manual resource investment.

C. Globalized Environment

As cyber security and cyber crime have long become a global issue, underground forums with relevant pieces of information are available in practically all languages with a significant number of speakers. Most existing studies of Dark Web content have focused on English or another single language (e.g., [2]). However, the ability to gather and combine information independent of the forum language broadens the scope and the scale of BlackWidow significantly. By employing automated machine translation services, we are able to not only increase the range of our analysis but also detect relationships and common threads and topics across linguistic barriers and country borders.

Naturally, this approach comes with several downsides. For example, it is not possible to employ sentiment or linguistic analysis on the translated texts nor is the quality of state-of-the-art machine translation comparable to the level of a human native speaker. However, given BlackWidow's aims of scalable and automatic intelligence gathering, these disadvantages can be considered an acceptable trade-off.

D. Real-Time Intelligence Extraction

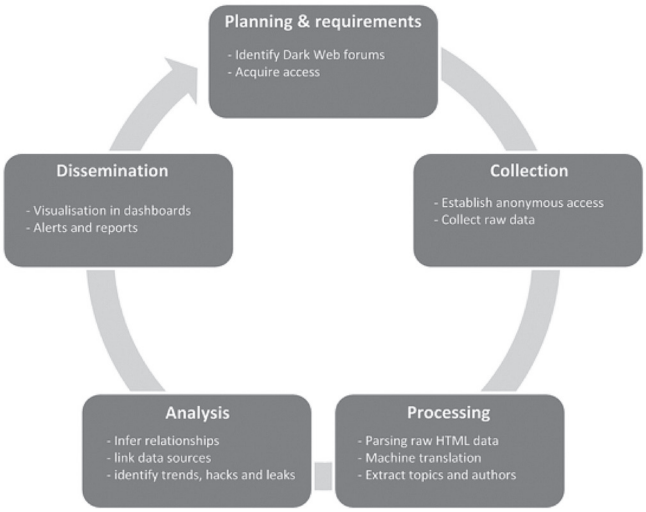
Beyond the previous issues, BlackWidow focuses in particular on the challenges posed by the nature of a real-time intelligence extraction process. Whereas previous studies have collected data from the Dark Web for analytical purposes, they have typically concentrated on a static environment. In contrast to collecting one or several snapshots of the target environment, BlackWidow aims to provide intelligence and insights much faster. Real-time capability is a core requirement for the longer-term utility of the system, due to the often very limited lifetime of the target forums.

To enable these functionalities, a high grade of automation is required, from the collection to the live analysis of the data. After the initial bootstrapping of sources and creating a working prototype, it is imperative that the processes require less manual input beyond normal human oversight tasks.

4. ARCHITECTURE OF BLACKWIDOW

In this section, we describe the basic architecture of BlackWidow. We largely abstract away from the exact technologies used and focus on the processing chain and the data model that enabled us to analyze the target forums in real time. Figure 2 shows the processing chain, including five phases defined as a recurrent cycle. The phases of the cycle are highly inspired by the conceptual model of the intelligence cycle [11]. Like the intelligence cycle, theses phases are continuously iterated to produce new insights.

FIGURE 2. BLACKWIDOW PROCESS CYCLE.



A. Planning and Requirements

The key focus of BlackWidow is on automation; manual work should only be needed for the integration of target forums in the initial planning and requirements phase, while all other phases are highly automated.

1) Identifying Dark Web Forums

The first suitable target forums are identified by hand to bootstrap the process and overcome the challenges described in Section 3.A. After obtaining a foothold, BlackWidow then aims to analyze the content of these forums in order to obtain further links and addresses to other targets in a more automated fashion in later iterations.

2) Gaining access

Since most forums require some sort of login to access the site, BlackWidow needs personal accounts to authenticate on each site. The way to acquire such logins differs on each site. While certain sites only request new users to provide a valid email address, others have higher entry barriers with reputation systems, measures of active participation, or even requiring users to first buy credits.

B. Collection

After the planning and requirements phase, all steps are fully automated. The collection phase deals with establishing anonymous access to the forums over Tor and the collection of raw data.

1) Establishing anonymous access to forums

We establish anonymous gateways to the identified forums using Docker containers, Tor to access Hidden Services and Virtual Private Networks (VPN) for regular Deep Web sites. Here, it is necessary to add custom functions to BlackWidow, which emulate typing and clicking behavior in order to log in automatically and subsequently detect whether the gateway has successfully logged into the target or not.

2) Collection of raw data

For the actual collection of the forum content and metadata, we employed the node.js headless Chrome browser puppeteer [12] as a crawler within the Docker containers. While it requires more resources than other collection methods, it more closely emulates the behavior of real forum users, meaning it more easily avoids defensive action by the Dark Web marketplace operators. In order to improve the speed, the collection is distributed across multiple containers and parallelized.

C. Processing

The processing phase deals with parsing the collected raw HTML data from the previous phase, translating the content into English and extracting the entities of interest to feed a knowledge graph.

1) Parsing raw HTML data

Since BlackWidow retrieves data over a headless browser, the data to process is in the Hypertext Markup Language (HTML). Extracting structured information from HTML data can be quite challenging depending on the layout of the forums. BlackWidow implements a standard HTML parser that we adapt to the layout of each forum. While this approach may seem expensive at first, many forums have a similar layout such that the same parsers can be reused for different forums. The output of the HTML parser for each page is a structured file including only the text information from the HTML page.

2) Translation of raw data in foreign languages

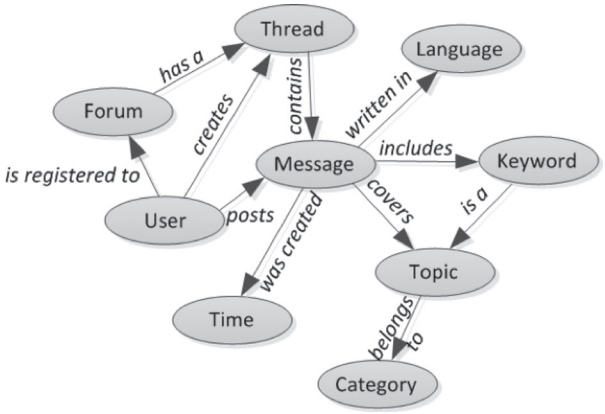
As much of the collected raw data contains content in several languages, we used automated machine translation to convert all non-English content into English. Through the use of Google's translation API, we obtain state-of-the-art translations, which enable the more complex data modeling and relationship analysis over forums in different languages in the follow-up phases.

3) Information extraction

To extract relevant information from the translated text from the gateways, we developed so-called extractors in Scala, which were also processed in a distributed fashion using the Apache Spark analytics framework. BlackWidow extracts

information about the forum writers and their content, i.e. the titles of forum threads and the posted messages. It then constructs a knowledge graph that connects threads, actors, messages and topics. Figure 3 shows the underlying data model of the knowledge graph of BlackWidow. The collected raw data and the knowledge graph is then put into Elasticsearch, a search engine based on Lucene [13]. As a tool for data exploration, it reads structured data and interprets timestamps and locations.

FIGURE 3. DATA MODEL REPRESENTING THE KNOWLEDGE GRAPH IN BLACKWIDOW.



D. Analysis

While inferring simple relationships between messages and authors is a relatively easy task given the HTML structure of the forums, other types of relationships and information extraction steps for the knowledge graph require advanced data analysis techniques. BlackWidow's goal is to automatically find relationships and trends across different threads and forums; the following processing steps are thus executed in this phase.

1) Infer user relationships

Relationships between users are mainly inferred in BlackWidow through the analysis of threads, since users of Dark Web forums barely link to each other explicitly as in classical social networks such as Facebook or LinkedIn [6]. A thread is always created by a single user and many different users then start posting messages on this thread. BlackWidow infers a relationship between users by ordering all messages in the same thread by their message times. We define a relationship from user A to user B if user B posted a message after user A in the same thread. The intuition is that user B is interacting with user A when he replies to his messages.

2) Identify topics

While messages in forums are commonly structured in threads and categories, it is not always obvious to see which threads cover the same topics. To facilitate trend analysis across different threads and forums, BlackWidow automatically identifies topics by means of automatic topic modeling. BlackWidow implements unsupervised text clustering techniques based on Latent Dirichlet Allocation (LDA) to classify messages into groups. These groups are then assigned to higher-level categories of interests such as botnets, databases, exploits, leaks and DDoS.

3) Identify cyber security trends

To identify cyber security trends, BlackWidow fuses the messages, topics and categories from the different forums and computes aggregated time series. These time series form the basis to identify trending topics, e.g. when the time series experiences a high growth or decline over short periods. Long-term trends are also detectable given that all collected messages are time-stamped and thus provide information over the whole lifetime of the forum.

E. Dissemination

Finally, it is important to disseminate the extracted information so that it can be easily processed by human intelligence analysts. To serve this purpose, BlackWidow supports various types of data visualizations and data query interfaces for exploratory analysis. For example, customized Kibana dashboards provide real-time views of the processed data that is stored in the Elasticsearch database. These dashboards can be generated and customized easily by the users allowing different views depending on the question of interest.

Finally, users may realize that some data is missing or that the additional forums should be integrated. The cycle of BlackWidow's architecture supports users to refine the planning and data collection requirements, thus closing the loop of the intelligence process.

5. STUDY DESIGN

After describing the architecture of BlackWidow, we now explain the goals of the study conducted for this paper. The study was designed to show the power and effectiveness of our automated data extraction and analysis efforts for the Dark Web.

A. Information Extraction

Forum contents are usually structured hierarchically. Users provide or exchange information by posting messages, known as "posts". Collections of posts belonging to

the same conversation are called threads. Threads can be separated by categories such as “Drugs”, “Exploits”, or “Announcements”. Besides the actual message, posts also provide meta information on the author (e.g., username, date of registration) and the exact date and time when the message was posted.

While posts are certainly the most interesting source of information in a forum, it is worth taking other parts of the forum into account for information retrieval as well. For example, most forums have a publicly available list of members which provides links to the profiles of all users registered in the forum. By additionally crawling the public profiles of all registered users, it is possible to gather information on passive users and the overall community as well. User profiles often provide useful information, such as registration date and time of last visit.

To extract all this information from the HTML-based forum data collected by BlackWidow, we implemented HTML parsers for each forum based on jsoup. Although forums generally have a very similar structure, the underlying HTML representations differ significantly depending on the platform. The consequence is that for each different forum platform (e.g., vBulletin), a separate forum parser is required.

For this analysis, we limit our implementation to parsing posts and user profiles. Our parsers transform the HTML-based representation of posts and user profiles into a unified JSON-based format. More specifically, each post is transformed into a JSON object with attributes forum, category, thread, username, timestamp and message. Objects from non-English forums are extended with the English translations of categories, threads and messages. User profiles are parsed into JSON objects with attributes forum, username, registration date and (where available) last visit date.

B. Forum Selection

For the purpose of this study, we collected data from seven forums as a proof of concept, as the manual integration of new forums can require significant time investment. At the time of writing, roughly one year after collecting the data, only four of the scanned forums are still online, confirming the short lifetime and high volatility of such forums. Overall, three of the seven forums were only accessible in the Dark Web and four were Deep Web forums. The languages used in the forums were Russian, English and French. An overview over the considered forums and the most popular categories (by number of posts) is provided in Table 1.

TABLE 1: OVERVIEW OF THE FORUMS CONSIDERED IN OUR ANALYSIS.

#	Type	Language	Top Categories	Online as of 12/ 2018
Forum 1	Deep Web	English	News, Porn, Software, Drugs	Yes
Forum 2	Deep Web	Russian	Marketplace, Electronic Money, Hacking	No
Forum 3	Dark Web	French	Drugs, News, Porn, Technology	No
Forum 4	Dark Web	Russian	Marketplace, General Discussions, Hacking, Security	Yes
Forum 5	Deep Web	English	Gaming, Leaks, Cracking, Hacking, Monetizing Techniques, Tutorials	Yes
Forum 6	Dark Web	French	News, Frauds, Conspiracy Theories, Drugs, Crime	No
Forum 7	Deep Web	Russian	Software, Security & Hacking, DDoS Services, Marketplace	Yes

6. STUDY RESULTS

A. Target Analysis

The size of each forum can be determined either in the number of posts or in the number of users. Both metrics for the crawled forums are shown in Figure 4 and 5. Forum 5 has by far the largest community with 67,535 registered users, while Forum 3 has (also by a considerable margin) the most content with over 288,000 posts. Forum 3 is also the forum with the most active community in terms of average posts per user. On average, each user had posted 22.74 messages in Forum 3. In contrast, the community of Forum 5 seemed to consist largely of passive users, since for each user, there were only 2.28 messages, roughly one tenth of those in Forum 3.

FIGURE 4. NUMBER OF USERS EXTRACTED FROM EACH FORUM.

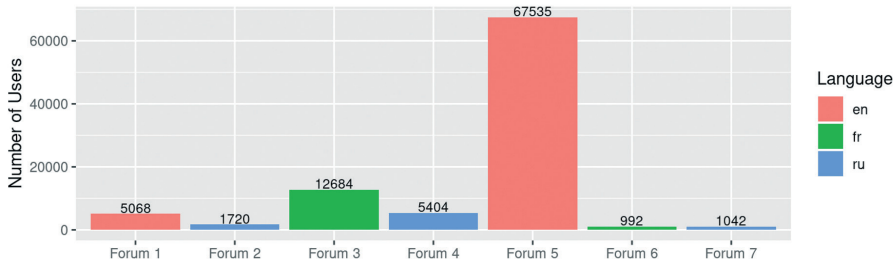
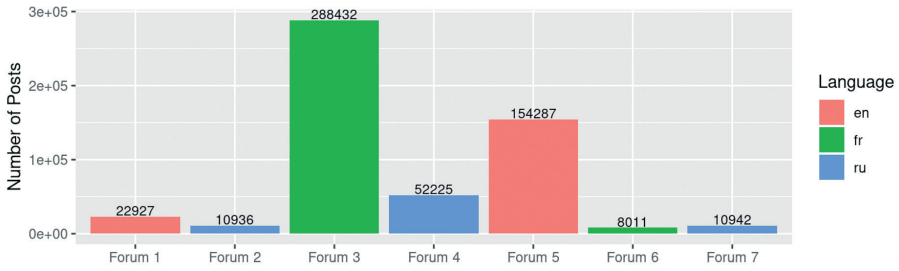


FIGURE 5. NUMBER OF POSTS EXTRACTED FROM EACH FORUM.



We hypothesize that the extremely large number of passive users in Forum 5 comes from the fact that the forum is a Deep Web forum, meaning that it does not require users to use additional software (such as the Tor browser) to sign up. As a consequence of this significantly lower technical hurdle, is can be accessed much more easily than Dark Web forums and is therefore open to a broader, less tech-savvy audience.

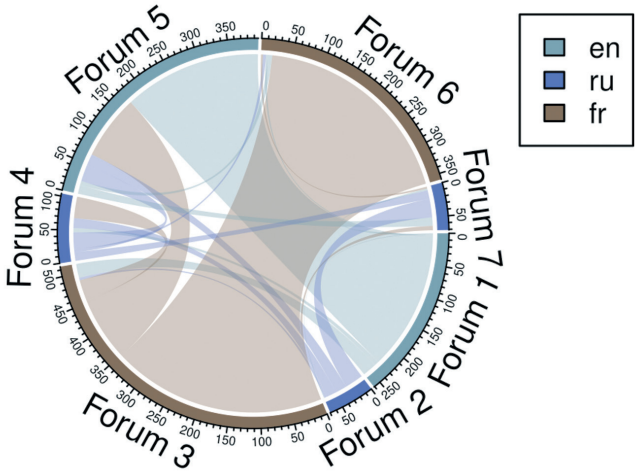
B. Forum Relationships

In order to get some insights on the relationships between the forums, we compared the sets of usernames of the forums. More specifically, we were interested in the intersections of these sets to see whether these forums host separate communities or whether there are significant overlaps. Surprisingly, those usernames that appeared most often were very specific, suggesting that they actually belonged to the same person. In fact, generic usernames such as "admin" or "john" were very rarely seen. Instead, users tended to individualize their usernames, for example by using *leetspeek*,[1] most likely as a means of anonymous branding. This tendency benefits the social network analysis conducted in this section since it provides us with reliable information about individual users, even across forums.

The result of this analysis is depicted as a chord diagram in Figure 6. Unsurprisingly, there are significant overlaps across forums in the same language. More interesting, however, is the fact that Forum 5, the forum with the largest community, has significant overlaps with most other forums, even if they are in a different language. By looking at these intersections as information dissemination channels, Forum 5 certainly provides the best entry point to spread information across the deep and dark side of the web.

1 A system of modified spelling, whereby users replace characters with resembling glyphs.

FIGURE 6. RELATIONSHIPS BETWEEN THE FORUMS IN TERMS OF COMMON USERS.



C. Author Relationships

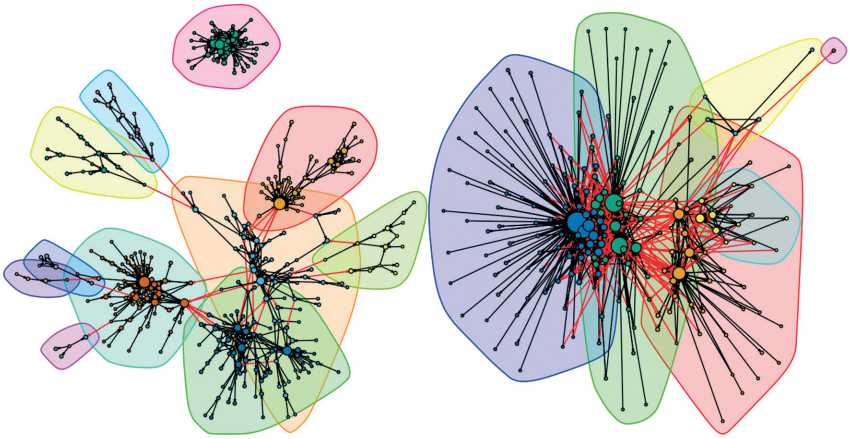
In order to analyze the internal relationships between users of forums, we first need to establish a reasonable definition of user relationships. While there are clearly defined relationships in social networks such as Facebook or Twitter, forum users do not have natural links such as friendships or followers. Given the hierarchical structure of forums, however, we can identify users with common interests by looking at the threads in which they are active together. We therefore define the relationships between two users in a forum by the number of threads in which both users posted messages.

Based on the creation timestamp of each post, we can also add a direction to this relationship by acknowledging which user merely reacted to a post of another user; i.e., which user posted a message in the same thread at a later point in time. This directed relationship will help us distinguish information or service providers from consumers. This is possible since a common communication pattern, for example in forum-based marketplaces, is that someone shares data or services in a new thread and interested users must post a reply (e.g., “thank you”) in order to access the shared content.

After these relationships were established, we used the Walktrap community-finding algorithm [14] with a length of 4 to determine sub-communities in the forums. These sub-communities evolve naturally since forums often cover many different unrelated

topics. For example, users interested in drugs might not be interested in hacking and vice versa, resulting in two sub-communities.

FIGURE 7. NETWORK SHOWING THE RELATIONSHIPS BETWEEN USERS OF FORUM 4 (LEFT) AND 5 (RIGHT).



The result of this analysis for Forums 4 and 5 is shown in Figure 7. The vertices in the graphs represent the individual users, while the (directed) edges show the relationships as defined above. Each sub-community is indicated by a color. The size of each node in the network represents the number of incoming edges, i.e., its degree. In comparison, the structural differences of the communities of the two forums are clearly visible. Forum 4, which has a much smaller community, is much denser, meaning that there are many more relationships between users, even across the different sub-communities.

The network analysis enabled us to select sub-communities and identify their key users, i.e., the most active information or service providers. For instance, the completely separate sub-community in Forum 5 is a group of so-called skin gamblers, i.e., people who bet virtual goods (e.g., cryptocurrencies) on the outcome of matches or other games of chance. Another sub-community in Forum 5 deals with serial numbers of commercial products, with one user being a particularly active provider.

It is worth noting that, besides active providers, forum administrators and moderators also stick out in terms of node degree (activity) as they post a lot of administrative messages. For example, one user was very prominent in a sub-community and by manually checking his posts we found that he was a very active moderator who enforced forum rules very strictly and made sure transactions were being handled correctly. His power to enforce rules and certain behavior was established by a system

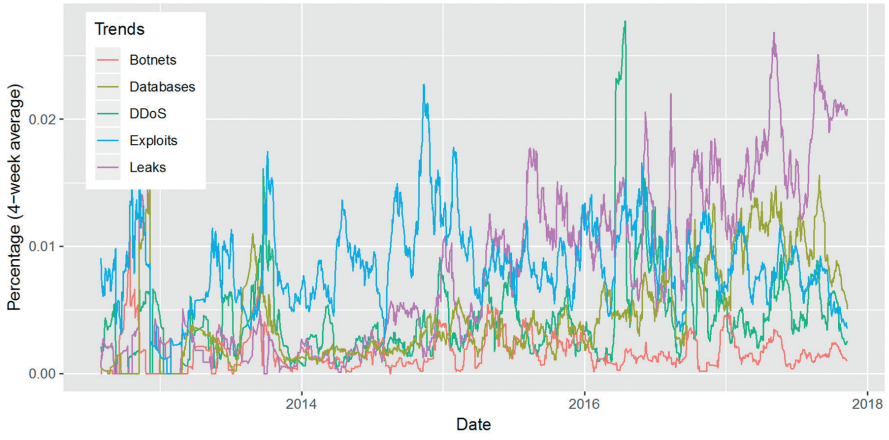
of reputation, in which users must gain hard-earned reputation points, for example by posting free content or being an active community member over a long period of time. Once a certain reputation is earned by a user, it becomes much easier for her or him to sell products on the marketplace; or they can charge higher prices as the risk of scam for buyers is lower. This system provides administrators and moderators with a certain leverage, since a ban from the forum would mean a complete loss of hard-earned reputation.

7. CYBER SECURITY INTELLIGENCE EXAMPLES

After conducting our quantitative study, we now discuss some exemplary trends and case studies that we noticed using BlackWidow during its initial deployment in 2017 to collect and analyze forum datasets dating back to 2012.

A. Forum Trends

FIGURE 8. CYBERSECURITY TRENDS BETWEEN 2012 AND 2017 IN SEVEN FORUMS AS OBSERVED BY BLACKWIDOW.



It is possible to use BlackWidow's functionality to look at the popularity of different concepts over time, which can aid the intelligence analyst in finding sudden anomalies or identifying trends that suggest increased or decreased importance. Figure 8 shows the fraction of all posts for the five most popular identified cyber security categories over a time frame of five years from the end of 2012 to the end of 2017. The time series are generated using the running mean of the number of posts in the respective

topics over time and normalized with respect to the overall activity in the considered period. The topic assignment is based on regular expressions and string matching.

From this, we can see a substantial change in the number of times that forum actors were discussing *leaks*, which increases roughly ten-fold in 2017 and outpaces the other groups in number of mentions significantly by the end of the period. Related to leaks, posts on *databases* seem to become increasingly popular, while talk of *exploits* remains more or less constant as a trend, with several peaks, e.g. at the end of 2014. *DDoS* and *botnets* are the least popular of the five; the significant DDoS peak in the beginning of 2016 was caused by one of the analyzed forums itself being the victim of a DoS attack.

B. Discovered Leaks and Exploits

During the course of our study, we encountered various data leaks consisting of usernames and passwords. As an example, BlackWidow crawled links to a list of half a million leaked Yahoo! accounts, a well-known dataset from a hack in 2014 (officially reported by Yahoo! in 2016). Perhaps surprisingly, these leaked datasets were accessible for free through direct links, highlighting again that security-relevant information can indeed be automatically collected by BlackWidow without interacting personally with criminal data brokers.

Exploits for various platforms were also found abundantly. Again, the open nature of the forums makes it possible to collect large amounts of exploits for free. While a systematic analysis on the quality and novelty of the individual exploits is outside the scope of this paper, we are confident that BlackWidow constitutes a very useful data source to better understand the cyber threat landscape and anticipate exploits that may be expected in the wild. Security professionals and defenders should therefore aim at analyzing such information to anticipate emerging threats.

8. RELATED WORK

Web forums inside and outside the Dark Web have been an active field of research in the recent past, with authors approaching them from a wide variety of angles, including cyber security and intelligence.

The closest works to ours relate to underground crawling systems. Pastrana et al. [6] recently built a system that looks at cyber crime outside the Dark Web. The authors discuss challenges in crawling underground forums and analyze four English-speaking communities on the Surface Web. In contrast, Nunes et al. [15] mine Dark Web and Deep Web forums and marketplaces for cyber threat intelligence. They show that it is

possible to detect zero-day exploits, map user/vendor relationships and conduct topic classification on English-language forums, results that we have been able to reproduce with BlackWidow.

Benjamin et al. [16] explore cyber security threats in what they call the "hacker web", with a focus on stolen credit card data activity but also potential attack vectors and software vulnerabilities. The authors extract data from carding shops, the Internet Relay Chat (IRC) and web forums, but do not investigate Tor Hidden Services.

In [17] and [18], the authors look at major hacker communities in the US and China, aiming to identify key players, experts and relationships in open web forums. They base their approach on a framework for automated extraction of features using text analytics and interaction coherence analysis. Similarly, Motoyama et al. [19] look at six different underground forums on the open web, providing a measurement campaign on historical data. The extensive quantitative data analysis covers features from the top content over the size of the overlapping user base to interactions and relationships between the users. However, their analysis is based on leaked SQL dumps of the forums, while BlackWidow is a framework that collects information in real time through the frontend of the forums.

Outside the academic literature, we find several commercial enterprises which aim to conduct automated analysis of cyber security intelligence from the Dark Web, among other sources. Two examples are provided by DarkOwl [20] and Recorded Future [21], which monitor the Dark Web in several languages and offer to detect threats, breached data and indicators of compromise.

To the best of our knowledge, this paper is the first to discuss real-time data collection in the Deep and Dark Web and the integration of external translation capabilities in a scalable way. Additionally, our results have been able to show that there is substantial overlap between actors across forums, even if they are not in the same language.

9. CONCLUSION

It is imperative in the current cyber security environment to have a real-time monitoring solution that works across languages and other barriers. We have shown in this paper that early detection of cyber threats and trends is feasible by overcoming several key challenges towards a comprehensive framework.

While we can be fairly certain that techniques similar to ours are being used by both governmental and private intelligence actors around the world, it is important to

analyze their power in a more open fashion, giving rise to possible scrutiny and further development. By implementing BlackWidow as a proof-of-concept collection and analysis tool, we show that monitoring of the Dark Web can be done with relatively little resources and time investment, making it accessible to a broader range of actors in the future.

REFERENCES

[1] Intelliagg, "Deeplight: Shining a Light on the Dark Web. An Intelliagg Report," 2016.

[2] M. W. Al Nabki, E. Fidalgo, E. Alegre and I. de Paz, "Classifying illegal activities on TOR network based on web textual contents," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017.

[3] A. Biryukov, I. Pustogarov and R.-P. Weinmann, "Trawling for tor Hidden Services: Detection, measurement, deanonymization," in *IEEE Symposium on Security and Privacy (S&P)*, 2013.

[4] Hyperion Gray, "Dark Web Map," [Online]. Available: https://www.hyperiongray.com/dark-web-map/. [Accessed 7 1 2019].

[5] V. Griffith, Y. Xu and C. Ratti, "Graph Theoretic Properties of the Darkweb," *arXiv preprint arXiv:1704.07525*, 2017.

[6] S. Pastrana, D. R. Thomas, A. Hutchings and R. Clayton, "CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale," in *Proceedings of the 2018 World Wide Web Conference*, 2018.

[7] A. Pescape, A. Montieri, G. Aceto and D. Ciuonzo, "Anonymity Services Tor, I2P, JonDonym: Classifying in the Dark (Web)," *IEEE Transactions on Dependable and Secure Computing*, 2018.

[8] K. Bauer, D. McCoy, D. Grunwald, T. Kohno and D. Sicker, "Low-resource routing attacks against Tor," in *Proceedings of the ACM Workshop on Privacy in Electronic Society*, 2007.

[9] A. Biryukov, I. Pustogarov, F. Thill and R.-P. Weinmann, "Content and popularity analysis of Tor Hidden Services," in *IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW)* , 2014.

[10] I. Sanchez-Rola, D. Balzarotti and I. Santos, "The onions have eyes: A comprehensive structure and privacy analysis of Tor Hidden Services," in *Proceedings of the 26th International Conference on the World Wide Web*, 2017.

[11] L. K. Johnson, Ed., *Handbook of intelligence studies*, Routledge, 2007.

[12] "Puppeteer," [Online]. Available: https://pptr.dev. [Accessed 7 1 2019].

[13] Elastic, "Elasticsearch," [Online]. Available: https://www.elastic.co/products/elasticsearch. [Accessed 7 1 2019].

[14] P. Pons and M. Latapy, "Computing communities in large networks using random walks.," *Journal of Graph Algorithms and Applications*, vol. 10, no. 2, pp. 191-218, 2006.

[15] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart and P. Shakarian, "Darknet and deepnet mining for proactive cybersecurity threat intelligence," in *IEEE Conference on Intelligence and Security Informatics (ISI)*, 2016.

[16] V. Benjamin, W. Li, T. Holt and H. Chen, "Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops," in *IEEE International Conference on Intelligence and Security Informatics (ISI),* 2015.

[17] A. Abbasi, W. Li, V. Benjamin, S. Hu and H. Chen, "Descriptive analytics: Examining expert hackers in web forums," in *IEEE Joint Intelligence and Security Informatics Conference (JISIC)*, 2014.

[18] V. Benjamin and H. Chen, "Securing cyberspace: Identifying key actors in hacker communities," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2012.

[19] M. Motoyama, D. McCoy, K. Levchenko, S. Savage and G. M. Voelker, "An analysis of underground forums," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference,* 2011.

[20] "DarkOwl," [Online]. Available: https://www.darkowl.com. [Accessed 7 1 2019].

[21] "Recorded Future," [Online]. Available: https://www.recordedfuture.com. [Accessed 7 1 2019].