CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale

Sergio Pastrana
Cambridge Cybercrime Centre, Computer Laboratory
University of Cambridge
Sergio.Pastrana@cl.cam.ac.uk

Daniel R. Thomas
Cambridge Cybercrime Centre, Computer Laboratory
University of Cambridge
Daniel.Thomas@cl.cam.ac.uk

Alice Hutchings
Cambridge Cybercrime Centre, Computer Laboratory
University of Cambridge
Alice.Hutchings@cl.cam.ac.uk

Richard Clayton
Cambridge Cybercrime Centre, Computer Laboratory
University of Cambridge
Richard.Clayton@cl.cam.ac.uk

ABSTRACT

Underground forums allow criminals to interact, exchange knowledge, and trade in products and services. They also provide a pathway into cybercrime, tempting the curious to join those already motivated to obtain easy money. Analysing these forums enables us to better understand the behaviours of offenders and pathways into crime. Prior research has been valuable, but limited by a reliance on datasets that are incomplete or outdated. More complete data, going back many years, allows for comprehensive research into the evolution of forums and their users. We describe CrimeBot, a crawler designed around the particular challenges of capturing data from underground forums. CrimeBot is used to update and maintain CrimeBB, a dataset of more than 48m posts made from 1m accounts in 4 different operational forums over a decade. This dataset presents a new opportunity for large-scale and longitudinal analysis using up-to-date information. We illustrate the potential by presenting a case study using CrimeBB, which analyses which activities lead new actors into engagement with cybercrime. CrimeBB is available to other academic researchers under a legal agreement, designed to prevent misuse and provide safeguards for ethical research.

CCS CONCEPTS

• **Information systems** → **Web crawling**; • **Security and privacy** → *Social aspects of security and privacy*;

KEYWORDS

Underground Forums; Cybercrime; Money Laundering; Ethics; Data Sharing; Web Crawling; CrimeBot; CrimeBB

ACM Reference Format:
Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. 2018. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale. In *Proceedings of The Web Conference 2018 (WWW 2018), Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186178

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186178

1 INTRODUCTION

Cybercriminal communities bring together individuals interested in hacking, or trading in illicit materials. They often use online forums for communication, where knowledge and material about various illicit and deviant topics are shared. Examples include trading in stolen accounts or credit card details [17], fraudulent monetizing techniques such as e-whoring,[1] or trading of virtual game items acquired through bots and cheats. Interest in these communities has increased in recent years and there is a wide literature on the analysis of cybercriminal forums, from different fields such as criminology, economics, cybersecurity and sociology (see Section 2.1). Research topics have included: how trust is managed [11], marketplace trade [35], and sharing of hacking material [37]. However, the interests and motivations of community members are widespread and evolve over time. Thus, being able to analyse these communities at scale is important to understand pathways into crime [20], or to evaluate interests beyond the fraudulent activities the cybercriminals are undertaking, and to study their psychological profile [19].

Large scale analysis of social behaviours within these communities has been limited due to a lack of datasets that are maintained and kept up-to-date [4]. Some researchers opt not to share data due to privacy concerns and ethical issues. Thus, researchers wishing to conduct similar research need to first design their own crawler. This is time consuming and requires technical expertise not necessarily found in the social sciences, thereby deterring research in this field. Most of the datasets that are available are either incomplete due to partial crawling [35], or come from leaked databases [32], which contain outdated data. Due to the dynamic nature of cybercriminal communities, we need tools that maintain updated datasets.

In this work we first describe CrimeBot, a tool to scrape online forums, particularly cybercriminal communities. The problem of crawling online forums has been addressed in previous works [6, 23] and there are several tools available for such purposes. However, the adversarial settings on which criminal communities operate pose additional challenges [15]. For example, stealthy crawling is required to avoid interfering with the natural behaviour of users [40]. Additionally, underground forums require the use of techniques to bypass access control restrictions, and due to the dynamic nature of these communities, it is necessary to collect updates efficiently [15].

[1] A social engineering technique whereby partners in cyber-sex encounters are imitated, and victims are sold pictures or videos, usually obtained from underground forums.

The main characteristics, challenges and goals when crawling such forums are presented in Section 2.3.

CrimeBot uses a rich configuration language to adapt its crawling behaviour to particular settings, and relies on the use of external proxies (e.g. using different Tor circuits), which provides anonymity and scalability. If needed, CrimeBot manages manually registered accounts and session cookies to access the different forums.

CrimeBot regularly updates CrimeBB, a database of information gathered from different online communities. CrimeBB currently includes more than 48m posts, 4.5m threads and 1m accounts from 4 sites. The dataset has been collected over 9 months, and data spans more than one decade, from 2005 to the present (our collection is ongoing). The bulk of the dataset is from *hackforums*, a popular hacking community that has gained recent attention due to the Mirai botnet [2] (the source code of Mirai was released by a user of this forum in 2016 [26]), and the arrest of the alleged author of banking malware, who was apparently an active seller in this forum during his adolescence [27]. The dataset presents a unique opportunity to understand these communities at scale, and allows for longitudinal data analysis. The dataset is available for other academic researchers to use through data sharing agreements via the Cambridge Cybercrime Centre,[2] which addresses legal concerns and prevents potential misuse of the data (see Section 3.3).

Finally, to illustrate the potential of the dataset, in Section 5 we provide a case study analysing currency exchanges from the community in *hackforums*. We show how currencies have evolved over the last decade, and note the increase of exchanges involving Amazon gift cards in the last two years. We then analyse key actors involved in currency exchange and track their previous activity in the community. This analysis would not be possible without CrimeBB, since it includes data from the beginning of *hackforums* and thus allows us to track the historical posts of these key actors.

2 SCRAPING UNDERGROUND FORUMS

Underground forums enable different actors to share knowledge and illicit material. While not all the contents and goods posted on these forums are illegal, their origin or use may be. Interacting within these communities can be a stepping stone towards more serious online criminal activities [18, 20, 21]. For example, following the leak of the Mirai source code, DDoS attacks have used this botnet and its variants [2]. Currency exchanges (e.g. converting bitcoin to PayPal) can be used for money laundering [35]. Other illicit activities include trading online stolen accounts [18], hacking Massively Multiplayer Online (MMO) games for profit [13], or advertising booter services, technically offered as "Service Stress Tools", but which are actually used for DDoS attacks [41].

Different forums share common structure and functionality and are usually based on commodity forum software. Normally forums are structured into sub-forums and categories like "Hacking" or "Gaming". Members of a forum initiate new topics of conversation (threads) by writing an initial post. Other members can reply to threads with additional posts. Members also have a public profile.

These forums are public, but their access can be restricted to registered members. In most cases, they operate in the surface web, i.e. they do not rely on the Tor network, and thus they could be traced and shut down by law enforcement agencies. Common characteristics of underground forums are:

- They usually require registration to enter the site, or at least to have full access. In general, there are two roles for users navigating through online communities: visitors, who can access the forums without registering; and members, registered users who are logged onto the system. Some forums restrict visitors' actions, for example, they might not be able to view attachments or use certain functionalities.
- There may be restricted sections for upgraded (or VIP) accounts. Members can upgrade their status by paying a fee or earning activity rewards (e.g. quality of posts, reputation, etc.). Additionally, viewing some content requires members to have written a minimum number of posts.
- There are specific sections for commerce (marketplaces). Since members are pseudonymous, trust is managed by means of active reputation systems or special sections aiming at dealing with disputes.
- To avoid obvious illegality, they operate policies on what is not allowed. Thus, it is common to observe removed content and banned members.
- They might employ anti-scraping techniques.

The remainder of this section starts with a review of the literature on scraping underground forums. Then we discuss the ethical issues involved with this practice. Finally, the technical challenges to be overcome to successfully scrape these forums are presented.

2.1 Related work

A number of general-purpose crawlers have been developed for scraping online forums for research applications, such as iRobot [6], FoCUS [23], or more recently Vigi4Med [3]. While they aim to collect data at scale, crawling and retrieving information from underground forums presents different challenges to many other forums, such as overcoming anti-crawling techniques. Indeed, general-purpose crawlers assume cooperation from forum administrators to avoid being banned. This assumption does not hold for underground forums. Seminal work by Fu et al. provides insights on how to crawl web forums which are not easily accessible through regular search engines [15]. Most of the challenges posed by Fu et al. are applicable to retrieving forum data under adversarial conditions (these and other challenges are described below).

Prior research analysing data from underground forums either use leaked datasets [32] or use custom crawlers [28]. The use of leaked datasets have two main drawbacks. First, since users mentioned in these datasets may know about the leak, they may move to other online communities, changing their online behaviour and aliases. Thus, these datasets are outdated and may not be representative of current practices. Second, where datasets have been obtained illicitly, legal and ethical issues may arise when they are used for research [42]. In particular they may contain private data that would not be available to a web crawler.

Custom crawlers in the literature either lack technical solutions for the challenges posed by Fu et al. in 2010, or they do not clearly explain how they deal with them, for example, when dealing with login functionality [12] or with CAPTCHA challenges that prevent

[2] https://www.cambridgecybercrime.uk

auto-login [22]. Others opt to exclude forums that require registration [43]. In some cases the datasets obtained by custom crawlers are incomplete. Portnoff et al. present tools to analyse underground forums [35]. Their experiments relied on data from 8 forums, from which they partially scraped 3 of them and used complete dumps leaked for the remaining 5. While their dataset is publicly available, their focus was on the tools rather than the data and thus forums were only partially crawled [35].

The Open Discussion Forum Crawler (ODFC) use rules in the form of *[path, pattern]* to retrieve data from web forums. Research with data obtained using ODFC includes analysing threat indicators against critical infrastructures from a hacker forum [29], finding communities focused on different aspects of malware development [28], and analysing money laundering activities in two Russian forums [31]. While capable, the ODFC does not crawl forums that require registration due to ethics complications [14].

AZSecure is a tool that provides "cyber threat intelligence" by crawling and scraping hacker forums. The tool includes machine learning capabilities to identify assets being exchanged in the forums, such as exploits or malware source code [37]. The crawler module relies on the Tor network to prevent IP blacklisting, but it does not consider forums that require registration. Datasets are available via the DIBBs-ISI project [7].

Nunes et al. use a custom crawler to extract information from several hacker forums and marketplaces of hacking assets [33]. Authors indicate that the crawler "addresses design challenges like accessibility, unresponsive server, repeated links, etc.", but do not detail how the system manages access control and the potential anti-crawling techniques implemented by underground forums.

The "Darknet Market Archives" is a dataset of underground marketplaces collected by different authors [5]. These include some of the daily scrapes of Silk Road performed by Christin for a period of 2 years [8], and 35 other marketplaces from later work with Soska [40]. Due to the volatility of information from marketplaces, incremental crawling was required. The 35 marketplaces were crawled and scraped 1,908 times, thus the database includes 3.2TB of data [40]. While focused on marketplaces rather than forums, Soska and Christin describe many of the same challenges that are faced by CrimeBot, such as the need for stealthiness (to avoid blocking), the need for being registered in the site to gain access (they manually logged in to complete CAPTCHAs), and the need for a flexible design (during the crawling period the structure of the site was modified, and they had to modify their crawler).

2.2 Ethical issues

There are a variety of ethical issues that arise from research collecting and analysing data from underground forums which require careful review. The purpose of ethical review for research involving human participants is to consider potential harms, and identify ways that these may be mitigated or avoided. Ethical review may involve consideration by a Research Ethics Board (REB), often referred to as Institutional Review Boards (IRBs) in the US, or Ethics Committees in the UK. It is important that researchers obtain approval from their REB, not only to ensure that their research is ethical, but also to have some protection from liability [42]. For this work, we have followed the procedures established by our REB. We were exempted from REB approval for the collection of the data, but we require approval for analysing such data. We have obtained approval for various projects analysing the CrimeBB dataset, including the analysis presented in Section 5.

Despite the importance of ethical review, few researchers have explicitly addressed the ethics of data gathering or analysis in their papers, or disclosed that their research was reviewed by a REB. Research can be considered ethical if the benefits outweigh the potential harms. These considerations are not straightforward. For example, research that identifies ways to prevent crime is of public interest: reducing victimisation and the associated costs, and benefiting would-be offenders who, if deterred, will not be caught up in the stigmatising criminal justice system. However, it may be counter-argued that online markets (such as those trading in drugs) reduce the violence associated with offline markets [1].

We consider the ethical issues with collecting data separately from those issues relating to the analysis of data. This distinction is important, as collecting data involves understanding the behaviour of the forum as a computer system, rather than its users as human beings. However, the researcher faces some risks, as scraping the data may require them to break the terms of service associated with the accounts that they use. They may also circumvent technical measures designed to prevent scraping, such as the use of CAPTCHAs [30]. Martin and Christin [30] argue that terms of service on criminal marketplaces are legally unenforceable, and it is ethically justified to break these as the benefits outweigh the potential harms.

Christin, first in [8] and then in the collaboration with Soska [40] justify the user of crawlers for data collection. They argue that as they do not compromise the site itself, they cause no additional harms to the individuals (either site administrators or users). They claim that it is ethical to bypass the CAPTCHA by providing session cookies to the crawler, since this is a feature offered by site administrators for their visitors. Another ethical issue raised by Christin is the potential abusive use of the Tor network, which is compensated by deploying a fast Tor relay in the author's institution.

Analysing the data illuminates the behaviour of people, rather than computers. Therefore, research that analyses data scraped from forums should undergo ethical review to consider potential harms, ensure safeguards are in place, and to protect the researcher. While many ethical guidelines intended for offline settings are also applicable online, some can be more difficult to implement, such as obtaining informed consent [9, 44]. It is particularly unlikely that researchers will be in a position to obtain informed consent from those involved in illegal activities. Furthermore, contacting all participants could be considered spamming, and not all accounts will be active. While it may be possible to seek informed consent from the forum administrators [9], this could affect the results [40]. However, according to established ethical principles and guidelines (for example, the British Society of Criminology's statement on ethics [34]), informed consent may not be required when a) the dataset is collected from the Internet and thus it is publicly accessible, and b) the data will be used for research on collective behaviour, without aiming to identify particular members. Where informed consent is not obtained, the role of the REB is particularly important, as this is the oversight mechanism that protects the interests of the research participants [10, 34].

Whether a website is considered public or private is a subject of debate in the research community [30]. For example, if you have to register an account or solve a CAPTCHA to access a forum, is it public or private? Is there a difference if membership is restricted, or if registration is open to anyone? Is it reasonable to expect that users are aware that their communications are not private? The conclusion of Décary-Hétu and Aldridge [9], which coincides with other authors [8], is that what is considered private should coincide with the norms of the community. Therefore, content gathered from forums or markets with "crypto-anarchist and radical libertarian principles" would be publicly accessible.

A number of safeguards may be implemented in order to reduce the likelihood of harm. These include: not identifying individuals (including not publishing usernames); taking care to present results objectively; dealing appropriately with personal data (such as credit card data belonging to victims); and taking steps to protect the researchers. Some researchers also take the step of not disclosing which forums have been analysed [22, 29]. When it comes to presenting results, the researchers can ensure that they do not make comments that are likely to offend the community being studied, to protect themselves as well as the research participants. The researchers can further protect themselves by ensuring that they do not unintentionally download malware, child exploitation material, or terrorist materials; which can create security and legal issues. These risks can be mitigated at the data collection stage, by only collecting text data, excluding all files and images. Another mitigation is having a standard procedure to report such material if it is encountered, such as to law enforcement or an INHOPE member hotline[3] that responds to reports of child exploitation images.

This discussion relating to ethics is specific to the use of scraped forum data. We note the analysis of backend forum databases that have been leaked may require additional considerations. For example, these datasets may include additional data, such as private messages, registration email addresses, IP addresses, or posts shown only to certain forum members. Thomas et al. [42] further discuss the ethical issues researchers may face when using leaked datasets.

2.3 Crawling challenges and goals

Gathering data from underground forums poses different challenges than from other online sites [15]. Operators on these forums may not be willing to cooperate, and they usually deploy anti-crawling techniques such as CAPTCHA [9]. Due to the use of pseudonyms, and closed or banned accounts, asking for informed consent is not possible. Indeed, performing a stealthy scrape is necessary to prevent members changing their behaviour, because they know they are being monitored [40]. The main goal is to reverse-engineer the public contents of the internal database by scraping publicly available information from the websites. Following previous work by Fu et al. [15], we define the following crawling goals:

Completeness Ideally, all the boards, threads, posts and members should be included in the database. Moreover, all the information contained should be effectively retrieved.

Incremental crawling The tool should provide a means to revisit the forum and get only the new or modified boards, threads, posts or members.

Accessibility Forum content may not be easily reachable. Material may be restricted to members or require upgrades involving payment of fees or gaining reputation within the community [4]. The crawler must handle the access control imposed by the forums. Moreover, as far as it is possible, it should be able to bypass anti-bot techniques that may be encountered.

Flexibility Each forum has its own structure, and this also changes within forums. It is necessary to incorporate knowledge of the peculiarities of each forum and their changes over time [9]. Thus, the crawler should be extensible through new modules as new forums are encountered or existing forums are modified.

Verbosity The tool must log and inform about the crawling status, differentiating errors to allow for re-crawling the forum and, if needed, adaptation of the pertaining module.

Stealthiness In order to avoid being blocked, the crawler should mimic human behaviour. For example, the speed of the crawls can be limited [9]. This is one of the most challenging features, as imitating human behaviour is detrimental to the efficiency of the crawl.

Efficiency The crawler should not visit the same page twice, and every crawled HTML document should be stored locally for later analysis in case new information is required.

Non-textual content Underground forums often contain *non-textual content*, like attached files, multimedia, or source code [37]. These may contain content that would put the researcher at risk [43], and must not be downloaded. However, the tool should be able to detect and annotate the type of content included in the posts.

Most current work does not meet all these goals, which affects the quality of the data retrieved, as shown in Table 1. The columns detail the above requirements. Also shown are indicators of the size of the dataset used, whether the authors documented ethical issues, and whether the tools or datasets are publicly available.

3 DESIGN AND IMPLEMENTATION

CrimeBot is implemented in Python and Bash scripts, using Selenium[4] to automatically fetch and store HTML pages, and XPath[5] to scrape the content. A PostgreSQL database with an encrypted filesystem running on FreeBSD is used for storage.

The database schema covers the common structure of online forums. Sites are typically divided into a set of sub-forums or boards, composed of threads initiated by members of the community who write an initial post. Then, other users contribute to these threads by posting replies. Threads and posts have an author, who has a profile page, which is also scraped to retrieve information about the community members. An abstract flow diagram of CrimeBot is shown in Figure 1. First, given the main URL of a site to be crawled, the URL and ID of all the sub-forums of a site are retrieved and added as new tasks in a list. Then, each crawl starts by getting a task from the list, fetching the page from the URL, scraping its contents (updating the database) and storing the raw HTML. CrimeBot uses a logging system to detect failures while fetching or scraping a page.

[3] http://www.inhope.org/gns/our-members.aspx

[4] http://www.seleniumhq.org/

[5] https://www.w3schools.com/xml/xml_xpath.asp

Table 1: Summary of goals and ethical issues considered in previous work (✓ documented . not documented – not specified or not applicable ✳ last row represents this work)

Related work	Year 20XX	*Completeness*	*Incremental crawling*	*Accessibility*	*Flexibility*	*Verbosity*	*Stealthiness*	*Efficiency*	*Non-textual content*	*Ethics discussion*	*Data/tool sharing*	*REB involved*	*Size*
[15]	10	✓	✓	✓	✓	✓	✓	✓	✓	.	.	.	172GB data
[12]	10	.	.	✓	.	.	✓	✓	.	.	.	.	1m posts
[8]	13	✓	✓	✓	✓	.	✓	✓	.	✓	✓	.	24k products
[37]	15	.	.	✓	.	.	✓	.	✓	.	✓	.	671k posts
[43]	15	✓	.	.	.	.	✓	✓	✓	✓	.	.	–
[29]	15	✓	.	.	.	.	.	✓	.	✓	.	.	25k posts
[40]	15	✓	✓	✓	✓	.	✓	✓	.	✓	✓	.	3.2TB data
[28]	16	✓	.	.	.	.	.	✓	.	.	.	.	150k posts
[4]	16	.	.	✓	.	.	✓	✓	✓	✓	.	.	–
[33]	16	✓	✓	✓	✓	.	.	✓	.	.	✓	.	5k posts
[22]	16	✓	.	✓	✓	.	.	✓	.	✓	✓	.	–
[11]	16	.	.	.	.	.	.	.	.	.	.	.	450k posts
[35]	17	.	.	✓	.	.	.	.	.	.	✓	.	61k threads
[17]	17	.	.	.	.	.	.	.	.	.	.	.	388 threads
[16]	17	.	.	.	✓	.	.	.	.	.	.	.	600k posts
✳	18	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	48m posts

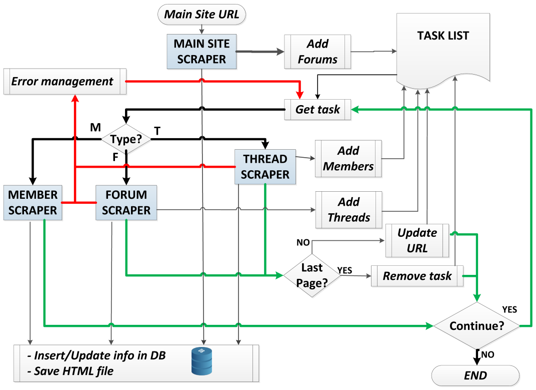

Figure 1: Flow diagram of CrimeBot

Sub-forums and threads can be split in several pages. CrimeBot checks when the last page has been scraped to remove the task. Finally, it checks whether the crawling should continue or not based on a set of configuration parameters, as explained below. The following sections present details of CrimeBot regarding the goals described in Section 2.3.

3.1 Completeness

To achieve completeness, CrimeBot includes a task to visit all the pages of a sub-forum and store its threads in the database. Then each thread is scraped to retrieve the posts. When parsing the posts, CrimeBot also retrieves the author and later crawls their profile page. Note that CrimeBot only retrieves information from members that are active in the forums, i.e. those that write at least one post.

Each forum provides different information and has a different structure. However, there are particular fields that are common in all the forums. Concretely, CrimeBot scrapes the following data from each page:

- **Sub-forum**. Title, ID, number of threads and number of posts.
- **Thread**. Heading, ID, Author (name and ID), number of posts.
- **Post**. ID, Author (name and ID), content, timestamp, cited posts (i.e., a reference to posts which are explicitly referred).
- **Member**. Name, ID, date of birth/age, avatar image (link), last visit, time spent online, registration date, signature, local time, reputation, prestige, home page, and total posts. Many of these fields are hidden or do not exist in particular forums, in which case we use default null or zero values.

3.2 Incremental crawling

Monitoring online communities and their evolution, and understanding new trends and longitudinal evolution requires up-to-date information. CrimeBot manages two modes of operation, i.e. *initial* and *incremental*. In the *initial* mode, all the information is gathered assuming that the database contains no previous data from the forum. In the *incremental* mode, the crawl is performed starting from the most recent items, and stops if it finds an item which was already updated in the database (the date of the last crawl is included for each item in the database). Thus, it is assumed that the threads and posts are chronologically sorted. For example, in the case of threads, the newest replies are posted in the last page of the thread. Accordingly, in the *initial* mode, the crawling starts from the beginning and proceeds forwards, while in the *incremental* mode it starts from the last page, crawls in backward direction, and stops when it finds a post that was already present in the database. The same strategy is implemented for the sub-forums. However, many sub-forums contain threads that are considered "hot" or are "pinned" at the beginning, no matter their age. Thus, CrimeBot always parses the first page completely, and stops when it finds a thread that was up-to-date from the second page onwards. Since threads show the timestamp of the last post in the heading, CrimeBot only adds to the task pool threads that are outdated.

3.3 Accessibility

Many forums require users to be registered in order to have full access, i.e. they do not allow visitors to view some content. For these forums, CrimeBot manages session cookies obtained from a previous registration. Both the registration and the login steps are performed manually by a human, since it usually requires account activation and solving *CAPTCHA* challenges. To this end, CrimeBot provides a session management module. This module opens a non-headless web browser and connects to the desired site using a given proxy or Tor circuit. After a human operator registers and enters the community, the module automatically stores the session cookies to be loaded during the crawling process. Many web forums discourage users that constantly login from different IPs or use different browsers. Thus, both the HTTP *useragent* header and the proxy used for registration of the account are stored and used during the crawling.

During our experimentation with CrimeBot, we have used two different proxy sets: Tor and a set of hosted servers. The main drawback of using Tor is that the availability of exit nodes is not guaranteed, due to connectivity failures. Moreover, certain nodes might be blocked [24, 39]. Thus, we first obtain a list of non-blocked circuits using a similar approach as Khattak et al. [24]. We first use Exitmap [45] to get current available exit nodes, from which we build Tor circuits and fetch the index page of the forums. Then, we perform an HTTP request from a non-Tor IP address (using one of our hosted servers) and compare both results to filter out blocked relays. Still, there are some blocked nodes that are not filtered since the blocking message appears after solving the *CAPTCHA* [39]. Since the registration of accounts is done manually, these nodes can be manually filtered by a human operator.

3.4 Flexibility

Forums can change their structure by adding new features or changing and removing existing ones. The scraper should be easily adapted to these changes. CrimeBot has a modular design where the main functionality is implemented in generic modules, e.g. for the interaction with the database or management of proxies. These modules do not require modification when adding new forums or modifying existing ones. Adding new scrapers for new communities is straightforward, since only few forum-specific modules should be added (see the blue boxes in Figure 1). Indeed, since many web forums use off-the-shelf software toolkits with similar structures, scraping modules from previous forums can be reused. Currently, CrimeBot contains scrapers for four of the most popular toolkits (i.e. vBulletin, phpBB, MyBB and SMF), which are used by more than 65% of all the forums in the Internet [46].

3.5 Verbosity and robustness

A log system helps to track the crawling status to detect failures [15]. CrimeBot implements a logging system based on log levels in order to record events.

The logging system allows for automatic processing of the output of a crawler. Most errors and warnings can be handled automatically by the tool (e.g. connectivity problems or removed threads). However, some errors require manual inspection, for example when there is a change in the structure of the forum or when the account has been banned. CrimeBot contains error management and self-recovery capabilities to detect when an error has occurred, and if possible, to continue with the regular crawl.

3.6 Stealthiness

Some web forums implement anti-crawling techniques and bot detection. Next, we enumerate some issues related with human behaviour on a forum, and the techniques implemented in CrimeBot to mimic such behaviours.

Client software. CrimeBot uses PhantomJS[6], a headless web browser which introduces some fingerprints that can be used by web servers to detect its presence [38]. Accordingly, CrimeBot modifies some HTTP headers to resemble non-headless browsers. Moreover, the user agent used during the registration of new accounts is used during the crawling.

Connection times. CrimeBot allows the crawling times for each user to be limited according to the timezone of the proxy or Tor exit relay used, for example to fix connection times to only day/night periods. Moreover, there is a time limit per day for crawling for a single user and a maximum number of pages visited per crawl, which both can be configured. While this negatively affects the efficiency of the crawling, it helps to avoid being banned and it can be mitigated by using more, parallel, crawlers.

Navigation patterns. There are several patterns for humans accessing threads [23]. Most commonly, they access a board and spend some time looking at new threads. Then, they click on the thread to read the posts, or in member to see their profile. Different navigation patterns are implemented and can be adapted in CrimeBot. Additionally it adapts the waiting times between fetching one page and another according to the amount of text being posted.

Since we cannot know in advance the level of moderation in each forum, these tools can be easily turned on/off or tuned and the crawling patterns can be adapted when needed, e.g. when either a user account or a proxy is banned [15]. For example, during our 9 months of operation we have seen forums that do not prevent bots at all, while others vary, with long periods where no accounts were banned, and periods where accounts were frequently banned, even when reducing the connection times to a minimum.

3.7 Efficiency

CrimeBot manages a task list with the items that must be parsed. Each task is composed by a unique ID, type of item (Thread, Post, etc.) and the URL. For example, when parsing the page of thread, all the members that have posted are added to the list so their profile pages can be crawled later. This allows for efficient distribution of tasks to different processes, which can be launched in parallel using different proxies and session cookies. For example, many web forums allow visitors to view the threads but not the actual posts and replies contained within them. Thus, it is possible to set up a crawler process without a registered account to crawl the boards and retrieve the thread information, and then use other processes with registered accounts to crawl and scrape the threads.

A common issue in automatic crawling is to prevent the retrieval of invalid or duplicate pages [6]. When CrimeBot fetches an HTML page, it also scrapes it to retrieve the interesting information, which filters out non-site URLs such as ads or external sources. Thus, only useful links are followed. Moreover, when any item (e.g. Thread, Member, etc.) is scraped, it is marked as "parsed" in the database, so it won't be retrieved twice even if its link appears in other pages (for example, when a thread is moved to another board). All the HTML pages are locally archived, and can be re-scraped if new information is required.

Another concern with crawlers using external proxies, is the overload of their bandwidth. Moreover, some of these proxies might be banned (for example, until June 2017, *hackforums* banned IPs belonging to datacenters). CrimeBot can be configured to use different proxies and crawling times to balance the use of bandwidth.

[6] http://phantomjs.org

Table 2: Summary of CrimeBB contents: (HF=Hackforums, KM=Kernelmode, OC=Offensive Community, MPGH=Multiplayer Game Hacking)

Forum	Boards	Members	Threads	Posts	Oldest
HF	175	559 671	3 789 274	39 448 526	01/07
KM	16	1 430	3 091	24 885	03/10
OC	63	9 786	11 460	49 426	06/12
MPGH	712	443 188	729 565	8 798 092	12/05

3.8 Non-textual content

Posts in online forums are mostly composed of plain text, but they can also include other content like images, video or attachments. One of our ethical safeguards is to not automatically download multimedia content (which could contain sexual abuse images or videos). However, analyzing legal multimedia content information might be useful for other researchers, if appropriate safeguards are maintained. Thus, CrimeBot detects and annotates the presence of images, videos (in form of iframes), source code snippets, links to other sites, links to other posts within the same thread, and attachments.

4 THE CRIMEBB DATASET

Using CrimeBot over a period of 9 months we have collected data from 4 different communities.[7] Table 2 summarizes the dataset gathered, showing the total number of boards, posts, members, threads, and the oldest post for each site. Figure 2 shows the number of posts and registered members over time. Next, we present some details about the different communities.

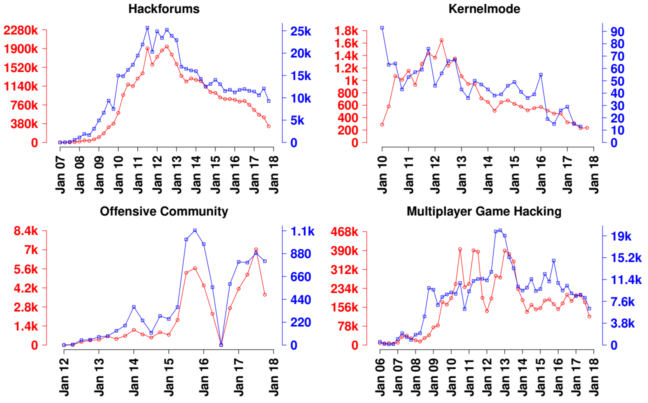

Figure 2: Number of posts (red, left axis) and new members (blue, right axis) per quarter in the forum. The x and y-axes scales vary.

Hackforums (HF)[8] is one of the largest and oldest ongoing hacking communities. Accessible from the surface web, it is indexed by common search engines. It is listed as the top hacking forum in the Alexa ranking, although from Figure 2 we can see that the number of posts and new members is decreasing. As mentioned, this community has been connected to several high profile events, including the release of the Mirai botnet source code [2], and the early activities of an accused who is alleged to have authored banking malware [27]. Table 3 shows the categories into which *hackforums* is divided, as well as the amount of data scraped from each. Most forums (including *hackforums*) show the number of posts and threads per sub-forum. We use this information as a ground-truth to test the coverage of CrimeBot. We can see that all the categories have been crawled entirely and CrimeBot retrieves nearly 100% of the data. This confirms the completeness of the crawler. Hackforums uses the toolkit MyBB.

Offensive Community (OC) is an "underground hacking forum that provides tutorials, latest hacking techniques, free tools and a online teaching to members".[9] It has a similar structure to *hackforums*, but contains special sections that are not permitted in the former, like a cracking sections and a forum for exploits and 0-days. It is poorly moderated and contains a lot of spam. The lack of posts and new members from June to November 2016 suggest that the site was temporary offline, although this is unverified. Offensive Community uses the toolkit MyBB.

Kernelmode (KM) is a community "to discuss rootkits, debugging, reverse-engineering, malware analysis, and other related topics".[10] This is not an underground forum, and its primary intent is to cater for researchers and malware analysts. However, it has an active forum for sharing malware binaries and has a special section for malware-related tools like anti-debuggers or crypters, which can be abused by malicious actors. Thus, although the site operators clearly discourage illegal content, it might be potentially frequented by offenders. Kernelmode uses the toolkit phpBB.

Multiplayer Game Hacking (MPGH) is a community focused on "Game Hacks, Game Cheats and Trainers",[11] with specific sections for different games. It contains an entire section for general hacking in particular and also contains a marketplace, particularly focused on trading accounts and items from virtual games. MPGH uses the toolkit vBulletin.

Limitations: The CrimeBB dataset contains data crawled from May 2017, and thus does not include content that has been removed previously from forums. For example *hackforums* used to have a "Booter Service Bazaar" forum, which was removed due to increased scrutiny shortly after the release of Mirai [25]. While we do not have the complete snapshot of this forum, thank to a previous scrape [35] we were able to include it partially in CrimeBB.

4.1 Data sharing and reproducibility

Reproducibility is an important principle of scientific research, as replicating findings can lead to robust conclusions. Martin and Christin [30] argue that data sharing is important to enable reproducibility, but also for the responsible use of resources required for scraping data, such as the traffic load on anonymity networks. However, due to privacy and ethical concerns, and to prevent misuse by malicious actors, such datasets should not necessarily be publicly released. While the forums themselves are publicly available, and

[7] The crawling process is ongoing and we are updating CrimeBB with new data and more communities

[8] https://hackforums.net/

[9] https://offensivecommunity.net/

[10] http://www.kernelmode.info/forum/

[11] https://www.mpgh.net/

Table 3: Summary of Hackforums in CrimeBB by category. The percentage calculations are affected by the deletion of threads and posts.

Category	Forums	Posts	Oldest	Threads	Coverage
Gaming	32	4 371 268	02/07	424 826	99.82
Web	9	627 484	01/07	87 582	99.27
Money	9	2 006 809	11/07	154 061	96.41
Hack	23	5 869 600	02/07	666 882	96.13
Coding	15	1 470 806	05/07	173 286	99.43
Tech	17	1 799 708	01/07	215 654	99.6
Common	27	12 735 925	01/07	857 006	99.07
Graphics	10	1 025 316	02/07	138 197	99.46
Market	28	9 541 610	11/07	1 071 780	98.9

the forum users are (or should be) aware they are publicly accessible, this data could be used for malicious purposes, for example to deanonymize users based on their posts.

To enable research, both the tool, CrimeBot, and the dataset, CrimeBB, are available to other academic researchers from the Cambridge Cybercrime Centre. Due to ethical concerns and to prevent misuse, before accessing the data researchers are required to sign a data sharing agreement.

5 CASE STUDY

To illustrate the potential of the dataset, we present a case study using CrimeBB, analysing currency exchange patterns. We first describe the evolution of currencies exchanged and current practices. Next, we analyse the types of interests that precede members' engagement in these exchanges.

5.1 Evolution of currency exchanges

Previous work found that members of a Russian forum preferred Webmoney and Western Union for cashing out and transferring money [31]. However, Portnoff et al. demonstrated that Bitcoin and PayPal were the preferred method on a number of English and German forums, including *hackforums* [35]. Both works used datasets updated in 2015. However, cybercrime offenders continuously update their methods and currencies to launder money acquired from illicit activities [36].

We analysed the evolution of currency exchanges performed by members of *hackforums*. This forum contains specific sub-forums for currency exchange. We first apply the tools developed by Portnoff et al. [35] to CrimeBB to extract the currencies being exchanged. We also track the year of each exchange. Figure 3 shows the evolution of the currencies wanted (dashed lines) and offered (solid lines) over the last 8 years.

In support of Portnoff et al.'s research [35], we confirm that Bitcoin and PayPal are by far the two most popular currencies in *hackforums*. However, we are also able to show an increasing level of activity involving Amazon gift cards over the last two years. We have observed same patterns in the currency exchange section of *MPGH*. We also observe that the decrease of Liberty Reserve (which was shut down in May 2013) is correlated with an increase in demand for bitcoin exchanges.

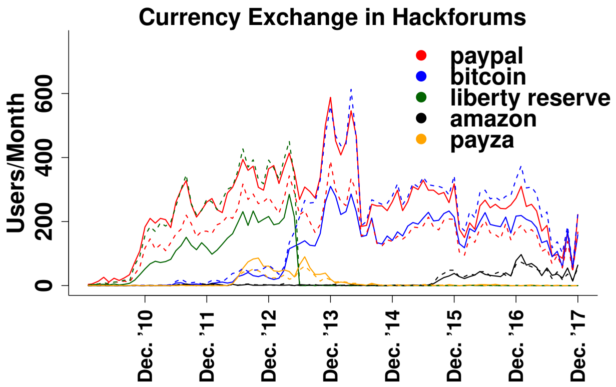

Figure 3: Evolution of the top five currencies wanted (dashed lines) and offered (solid lines) on hackforums

5.2 Pathways for those exchanging currencies

Currency exchange in underground forums is open and every member can make and reply to currency exchange requests. Currency exchange can potentially be used to launder money obtained from criminal activities. Understanding where the money comes from is not straightforward, since users have multiple accounts, or operate in different communities [32]. Indeed, we have detected several user names of *hackforums* members that repeat in other communities. Concretely, 8.3k names with *MPGH*, 1.3k with *Offensive Community* and 171 with *Kernel Mode*. While a portion of these coincidences might be accidental, others might not. Indeed, keeping the same pseudonym is a common practice on underground forums to maintain reputation across communities [19]. In these cases, the money-making and currency exchange cannot be correlated in a direct way. However, in other cases members operate in the same community, both the activities from which they profit, and the currency exchange to cash out the benefits. Thus, we focus on key actors to analyse their evolution across the community to learn about their pathways into currency exchange.

To this end, we collect from CrimeBB information about accounts that had offered currency exchange, and when. Additionally, we gather the registration dates, and the dates where members were last active. For each member, we define the variable *daysUntilCurrency* as the time elapsed between registration time and first post in currency exchange, and the variable *spanCurrency* as the months elapsed since the first and last post in currency exchange. We then select users that: 1) had been active for some time (i.e. $daysUntilCurrency > 720$) before they started posting into currency exchange; and 2) had continuously (i.e. on a monthly basis) asked for currency for more than a year (i.e. $spanCurrency > 12$). In *hackforums*, there are 44 users matching these criteria.

These members are of interest due to the extent of their activities. They have been members of the community for a long period of time, and have eventually engaged in currency exchange. It is possible that they have earned money using techniques learned from the forum. We focus on these members to analyse their pathways, particularly their interests before they started exchanging

Table 4: Sub-forums of the community *hackforums* most frequented by members engaged in currency exchange, including the category (M=Market, C=Common, G=Gaming, $=Money, H=Hacking, P=Coding, T=Technology)

Buyers Bay M	The Lounge C
Virtual Game Items M	Marketplace Discussions M
Secondary Sellers M	Online Accounts M
Premium Sellers M	Crypto Currency $
Shopping Deals M	SQL Injection H
Traders Topics M	Graphics Market M
Visual Basic/.NET P	Beginner Hacking H
Gamertags G	E-Whoring H
Botnets/IRC/Zombies H	Computer Customizing T

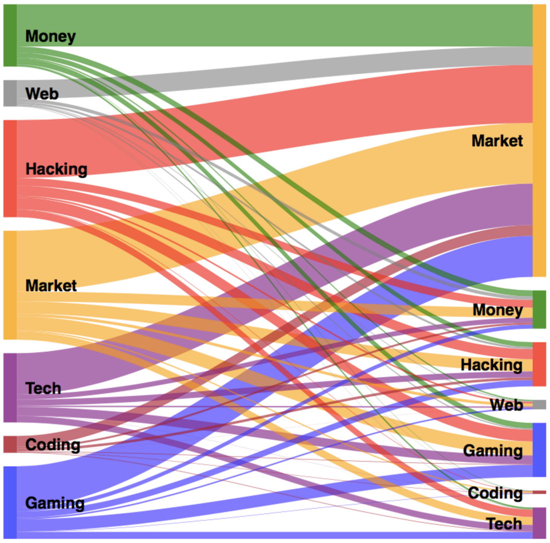

Figure 4: Mapping diagram of forums most frequented by *hackforums* members before (left) and after (right) they start posting about currency exchange.

currencies. Whether these activities account for the source of the currencies being exchanged requires further research.

To categorise common interests, we analyse the sub-forums these members were posting in before and during/after they start posting in the currency exchange sub-forum. The interest of a member M in a sub-forum F is calculated as:

$$\mathrm{I}(M,F) = \mathrm{N_T}(M,F) * 3 + \mathrm{N_P}(M,F)$$

Where $\mathrm{N}_{\{\mathrm{T,P}\}}(M,F)$ denotes the number of {threads, posts} written by M in F. We assign more weight to threads since initiating a thread represents a greater interest than a post. Table 4 lists the most frequented forums (including before and after currency exchange). Most of the forums are market related. Other particularly interesting forums include hacking forums, where botnets and SQL Injection attacks are discussed, as well as e-whoring. We then analyse how interests shifted, comparing their interests while being active in the currency exchange sub-forum, to their interests before this period. A transition of interest from forum F_i to forum F_j is defined as:

$$\mathrm{T}(F_i \rightarrow F_j) = \mathrm{I}(M,F_i) + \mathrm{I}(M,F_j) \leftrightarrow F_i \in \lambda_{\mathrm{B}}(M) \wedge F_j \in \lambda_{\mathrm{A}}(M)$$

λ_B and λ_A represents the set of the 5 top forums where user M is interested before and after starting with currency exchange. Figure 4 shows the transitions of interests aggregated by category. It can be observed that many members start with interest in hacking, gaming or technology, but these interests move to market and money-making forums once they start exchanging currencies.

6 CONCLUSIONS

The growth of new web technologies allows crime to spread through the online world. The proliferation of illicit services such as malware-as-a-service or online booters permit cyberoffences to be easily committed without the need for a deep technical background. Many online communities combine non-malicious topics, such as computer games or technology, with active black hat communities. Thus, users seeking to gain reputation in online communities or curious to explore black hat activities can be attracted by easy money-making methods, which normally implies fraudulent or illicit activity.

Underground forums constitute a rich source of information to understand these behaviours and analyse pathways into crime, and have gained the attention of the research community. However, most research has relied on old or incomplete datasets, and thus results become rapidly outdated. Since cybercrime is an unpredictable discipline that evolves rapidly, researchers require complete and up-to-date datasets. Tools are vital for collecting and maintaining these datasets.

In this work, we present the CrimeBB dataset, which we make available for other academic researchers. The dataset spans more than a decade and contains more than 48m posts made by 1m users in 4 different communities. We collect the data using CrimeBot, a focused crawler designed to update the dataset efficiently and stealthily. The collection is ongoing, and we are updating CrimeBB with data from more communities.

The CrimeBB dataset presents unique opportunities for large-scale analysis of underground forums which would not otherwise be possible. As an example, we presented a case study analysing the evolution of currency exchange. We find that Amazon gift cards are increasingly being exchanged for other types of currencies. Gift cards are a type of alternative currency, and are therefore vulnerable to abuse. Additionally, by analysing one of the largest and longest running hacker forums, we have empirically measured the pathways of those exchanging currencies believed to have been obtained illicitly, confirming that many of them were first interested in the gaming or technology communities.

ACKNOWLEDGMENTS

All authors are supported by the EPSRC [grant number EP/M020320/1]. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of any of the funders.

REFERENCES

[1] Judith Aldridge and David Décary-Hétu. 2015. Not an 'eBay for drugs': The cryptomarket 'Silk Road' as a paradigm shifting criminal innovation. https://ssrn.com/abstract=2436643. (2015). https://doi.org/10.2139/ssrn.2436643

[2] Manos Antonakakis, Tim April, Michael Bailey, Matt Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J. Alex Halderman, Luca Invernizzi, Michalis Kallitsis, Deepak Kumar, Chaz Lever, Zane Ma, Joshua Mason, Damian Menscher, Chad Seaman, Nick Sullivan, Kurt Thomas, and Yi Zhou. 2017. Understanding the Mirai Botnet. In *26th USENIX Security Symposium (USENIX Security)*. USENIX Association, Vancouver, BC, 1093–1110.

[3] Bissan Audeh, Michel Beigbeder, Antoine Zimmermann, Philippe Jaillon, and Cédric Bousquet. 2017. Vigi4Med scraper: A framework for web forum structured data extraction and semantic representation. *PloS one* 12, 1 (2017). https://doi.org/10.1371/journal.pone.0169658

[4] Victor Benjamin, Sagar Samtani, and Hsinchun Chen. 2016. Conducting large-scale analyses of underground hacker communities. *Cybercrime Through an Interdisciplinary Lens* 26 (2016), 56.

[5] Gwern Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard, Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Sohhlz, Delyan Kratunov, Vince Cakic, Van Buskirk, Whom, Michael McKenna, and Sigi Goode. 2015. Dark Net Market archives, 2011-2015. https://www.gwern.net/DNM-archives. (July 2015).

[6] Rui Cai, Jiang-Ming Yang, Wei Lai, Yida Wang, and Lei Zhang. 2008. iRobot: An intelligent crawler for Web forums. In *Proceedings of the 17th international conference on World Wide Web (WWW)*. ACM, 447–456.

[7] Hsinchun Chen, Ahmed Abbasi, Bhavani Thuraisingham, Chris Yang Drexel, Paul Hu, and Resha Shenandoah. 2017. Intelligence and security informatics dataset. http://www.azsecure-data.org. (2017).

[8] Nicolas Christin. 2013. Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd international conference on World Wide Web (WWW)*. ACM, 213–224.

[9] David Décary-Hétu and Judith Aldridge. 2015. Sifting through the net: Monitoring of online offenders by researchers. *European Review of Organised Crime* 2, 2 (2015), 122–141.

[10] David Dittrich, Michael Bailey, and Erin Kenneally. 2013. *Applying ethical principles to information and communication technology research: A companion to the Menlo Report*. Technical Report. U.S. Department of Homeland Security. https://doi.org/10.2139/ssrn.2342036

[11] Benoît Dupont, Anne-Marie Côté, Claire Savine, and David Décary-Hétu. 2016. The ecology of trust among hackers. *Global Crime* 17, 2 (2016), 129–151.

[12] Hanno Fallmann, Gilbert Wondracek, and Christian Platzer. 2010. Covertly probing underground economy marketplaces. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*. Springer, 101–110.

[13] Lorenzo Franceschi-Bicchiera. 2017. For 20 years, this man has survived entirely by hacking online games. https://motherboard.vice.com/en_us/article/59p7qd/this-man-has-survived-by-hacking-mmo-online-games. (July 2017).

[14] Richard Frank. 2017. personal communication. (2017).

[15] Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen. 2010. A focused crawler for Dark Web forums. *Journal of the American Society for Information Science and Technology* 61, 6 (2010), 1213–1231.

[16] Joobin Gharibshah, Tai Ching Li, Maria Solanas Vanrell, Andre Castro, Konstantinos Pelechrinis, Evangelos E Papalexakis, and Michalis Faloutsos. 2017. InferIP: Extracting actionable information from security discussion forums. In *International Conference on Advances in Social Networks Analysis and Mining*. IEEE/ACM.

[17] Andreas Haslebacher, Jeremiah Onaolapo, and Gianluca Stringhini. 2017. All your cards are belong to us: Understanding online carding forums. In *APWG Symposium on Electronic Crime Research (eCrime)*. IEEE. https://doi.org/10.1109/ECRIME.2017.7945053

[18] Thomas J. Holt, Olga Smirnova, and Yi-Ting Chua. 2016. The social organization of actors in stolen data markets. In *Data Thieves in Action*. 73–95. https://doi.org/10.1057/978-1-137-58904-0_4

[19] Thomas J. Holt, Deborah Strumsky, Olga Smirnova, and Max Kilger. 2012. Examining the social networks of malware writers and hackers. *International Journal of Cyber Criminology* 6, 1 (2012), 891.

[20] Alice Hutchings. 2016. Cybercrime trajectories: An integrated theory of initiation, maintenance, and desistance. In *Crime Online: Correlates, Causes, and Context*. Carolina Academic Press, 117–140.

[21] Alice Hutchings and Richard Clayton. 2016. Exploring the provision of online booter services. *Deviant Behavior* 37, 10 (2016), 1163–1178. https://doi.org/10.1080/01639625.2016.1169829

[22] Christos Iliou, George Kalpakis, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2016. Hybrid focused crawling for homemade explosives discovery on surface and dark web. In *11th International Conference on Availability, Reliability and Security (ARES)*. IEEE, 229–234.

[23] Jingtian Jiang, Xinying Song, Nenghai Yu, and Chin-Yew Lin. 2013. Focus: learning to crawl web forums. *IEEE Transactions on Knowledge and Data Engineering* 25, 6 (2013), 1293–1306.

[24] Sheharbano Khattak, David Fifield, Sadia Afroz, Mobin Javed, Srikanth Sundaresan, Damon McCoy, Vern Paxson, and Steven J. Murdoch. 2016. Do you see what I see? Differential treatment of anonymous users. In *Network and Distributed System Security Symposium (NDSS)*.

[25] Brian Krebs. 2016. Hackforums shutters booter service bazaar. https://krebsonsecurity.com/2016/10/hackforums-shutters-booter-service-bazaar/. (October 2016).

[26] Brian Krebs. 2017. Who is Anna-Senpai, the Mirai worm author? https://krebsonsecurity.com/2017/01/who-is-anna-senpai-the-mirai-worm-author/. (January 2017).

[27] Brian Krebs. 2017. Who is Marcus Hutchins? https://krebsonsecurity.com/2017/09/who-is-marcus-hutchins/. (September 2017).

[28] Mitch Macdonald and Richard Frank. 2017. The network structure of malware development, deployment and distribution. *Global Crime* (2017), 1–21. https://doi.org/10.1080/17440572.2016.1227707

[29] Mitch Macdonald, Richard Frank, Joseph Mei, and Bryan Monk. 2015. Identifying digital threats in a hacker web forum. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE/ACM, 926–933.

[30] James Martin and Nicolas Christin. 2016. Ethics in cryptomarket research. *International Journal of Drug Policy* 35 (2016), 84–91.

[31] Alexander Mikhaylov and Richard Frank. 2016. Cards, money and two hacking forums: An analysis of online money laundering schemes. In *European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 80–83.

[32] 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.

[33] 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 *Conference on Intelligence and Security Informatics*. IEEE, 7–12.

[34] British Society of Criminology. 2015. Statement of ethics. (2015). http://www.britsoccrim.org/ethics/

[35] Rebecca S. Portnoff, Sadia Afroz, Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Damon McCoy, Kirill Levchenko, and Vern Paxson. 2017. Tools for automated analysis of cybercriminal markets. In *Proceedings of 26th International World Wide Web conference (WWW)*.

[36] Jean-Loup Richet. 2013. Laundering money online: A review of cybercriminals methods. *arXiv preprint arXiv:1310.2368* (2013).

[37] Sagar Samtani, Ryan Chinn, and Hsinchun Chen. 2015. Exploring hacker assets in underground forums. In *International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 31–36.

[38] Sergey Shekyan. 2015. Detecting PhantomJS based visitors. https://blog.shapesecurity.com/2015/01/22/detecting-phantomjs-based-visitors/. (January 2015).

[39] Rachee Singh, Rishab Nithyanand, Sadia Afroz, Paul Pearce, Michael Carl Tschantz, Phillipa Gill, and Vern Paxson. 2017. Characterizing the nature and dynamics of Tor exit blocking. In *26th USENIX Security Symposium (USENIX Security)*. USENIX Association, Vancouver, BC, 325–341.

[40] Kyle Soska and Nicolas Christin. 2015. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem.. In *USENIX Security Symposium (USENIX Security)*.

[41] Daniel R. Thomas, Richard Clayton, and Alastair R. Beresford. 2017. 1000 days of UDP amplification DDoS attacks. In *APWG Symposium on Electronic Crime Research (eCrime)*. IEEE. https://doi.org/10.1109/ECRIME.2017.7945057

[42] Daniel R. Thomas, Sergio Pastrana, Alice Hutchings, Richard Clayton, and Alastair R. Beresford. 2017. Ethical issues in research using datasets of illicit origin. In *Proceedings of the Internet Measurement Conference (IMC)*. ACM. https://doi.org/10.1145/3131365.3131389

[43] Bryce Westlake, Martin Bouchard, and Richard Frank. 2015. Assessing the validity of automated webcrawlers as data collection tools to investigate online child sexual exploitation. *Sexual abuse: A journal of research and treatment* (2015), 685–708. https://doi.org/10.1177/1079063215616818

[44] Ellen Whiteman. 2007. "Just Chatting": research ethics and cyberspace. *International Journal of Qualitative Methods* 6, 2 (2007), 95–105. https://doi.org/10.1177/160940690700600209

[45] Philipp Winter, Richard Köwer, Martin Mulazzani, Markus Huber, Sebastian Schrittwieser, Stefan Lindskog, and Edgar Weippl. 2014. Spoiled onions: Exposing malicious Tor exit relays. In *Privacy Enhancing Technologies Symposium (PETS)*. Springer.

[46] Built With. 2018. Forum software usage. https://perma.cc/XUX9-HKV8. (Feb 2018).