

Identifying, Collecting, and Presenting Hacker Community Data: Forums, IRC, Carding Shops, and DNMs

Po-Yi Du, Ning Zhang, Mohammedreza Ebrahimi, Sagar Samtani, Ben Lazarine, Nolan Arnold, Rachael Dunn, Sandeep Suntwal, Guadalupe Angeles, Robert Schweitzer, Hsinchun Chen
Department of Management Information Systems, The University of Arizona
Tucson, AZ 85721, USA
{pydu, zhangning, ebrahimi, sagars, benlazarine, nolanarnold, rajacobi, sandeepsuntwal, angeles, rschweitzer}@email.arizona.edu, hchen@eller.arizona.edu

Abstract— Cyber-attacks cost the global economy over \$450 billion annually. To combat this issue, researchers and practitioners put enormous efforts into developing Cyber Threat Intelligence, or the process of identifying emerging threats and key hackers. However, the reliance on internal network data has resulted in inherently reactive intelligence. CTI experts have urged the importance of proactively studying the large, ever-evolving online hacker community. Despite their CTI value, collecting data from hacker community platforms is a non-trivial task. In this paper, we summarize our efforts in systematically identifying and automatically collecting a large-scale of hacker forums, carding shops, Internet-Relay-Chat, and Dark Net Marketplaces. We also present our efforts to provide this data to the larger CTI community via the AZSecure Hacker Assets Portal (www.azsecure-hap.com). With our methodology, we collected 102 platforms for a total of 43,981,647 records. To the best of our knowledge, this compilation of hacker community data is the largest such collection in academia.

Keywords— *Hacker community data collection, Hacker forums, Internet-Relay-Chat, Dark Net Marketplaces, Carding Shops*

I. INTRODUCTION

With computer and information technology becoming more ubiquitous, cybersecurity has become a grand societal challenge. Today, malicious hackers commit numerous large-scale, advanced attacks on industry and government organizations. These cyber-attacks cost the global economy over \$450 billion annually [1]. Cyber Threat Intelligence (CTI), or the process of identifying emerging threats and key threat actors (i.e., hackers) to enable effective cybersecurity decisions, has emerged as a viable approach to mitigate this concern.

Fundamentally a data-driven process, CTI has traditionally focused on collecting data from internal network devices databases, IDS/IPS, routers, workstations, and others. Collected data is analyzed using malware analysis, forensics, event correlation, and other well-established methods. Despite the prevalence of these approaches, CTI experts from major cybersecurity firms, note that the reliance on data from past events results in inherently reactive intelligence [2]. As a result, cyber-attacks remain an unfortunate uptick.

To combat this issue, CTI experts have urged the importance of developing proactive CTI by directly investigating hackers within the online hacker community [2][3]. The international online hacker community attracts and motivates millions of hackers from the US, Russia, and China, to share or sell hacking tools, knowledge, and other illegal products and services [4]. Today, four major hacker community platforms exist: hacker forums, Internet-Relay-Chat (IRC), carding shops, and Dark Net Marketplaces (DNMs). Exploits found on these platforms have executed well-publicized attacks such as the BlackPOS malware for the Target breach or the Mirai botnet for the internet-scale Denial-of-Service (DoS) attack in 2016.

Collecting data from each of these platforms is a non-trivial task. Hacker community platforms carefully conceal themselves and employ numerous anti-crawling measures that prevent automated, large-scale data collection. These barriers force many researchers to manual collection efforts. Studies attempting automated collection are limited to one platform type. In this paper, we summarize our work in identifying and automatically collecting a large collection of hacker forum, carding shop, IRC, and DNM data. We also present our efforts to provide this data to the larger CTI community via the AZSecure Hacker Assets Portal. To the best of our knowledge, this collection of hacker community data is the largest in academia. Consequently, it can enable a multitude of novel and valuable proactive CTI research inquiries.

The remainder of this paper is organized as follows. First, we review each platform, discuss their CTI value, and note existing data collection strategies. Section III details our platform identification and collection methodology. Section IV summarizes our collected data and highlights promising CTI research directions. Section V illustrates key functions of the Hacker Asset Portal. Section VI concludes this work.

II. HACKER COMMUNITY PLATFORM REVIEW

A. Hacker Community Platforms Overview

To the best of our knowledge, four hacker community platforms exist: (1) forums, (2) IRC, (3) DNMs, and (4) carding

shops. Table I describes each platform and their CTI value.

TABLE I. HACKER COMMUNITY PLATFORM SUMMARY

Platform	Description	CTI Value
Hacker Forums	Message board allowing members to post messages that are archived	Key threat actor identification; sharing of hacking tools; indication of access to other hacker communities
IRC	Plain-text, instant messaging, communication that is not archived	Sharing of hacking knowledge and potential target; indication of access to other hacker communities
DNMs	Markets on Tor that sell illicit goods via cryptocurrency	Early indicator for breached companies; in-depth understanding of underground economy
Carding Shops	Shops selling stolen credit/debit cards and sensitive data	Monitoring trafficking of internet fraud industry; precaution of breaches before happen

The four hacker community platforms create an ecosystem of hacker activities. Hackers use forums and/or IRC to freely discuss and share Tools, Techniques, and Processes (TTP) and advertise hacking services. Hackers freely download these tools or navigate to DNMs to purchase exploits. These tools help hackers conduct cyber-attacks to attain sensitive data such as credit/debit cards, Social Security Numbers (SSN), and logins to sell on DNMs and/or carding shops for financial gain. Each platform is further detailed in the following sub-sections.

1) Hacker Forums

Hacker forums are the most common and largest platforms for hackers to share hacking resources [4]. Hackers use these message boards to post messages within threads of conversations related to hacking tools, techniques, and malicious source code. Among the four major platforms, forums are the only one allowing hackers to post malicious exploits for others to freely download [5]. Figure 1 illustrates an example of a hacker sharing ransomware. Sharing hacking assets enables individuals with limited hacking skills to become capable of conducting cyberattacks [4]. This characteristic, combined with the rich metadata (e.g., post content, post date) make forums a viable data source for monitoring the TTP of hackers, identifying key actors, and discovering emerging threats.

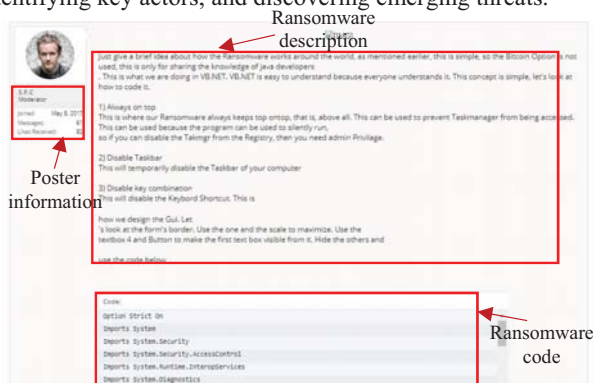


Figure 1. An example of a hacker forum member sharing ransomware code

2) Internet-Relay-Chat (IRC)

Built on a separate protocol, an IRC server can hold multiple channels, containing conversations about pre-defined topics [4]. Although not originally intended for hackers, IRC channels have become a popular medium for hacktivist groups to share hacking knowledge. Figures 2 and 3 illustrate two examples of user behaviors on hacker IRC. Figure 2 depicts hackers sharing links to forums with illegal contents. Figure 3 illustrates an IRC user demanding hacking service with a provided target IP.

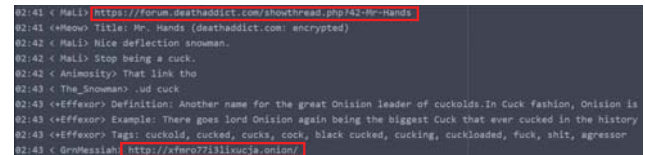


Figure 2. An example of hackers sharing links containing illegal contents

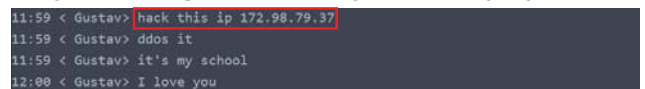


Figure 3. An example of an IRC user demanding hacker service

Unlike forums, IRC conversations are not archived and must be collected in real-time. Additionally, IRC messages are broadcast to all channel participants. If a user loses server connection, he/she cannot retrieve the conversation for that time period [6]. This allows hackers to share hacking knowledge and targets more freely. As a result, collecting IRC data can help understand hacker behaviors, targets, and emerging threats.

3) Dark Net Marketplaces (DNMs)

DNMs operate as Tor hidden services, and are one of the major platforms on which users make illegal transactions with cryptocurrency [7]. Products such as drugs, hacking tools, weapons, and stolen personal documents can easily be found on DNM. Access to DNMs are often indicated by other hacker communities such as hacker forums and IRC channels [4].

Two categories of DNM data exist: product and seller. Each product listing contains a name, description, price, delivery destinations, and product category. Seller information includes seller history, ID, rating and trust level, PGP keys, etc. Figure 4 illustrates an example of a product listing page on DNM, where a scam page of PayPal to conduct phishing attacks is sold.

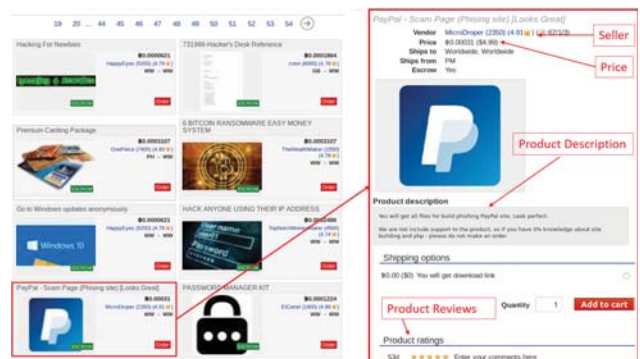


Figure 4. An example of a product listing page on DNM

Data and information stolen from breached companies are often sold on DNM. Thus, DNMs can serve as an early indicator for breached organizations. Also, past research indicate the thriving of DNM has raised concern in public health and law enforcement, for holding an abundant amount of drug listing [8][9]. Researchers observed that DNM users often share information about reliable seller and quality goods among DNM [9][10]. Moreover, DNM has been leveraged to explore the product distribution network [11]. While Tor’s untraceable nature makes linking DNM users to their true identity is a non-trivial task, studying their behavior can provide an in-depth understanding about the underground economy.

4) Carding Shops

Carding shops facilitate many underground economy activities by providing high quality carding services [4]. A large amount of stolen card data is traded and diffused through carding shops [12]. Since carding data are duplicable and recyclable, the rapid dissemination of card information can inflict significant financial losses on cardholders. Carding shop data can be divided into payment card data, identity data, and credential data [13]. Payment card data is the major product type in carding shop, and it can be further categorized into “Dumps” and “Fullz” based on the amount of information carried by the product. Dumps refer to the raw information retrieved from the magnetic stripe of the card. Fullz includes the full information of the victim, including name, address. Both Dumps and Fullz contain three sections: card information, source, and price. Beyond these two, SSNs and logins are also commonly found on carding shops.

Carding shops have unique CTI value as they provide a comprehensive view of carding fraud. Despite its importance, little academic literature exists about them [4]. Past researchers have analyzed relationships between each attribute and price by comparison to identify that card data is packaged, with a label, to periodically release on carding shop [12]. Despite this useful discovery, significant potential remains for carding shop data to be used as a source of identifying exploited individuals and financial institutions.

B. Hacker Community Data Collection: Existing Approaches and Challenges

Unlike traditional social media sites, researchers face numerous issues when collecting hacker community platforms. Web-based platforms (i.e., forums, shops, DNMs) employ anti-crawling measures such as drive-by malware, session timers, user-agent checks, CAPTCHA, and others. IRC data must be collected in real-time. These challenges have limited many researchers to manual collection efforts, or to studying old datasets (e.g., dumped SilkRoad, archived forums). While still valuable, such procedures result in incomplete and/or dated CTI insights. Studies employing automated collection procedures are often limited to one platform type (e.g., forums). However, each hacker community platform is interconnected with others. Thus, the focus on collecting one platform prevents

comprehensive CTI development. These issues motivate the development of large-scale, automated crawling approaches.

III. COLLECTION METHODOLOGY

We developed a systematic approach to gather a comprehensive collection of hacker community data. The process has three phases, platform identification, enhanced automated collection, and content parsing. We summarize each in the following sub-sections.

A. Hacker Community Platform Identification

The first stage in any data collection task is identifying the appropriate data sources. We use three approaches to identify hacker platforms: suggestions from cybersecurity experts, surface web and Tor search engines, and snowball identification. Using all three ensure a comprehensive, high-quality coverage. Irrespective of approach, we only collected platforms containing significant amounts of cybersecurity content. We deliberately avoided platforms specializing in weapons or pornography, as such content has minimal CTI value.

In the first strategy, our team consulted with the National Cyber-Forensics Training Alliance (NCFTA), a major non-profit organization focusing on the CTI sharing across private, public, and academic sectors, and Policing in Cyberspace (POLCYB), an internationally recognized law enforcement entity. Both suggested platforms providing valuable cybersecurity data and also recommended keywords as input for surface web and Tor-based (e.g., Grams) search engines to identify additional platforms. Since hackers often information on traditional social media platforms (e.g., Twitter, Facebook, YouTube), these keywords were also inputted into these sites to identify additional platforms. Figure 5 depicts a YouTube video of Anonymous recruiting for their IRC channel, #OpTestet.



[ENG VERSION] Anonymous ▶ #OpTestet ZAD Partout / Barrage de Sivens - TESTET

Figure 5. An example of a recruiting video of Anonymous on YouTube

The platforms identified from the first two approaches were used as “seeds” for our final strategy: snowball identification. Hackers within these platforms often post links to other platforms. We followed these links and identified if they contained valuable cybersecurity content. The newly identified platforms were used as new seeds to identify additional platforms.

B. Enhanced Automated Collection Procedures

Collection processes for hacker community data vary. Past web-based platforms (i.e., forums, DNMs, carding shops) research usually conducted undirected web crawlers to collect the raw data in the HTML format. To address the anti-crawling challenges detailed in our review, we upgraded our crawler to directly collect web pages and contents. By flexibly switching the types of HTTP request, we significantly reduced the time cost on crawling web-based hacker platform data. For IRC data, we employed two “bots,” similar to fake users, inside each channel, and used these bots to log in at their own routines, to avoid automatic disconnections.

C. Content Parsing

After data collection, the collected raw data requires further parsing to enable subsequent analytics. For forums, DNMs, carding shops, and IRC, parsing entails recognizing text patterns containing relevant attributes (e.g., post date, product description). We developed several custom parsers and leveraged Regular Expression (RegEx) to retrieve information from those platforms, and stored them into a relational database.

IV. DATA COLLECTION OVERVIEW

In our hacker community data collection, we collected 51 hacker forums, 13 IRC channels, 12 DNMs, and 11 carding shops. Table II summarizes our collection.

TABLE II. HACKER COMMUNITY DATA COLLECTION SUMMARY

Platform	# of Platform	# of Records	Languages
Forums	51 forums	32,266,852 posts	English/ Russian/ Arabic
IRC	13 channels	2,791,120 lines of conversation	English
DNM	12 markets	249,597 listings	English/ Russian/ French
Carding Shops	26 shops	8,674,078 listings	English

A. Hacker Forums

We focused on collecting 51 hacking oriented forums containing 32,266,852 posts in 2,961,363 threads in English, Russian, and Arabic. Generally, we observed a high frequency of hacking/security tools, for instance, online shopping site receipt generators for phishing purpose. Some forums specialize in other services such as breached data, mobile malware, cryptocurrencies, login dumps, and code for AI bots. The multilingual feature of our collection can facilitate cybersecurity research in cross-countries comparison. The prolific nature of forum as well as their dynamic and time-sensitive property enables the researcher to identify *trends* of cyber threats easier and sometimes earlier than other platforms. Hence, a promising research direction would be developing time-sensitive methods to analyze the trends of cyber threat landscape through constant monitoring of the forum data. Another direction would be cross-referencing the forum data with DNMs in order to have holistic

trend analysis. Moreover, due to the interactive structure of these platforms, they are capable of revealing the interaction network of cyber criminals.

B. IRC

We collected 2,791,120 lines of conversation from 13 IRC cybersecurity specific IRC channels between 9/2016 and 1/2018. The data collection is summarized in Table III. For space consideration, we only listed the top six channels.

TABLE III. IRC DATA COLLECTION SUMMARY

Channel	# of lines	Description
#anonops	1,696,024	General discussion of hacking-related topics
#ed	574,024	Discussion about current topics
#hackers	174,328	General discussion of tips and tricks for Anonymous hackers
#Evilzone	163,402	Casual discussions on cyber security
#ddos	23,172	Posts about current ddos tools recommended by Anonymous hackers
#tutorials	77,903	Offers selected members tutorials through a separate interactive IRC channel
Total (of all channels)	2,791,120	-

The most popular IRC channel “anonops” is the main channel of the famous hacker group, Anonymous. Anonymous also runs channels such as “ddos,” which focuses on Distributed Denial-of-Service (DDoS), and “hacker,” in which users share and ask for hacking tips. IRC users also demand/provide hacking services with target information to each other. While past studies have explored the IRC participant duration [14], the CTI value of IRC data is still undiscovered. The links, URLs, and named entities exchanged in IRC chatrooms can be used in a snowball sampling manner to expand cyber threat resources. After identifying resources, a promising research direction would be discovering conversation topics via topic modeling approaches (e.g., Latent Dirichlet Allocation). Techniques such as Named Entity Recognition (NER) and relationship extraction can identify the targets of hackers and hackers in IRC.

C. DNM

We collected 12 DNMs between September, 2016 and January, 2018. Table IV summarizes our DNM data collection.

TABLE IV. DNM DATA COLLECTION SUMMARY

DNM	# of listing	# of security listing	Language
0day	28,330	28,330+	English
Alphabay	25,118	N/A	English
Apple Market	2,012	N/A	English
Dream Market	120,962	1,916+	English
French Deep Web	1,536	134+	French
Hansa	14,149	N/A	English
Minerva	166	N/A	English
Russian Silk Road	488	N/A	Russian

SilkRoad3	1,798	70+	English
TradeRoute	35,504	547+	English
TOCHKA	1,958	300+	Russian/English
Valhalla	17,576	695+	English
Total:	249,597	31,992+	English/Russian/ French/Dutch

All DNMs, except for 0day, contain illicit product not limited to cybersecurity. Among these DNMs, 60% of the products are drug related. Cybersecurity-related products account for around 20%, while the remainder are weapon and stolen personal document. 0day data, on the other hand, contains only exploits. Such listings display information about the exploit category, description, the platform affected the exploit, and severity level. Most of the listings were originally priced. Once a patch or fix for the exploit comes out, the listing becomes free. In our collection, only 37 out of 28,330 exploits are priced, and the rest are free. Figure 6 is an example of 0day exploit listing.

Windows 10 RCE (Sandbox Escape/Bypass ASLR/Bypass DEP) 0day Exploit	
Full title	Windows 10 RCE (Sandbox Escape/Bypass ASLR/Bypass DEP) 0day Exploit [Highlight]
Date add	23-08-2017
Category	remote exploits
Platform	windows
Verified	✓
Price	0.369 BTC 6,000 USD
Risk	High [Security Risk Critical]
Rel. releases	R
Description	<p>1. Affected OS: Windows 10 x86 x64</p> <p>2. Vulnerable Target application versions and reliability. If 32 bit only, is 64 bit vulnerable? The vulnerability is present in the 32-bit and 64-bit versions of Windows 10 (1507, 1511, 1607, 1703). With this vulnerability, you can remote code execute in the target system via any browser.</p> <p>3. Tested, functional against target application versions, list complete point release range: Windows 10 x86 & x64 (1507, 1511, 1607, 1703), Google Chrome 58.0.3029.110, Mozilla Firefox 53.0.3, Opera 43.</p> <p>4. Does this exploit affect the current target version? <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p>

Figure 6. An example of a 0day exploit listing

Since DNMs are relatively new compared to other platforms, there is more untapped research opportunities in this area. The study by Nunes et al. (2016) is one of the few studies targeting this area in which 11,992 DNM products were collected and semi-supervised classification were employed to find the hacking-related products. Based on the collected DNM data we observe that there is a rise in the number of non-English marketplaces. More specifically, multilingual identification of threats across marketplaces with different languages such as Russian and French would add a global insight of threats. Cross-platform studies can be expanded beyond analyzing DNMs with different languages. We found some overlap between cyber threats being advertised in DNMs and forums. Studying the supply chain aspects of the threats between these two platforms is another possible promising research area that can provide insights about the dissemination patterns and flow of the cybersecurity related products in the span of time.

D. Carding Shops

From 5/2014 to 1/2018, we collected 26 carding shops data with 8,674,078 listing. The majority are credit card dumps

(6,596,093 listings), followed by Fullz and SSNs at 1,999,251 and 78,734 respectively. Table V summarizes our partial collection. For space considerations, we only listed the top nine shops and BuySSN, the only shop with personal information.

TABLE V. CARDING SHOPS DATA COLLECTION SUMMARY

Name	# of CVV and Fullz	# of Dumps	# of SSN	Total
Jokers Stash	566,600	446,642	0	1,013,242
DUMPS MANIA	42,311	777,634	0	819,945
Buybest	9,008	790,325	0	799,333
United Dumps	63,575	734,495	0	798,070
THE MONEY TEAM	20,462	750,642	0	771,104
EBIN CC	13,786	752,441	0	766,227
Golden Shop	19,732	727,888	0	747,620
Getcc Shop	609,770	0	0	609,770
BuySSN	0	0	78,734	78,734
Total (of all shops)	1,999,251	6,596,093	78,734	8,674,078

Nine shops focus on selling payment card information. Six of them, such as Getcc Shop and Diamond Dumps, contain only Fullz data. Generally, price for payment card data greatly varies, and similar card information is frequently found across different shops. BuySSN (78,734 records), specializes in selling SSNs. Such information is capable of causing serious personal losses. One promising CTI direction is identifying customers whose data have been breached. That is, having both names and zip code of each stolen card, would help identify the customer with an acceptable precision which is useful for credit companies and law enforcement agencies. Given that both DNMs and carding shops share credit card information, cross-referencing the stolen data on carding shops with DNMs provides new insights about the dissemination patterns of breached data in the ecosystem.

V. INTEGRATION INTO THE HACKER ASSETS PORTAL

A key aspect of the CTI process is presenting relevant collected data and selected analysis for user consumption. Adhering to key requirement for effective CTI, we integrate collected data into one of our projects, the Hacker Assets Portal (www.azsecure-hap.com). Funded by the National Science Foundation, the portal was designed to provide selected hacker community content to enhance education such that a specific group of CTI professionals can develop proactive CTI measures [15]. Initially hosting only forum data, the Portal has now includes all relevant records from DNMs and carding shops. Users can search, sort, and browse through content based on each platform's metadata. For example, users interested in identifying whether their credit/debit card was stolen can search their name. Should it appear, they can pinpoint which shop it is sold in, the card price, and others. Beyond these core functionalities, we developed visualizations for users to interactively explore the data. Carefully constructed based on

