DICE-E: A Framework for Conducting Darknet Identification, Collection, Evaluation, with Ethics

Victor Benjamin
Arizona State University
W.P. Carey School of Business
Victor.Benjamin@asu.edu

Joseph Valacich
University of Arizona
Eller College of Management
Valacich@arizona.edu

Hsinchun Chen
University of Arizona
Eller College of Management
hchen@eller.arizona.edu

ABSTRACT

Society's growing dependence on computers and information technologies has been matched by an escalation of the frequency and sophistication of cyber-attacks committed by criminals operating from the Darknet. As a result, security researchers have taken interest in scrutinizing the Darknet and other underground web communities to develop better understanding of cybercriminals and emerging threats. However, many scholars lack the capabilities or expertise to operationalize Darknet research and are thus unable to contribute to this increasingly impactful body of literature. This article introduces a framework for guiding such research, called Darknet Identification, Collection, Evaluation, with Ethics (DICE-E). The DICE-E framework provides a focused reference point and detailed guidelines for scholars wishing to become active in the Darknet research stream. Four steps to conducting Darknet forum research are outlined: (1) identification of Darknet data sources, (2) data collection strategies, (3) evaluation of Darknet data, and (4) ethical concerns related to Darknet research. To illustrate how DICE-E can be utilized, an example empirical study is reported. This exemplar illustrates how DICE-E can guide scholars through key decision points when attempting to incorporate the Darknet within their research.

Keywords: Darknet, Deepnet, Cybersecurity, Cybercriminal, Research Framework, Ethics

Author Biographies

VICTOR BENJAMIN

Victor Benjamin is an assistant professor and co-director of the Actionable Analytics Lab in the Department of Information Systems at ASU's W. P. Carey School of Business. His most recent works heavily on cyber-threat intelligence, Darknet community analytics, and detection of bots that manipulate social media conversations. The contributions of his work include detection of emerging cyber threats, improved cybercriminal attribution, understanding of the global cybercriminal supply chain, and more. Beyond the context of cybersecurity, Dr. Benjamin is interested in pursuing other domains of high societal impact while also contributing to advancement of the latest natural language processing (NLP) and machine learning (ML) methodologies. Recently, he has made significant contributions to shallow-layer neural networks used to generate state-of-the-art word embeddings. Dr. Benjamin developed a novel method for representing temporal attributes of data during model learning. Further, he introduced a new technique to boost SGNS model learning by incorporating information from existing knowledgebase.

Dr. Benjamin received his PhD in management information systems — with a PhD minor in linguistics — from the Eller College of Management at the University of Arizona. He is involved with numerous professional and service-oriented activities within the information systems field. He has recently served as a reviewer for *ACM Transactions on Management Information Systems*, *Communications of the International Information Management Association*, *Journal of the American Society for Information Science and Technology*, *IEEE Intelligent Systems*, *IEEE*

Transactions on Human-Machine Systems, *INFORMS Journal on Computing*, and the International Conference on Information Systems.

JOSEPH VALACICH

Joseph (Joe) S. Valacich is an *Eller Professor of MIS* within the Eller College of Management at the University of Arizona, a Fellow of the Association for Information Systems (2009), and the Chief Science Officer (CSO) of Neuro-ID, Inc. His primary research interests include deception detection, human-computer interaction, data visualization, cyber security, and e-business. Dr. Valacich is a prolific scholar, publishing more than 200 scholarly articles in numerous prestigious journals and conferences, including: *Academy of Management Journal, Communications of the ACM, Decision Sciences, Information Systems Research, Journal of Applied Psychology, Journal of the AIS, Journal of MIS, MIS Quarterly, Management Science, Organizational Behavior and Human Decision Processes*, and many others. His scholarly work has had a tremendous impact not only on the IS field, but also on a number of other disciplines, including computer science, cognitive and social psychology, marketing, and management. In March 2018, Google Scholar lists his citation counts as 22,000, with an H-index of 68. He was the general conference co-chair for the 2003 International Conference on Information Systems (ICIS) and the 2012 Americas Conference on Information Systems (AMCIS); both were held in Seattle. He is the Honorary Chair for the 2021 ICIS conference to be held in Austin, Texas.

HSINCHUN CHEN

Hsinchun Chen, Ph.D.; Arizona Regents' Professor, Thomas R. Brown Chair Professor in Management and Technology, University of Arizona; Director, Artificial Intelligence Lab;

Fellow, ACM, IEEE, and AAAS. Dr. Hsinchun Chen graduated with BS degree at the National Chiao-Tong University (Taiwan), MBA at SUNY Buffalo, and MS and Ph.D. at New York University. He is the University of Arizona Regents' Professor and Thomas R. Brown Chair Professor in Management and Technology. He is also a Fellow of ACM, IEEE and AAAS. Dr. Chen recently served as the lead Program Director of the Smart and Connected (SCH) Program at the NSF (2014-2015), a multi-year multi-agency health IT research program of USA. He is author/editor of 20 books, 300 SCI journal articles, and 200 refereed conference articles covering digital library, data/text/web mining, business analytics, security informatics, and health informatics. His overall h-index is 91 (25,000 citations for 900 papers according to Google Scholar), among the highest in MIS and top 50 in computer science. Dr. Chen founded the Artificial Intelligence Lab at the University of Arizona in 1989, which has received more than $50M in research funding from NSF, NIH, NLM, DOD, DOJ, CIA, DHS, and other agencies (100 grants, 50 from NSF). He has served as Editor-in-Chief of major ACM/IEEE, and Springer journals and conference/program chair of major ACM/IEEE/MIS conferences in digital library, information systems, security informatics, and health informatics. He is also a successful IT entrepreneur. His COPLINK/i2 system for security analytics was commercialized in 2000 and acquired by IBM as its leading government analytics product in 2011. Dr. Chen has served as an advisor to major federal research programs and was a Scientific Counselor of the National Library of Medicine (USA), National Library of China, and Academia Sinica (Taiwan). He is a visiting chair professor at several major universities in China (Tsinghua University) and Taiwan (National Taiwan University). He is internationally renowned for leading the research and development in the health analytics (data and text mining; health big data; DiabeticLink and SilverLink) and security informatics (counter terrorism and cyber security analytics; security big

data; COPLINK, Dark Web, Hacker Web, and AZSecure) communities. His recent research includes SilverLink for mobile health and AZSecure for advanced cyber threat intelligence. For more information, see: http://ai.arizona.edu/hchen

DICE-E: A Framework for Conducting Darknet Identification, Collection, Evaluation, with Ethics

INTRODUCTION

Cybersecurity has become an imperative societal problem with widespread implications for the public, industrial, and governmental sectors. Critical infrastructures and complex systems have become increasingly reliant on computing technologies that are threatened by cyber-attacks. Global markets lose an estimated $445 billion a year to cyber-based fraud and intellectual property theft (Graham, 2017). Overall, cybercrime is becoming an increasingly common part of daily life.

The availability of sophisticated technologies and methods for committing cybercrime has grown considerably. Many different cybercriminal assets can be found freely accessible online including hacking tools, malware, source code examples, tutorials, and more (Holt et al., 2012; Benjamin & Chen, 2013). The proliferation of such assets has enabled many lesser-skilled Internet miscreants to conduct advanced cybercriminal operations that may cause disruption and financial loss. The increased reliance on cyber infrastructure, as well as an ever-increasing number of threats, presents challenging problems for researchers, practitioners, and society.

As a result, there is growing interest in advancing current cybersecurity capabilities. In particular, a report by the National Science and Technology Council (NSTC) outlined a critical need to develop advanced methods for modeling cybercriminals (NSTC, 2011). Such research

could result in deeper knowledge of cybercriminal behaviors, the cybercriminal supply chain, emerging vulnerabilities, and so on. Since the release of the NSTC report, the National Science Foundation, Department of Homeland Security, DARPA, and other agencies have sponsored numerous funding opportunities that target extant gaps in cybersecurity literature, many of which can be addressed by IS scholars. In fact, there are many within the IS community that already have expressed interest in security-related topics, such as safe computing practices and privacy (Anderson & Agarwal, 2010; Sutanto et al., 2014; Jenkins et al., 2016). However, one major research gap that IS scholars have currently left unexplored is the investigation of cybercriminals to enrich our understanding of how to effectively combat cybercrime (Mahmood et al., 2010; Chen et al., 2012).

Today, little work has been reported on large-scale identification, collection, and analysis of cybercriminal-generated data that can be used to inform threat intelligence, attribution, target identification, and more. Within the Darknet exist many web forums operated by cybercriminals, but these data sources have largely gone untapped despite their ability to provide rich data. Though the lack of research seems paradoxical given the high societal impact of cybersecurity, this shortcoming may be explained by understanding that Darknet forums greatly differ from traditional online communities. There exist several unique challenges that necessitate new data identification and collection processes.

First, researchers that are interested in conducting research within this area may not know how or where to begin searching for cybercriminal-generated data. Darknet forums take great care to minimize and obfuscate their online presence to avoid surveillance and attracting the attention of law enforcement (Martin, 2013). Anonymity is a key safeguard to avoid real-world legal, ethical, and financial repercussions. Further, the social structure of Darknet forums can

differ greatly from more traditional communities, presenting additional challenge for researchers wishing to navigate this space. For example, many forums host secret subcommunities where only some participants are invited (Motoyama et al., 2011). Such characteristics break from expected behaviors observed from more traditional online communities, impeding efforts by researchers to comprehensively identify Darknet data.

Second, cybercriminal-generated data is much more difficult to collect than traditional web data. Many Darknet forums employ sophisticated anti-crawling mechanisms that make comprehensive automated data collection difficult (Benjamin et al., 2015a). For example, several forums have implemented "drive-by exploits" where JavaScript-based malware is implanted within webpages and executed against vulnerable web browsers to exploit unsuspecting visitors, including researchers. Similar with data identification challenges, collection efforts are quickly complicated by many technical hurdles It is important to recall the motivation of Darknet participants, and their desire to limit the accessibility and archiving of any criminally-related data. Such issues increase the barrier of entry for conducting this type of research.

Lastly, given the non-traditional and often illegal nature of cybercriminal data and cybercriminal community contents, it is necessary to establish research guidelines that can be closely followed by academics wishing to conduct their own explorations in this space. Even the most experienced online community researchers may lack preparedness to handle the potential ethics and legal concerns that can be encountered within the Darknet. This issue is exacerbated by the fact that there is currently no single comprehensive source that can help guide researchers and university administrators when attempting to address such concerns. Thus, a comprehensive roadmap for developing Darknet research projects with considering for the numerous ethical concerns would thus be of great importance.

This paper is organized into the following sections. First, a background on cybercriminal and Darknet forum research is provided. This section compares Darknet forum research with more traditional virtual community work and also highlights recent studies of relevance. Next, details are provided concerning how to operationalize a cybercriminal forum research project. This section includes methods to identify cybercriminal and Darknet forums, as well as techniques to automatically collect contents of such forums. A discussion of potential analytical directions is provided. Also, various ethical issues regarding Darknet research are explored. Next, a case example of cybercriminal forum research is presented to demonstrate the flavor of research that can be undertaken by following the guidelines described in this paper. Lastly, the contributions of this work are discussed. In sum, this paper provides a comprehensive roadmap for researchers to successfully conduct cybercriminal and Darknet forum research.

BACKGROUND

While the Darknet is not a traditional topic found in business literature, it is now of critical importance as cyber-based threats have become a significant factor that may cause unforeseen, large-scale disruption of business operations and continuity. Cyber-threat intelligence research is of relevance to both business researchers and security practitioners, as it can provide new and enhanced capabilities for detecting emerging threats, potential targets of future attacks, victims of existing attacks, and so on. The value of such research is exemplified in Figure 1, a recruitment advertisement for the hacktivist campaign called *Operation Green Rights*. The hacktivists in this example were targeting companies accused of causing vast environmental damage. Specific targets are declared in the hacktivist advertisement, and hyperlinks are provided to download data dumps of stolen employee e-mail account and password information from the targeted organization. This threat was a hacktivist campaign first detected within

Darknet communities and is just one case of many similar recruitment messages that exist as blogs, images such as this one, and video. Similarly, with breaches of major firms such as Equifax, it would be useful for credit monitoring firms to survey the Darknet for stolen customer information and to detect potential large-scale thefts of personal and financial information. The existence of such persistent threats facing real-world organizations creates opportunities for business researchers to contribute to an extremely high-impact and profound body of work. Exploration and development of new capabilities for Darknet and cyber threat intelligence could help businesses better defend themselves in this rapidly changing landscape.

Figure 1 – Encountered Cyber Threat *Operation Green Rights*, Environmental Hacktivism

Cybercriminals often congregate within Darknet virtual communities, creating a valuable repository of data. Such communities often exist as web forums, a commonly studied medium in more traditional virtual community research (Liu & Chen, 2013). Web forums allow participants to post messages and take part in numerous discussions simultaneously. Additionally, forum participants may share hyperlinks, pictures, videos, and other web resources.

Studies focused on cybercriminal and Darknet forums have often been limited by shortcomings in data and methodology despite the high societal importance of this work domain (see Table 1 for a sample of related works from recent years). For example, many studies utilize manual data collection methods that produce limited data repositories (Hutchings & Clayton, 2017). Others perform analyses that rarely venture past surface-level analytics of Darknet forum structure and contents (Yip et al., 2013). As a result, there is a large gap in research utilizing computational techniques that enable large-scale research, as commonly observed in more traditional virtual community or big data research.

Previous Studies	Data Sources	Research	Analytical Methods	Findings
Holt et al., 2017	Forums	Ideologically-motivated Attackers	Interview of 10 forum participants	Religious and political ideology is at the root of many ideologically-motivated attacks
Hutchings & Clayton, 2017	Forums	Case study of cybercriminal discussions regarding a specific attack tool	Performed keyword searches to identify relevant forum discussions; qualitative analysis of identified discussions	Illuminated how cybercriminals sold, traded, and shared configuration files for the focus attack tool
Macdonald & Frank, 2017	Forums	Develop method to estimate number of participants in stolen data markets	Qualitative and statistical analysis of samples collected from three forums that facilitate financial crimes and fraud	Identify method and develop framework for estimating the size of criminal populations within stolen data markets.
Van Hardeveld et al., 2017	Forums	Case study of tools cybercriminals use to trade securely	Qualitative analysis of 25 tutorials regarding trade of stolen data. Tutorials were found in cybercriminal forums	Illuminates how cybercriminals use tools to stay anonymous in the process of obtaining or crashing out on stolen payment card data
Décary-Hétu & Laferrière, 2015	Forums	Case study of how law enforcement has disrupted cybercriminal forums	Qualitative analysis of leaked data dumps belonging to cybercriminal forums previously targeted by law enforcement	Identify that even small disruptions that law enforcement cause to cybercriminal forums can cause significant impact a forum's popularity and ability for participants to trust each other
Yip et al., 2013	Forums	Cybercriminal black markets	Combination of manual analysis and automated network analysis of two cybercriminal carding forums	Underground trading facilitated by social networking, reputation, and quality control
Martin, 2013	Forums	Cybercriminal black markets	Manual analysis of the *Silk Road* cryptomarket and forums	*Silk Road* and similar cryptomarkets will assume greater share of global trade of illicit drugs
Holt & Kilger, 2012	Forums	Cybercriminal skill in global hacking	Manual qualitative analysis of contents and networks found	Cybercriminals practice a meritocratic culture, majority of

		community	within cybercriminal forums and other cybercriminal-related web pages	participants are unskilled
Motoyama et al., 2011	Forums and Internet-Relay-Chat	General exploration	Manual content analysis, some automated network analyses	General descriptions of cybercriminal interactions in forums and IRC, existence of meritocratic structure

Table 1 – Summary of Recent Cybercriminal Community Studies

Early studies have demonstrated that more comprehensive work using cybercriminal-generated data is critical to improve cyber-defense. However, advancing this research stream and generating actionable intelligence from cybercriminal and Darknet communities cannot be accomplished unless researchers move away from manual or otherwise non-scalable identification, collection, and analysis procedures. These shortcomings occur because cybercriminal forum data is often much more difficult to identify and collect than data from more traditional web forums. Additional complications arise when considering that researchers undertaking Darknet research may inadvertently expose themselves to numerous cyber threats by collecting and viewing potentially malicious contents. Overall, when compared to more typical web forum research, a greater level of planning and technical sophistication is required of researchers to conduct successful Darknet-related projects.

THE DICE-E FRAMEWORK FOR DARKNET RESEARCH

Due to the challenges and limitations facing current work, a research framework for conducting Darknet and cybercriminal community research would be of great value. The Darknet Identification, Collection, Evaluation, with Ethics (DICE-E) research framework described here serves to guide researchers through their own explorations of Darknet forums and other potential communities (Figure 2). DICE-E is intended to help scholars through the entire lifecycle of a Darknet research project.

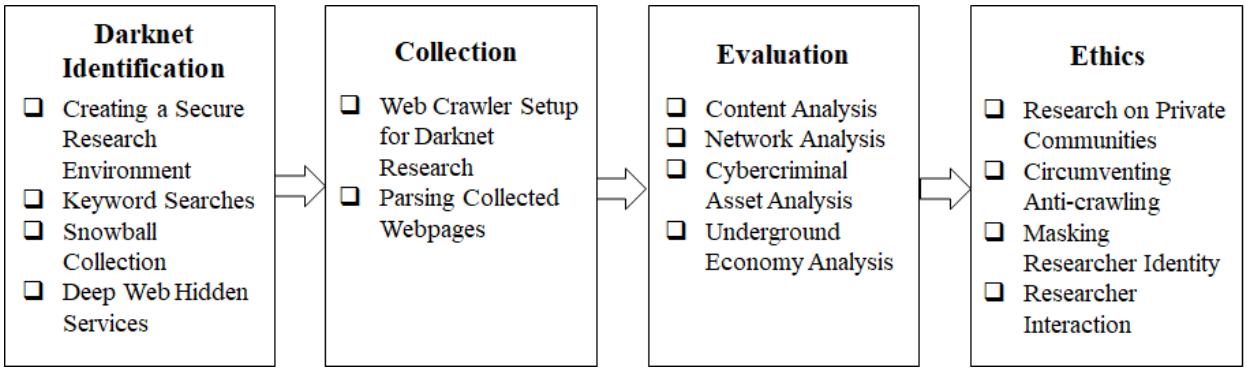

Figure 2 – The DICE-E Framework

Four different phases of research are described: (1) techniques for identifying Darknet forums suitable for use in research studies; (2) techniques for collecting forum data, including suggestions for safe collection practices; (3) potential analytical methods for interpreting forum content; and (4) ethical considerations when conducting research within Darknet and potential criminal communities.

Darknet Identification

To successfully conduct a Darknet forum study, quality data sources must first be identified. Forum identification is often a trivial process in/for more traditional research, but the Darknet context possesses unique properties of which researchers must be mindful. First, forums are used heavily by cybercriminals across the world, allowing researchers to examine cybercriminal activities in different geopolitical regions (Motoyama et al., 2011; Benjamin & Chen, 2015). Many of these forums utilize different languages depending on their origin, with the most frequently encountered languages being English, Chinese, and Russian. Further, forums will vary from one another in terms of size, activity level, and topical coverage (Macdonald & Frank, 2017). For example, some communities will include a wide variety of hacking-related discussions, while others are focused on discussing specific topics such as carding (i.e., credit card fraud). By utilizing appropriate cybercriminal forum identification techniques, it is possible to discover quality data sources in different languages and of diverse topical focus. In total, there

are three primary techniques for identifying cybercriminal forums that can be utilized: (1) keyword searches, (2) snowball collection, and (3) deep web hidden services. Each technique differs and yields unique forums that may not otherwise be found with alternate forum identification procedures. However, before beginning to search for data sources, it is important for researchers to take precautionary measures that help ensure their security.

Creating a Secure Research Environment

Due to unique challenges faced when attempting to research Darknet forums, it is worthwhile for researchers to plan and create a computing environment to securely conduct research in this space. Primarily, the suggested environment will fulfill two purposes: (1) to aid researchers in safely downloading and archiving relevant Darknet contents, and (2) to later provide a quarantined space for researchers to analyze and evaluate collected data. A properly setup research environment is an important component for successfully executing all Darknet-related research tasks.

Data identification and collection from Darknet forums can pose numerous threats to researchers. For example, many forums may embed malicious “drive-by” JavaScript code within web pages in an attempt to exploit outdated and vulnerable browsers (Cova et al., 2010). Unsuspecting users, including researchers, could become infected with malware when browsing underground forums. Because of this and other security hazards, it is important to make sure that if a computer used for data collection is infected with malware, that infection does not spread to other computing resources. One way to ensure safety is to take precautionary measures and perform all Darknet data collection on computers that are quarantined or otherwise removed from local networks shared with other computers and devices. This can be operationalized by

relying on virtualized operating systems and networks, or by physically segregating real-world networks.

Renting virtual private servers from cloud services would also provide the added benefit of easily being able to clone servers; one could simply set up research tools on one server and then clone it to scale collection for different forums. However, it is important for researchers to check for terms of service (TOS) violations when identifying potential cloud service providers. Most providers will state what types of usage behaviors they prohibit from their servers, with many only explicitly limiting criminal behavior. In general, cybersecurity research seems of acceptable use. However, security researchers must remain vigilant against accidental violations that may occur. For example, a plausible scenario would be to accidentally infect a cloud server with malware originating from a Darknet forum. The malware could to take control of the server and use it to launch cyber-attacks, thus leading to illegal usage activity that breaches the cloud provider's TOS. Lastly, many cloud providers openly invite their users to contact them regarding TOS questions; researchers are encouraged to contact cloud provider's directly to clarify any points of doubt.

Additionally, it is beneficial to establish a database for archiving collected Darknet forum data. Collected data can be processed and sanitized for secure long-term storage, as described in the collection procedures of the DICE-E framework. Further, the database should be set up on a network that is separate from the quarantined computers used to actually download the Darknet contents in order to avoid potential malware and other security threats from corrupting the archived data. A simple use-case would be to create a database table per identified forum, with each table record representing a forum post. The record could store forum message attributes such as the author's name, the post title, message body, date, and other available information of

interest. After such a database is created, data processing programs can be run against webpages collected by crawlers to extract relevant forum message attributes from surrounding HTML and JavaScript code. Extracted data could then be stored as plaintext within the database. This process also provides some additional security for researchers, as it allows forum data to be viewed without needing to view the original forum webpages that may contain JavaScript-based malware.

Archived data could then be accessed on-demand (using SQL or other means) for analyses and evaluation. For example, if the research objective is to extract popular topics discussed within a given cybercriminal forum, the following steps could be taken to ensure secure practices: (1) set up a new sandbox environment intended to house all analytics work, (2) import forum messages into the sandbox by pulling from the data archive, (3) conduct topic analysis using a relevant technique of choice, and finally, (4) evaluate results. The steps in this approach can be taken for all different types of Darknet community research, as the only changing component is the chosen analytical focus and methodology.

Keyword Searches

The first and most accessible method to identify possible forums is to conduct keyword searches (Décary-Hétu & Leppänen, 2016; Hutchings & Clayton, 2017). Certain types of Darknet-oriented forums will publicize themselves in order to attract new members to contribute to the cognitive advancement of the community and to also participate in underground markets. For example, searching for "carding forum" may yield a cybercriminal community focused on credit card fraud, while "black hat forum" may just return a more general-topic cybercriminal forum. Additionally, keyword searches can be tailored to find forums of a specific geopolitical region by searching for translated queries (e.g., 'хакер форум,'or Russian for 'cybercriminal

forum'), or by including the country or language of interest as a keyword within the query (e.g., "Chinese hacker community").

However, one limitation to keyword searches is that forums that openly publicize themselves are notorious for attracting benign users who generally possess only passing interest and are not deeply involved with Darknet or cybercriminal activities. For example, in the case of cybercriminal forums, many participants are known as "script kiddies," a term often used to describe individuals that possess little to no actual hacking skills. Script kiddies are entirely dependent on using hacking tools that more experienced cybercriminals release publicly; they contribute little to no valuable contents to their community, and generally are a source of noise for researchers wanting to use cybercriminal forums for threat intelligence, potential target identification, cybercrime attribution, and so on (Benjamin & Chen, 2012; Holt et al., 2012). To further exacerbate the issue, more knowledgeable cybercriminals are generally conscious to not reveal their ongoing activities or credible threats due to concern that law enforcement may be monitoring forum activity (Motoyama et al., 2011).

While many Darknet forums may be identified through simple keyword searches, they do not yield the highest quality content for research. Forums identified through keyword searches are generally at the surface level of the Darknet and are openly visible to those who look for them. As a result, such forums often ban discussion of many illicit contents to avoid potential intervention by law enforcement agencies. However, despite these shortcomings, they are still of use to Darknet researchers. After identifying an initial set of forums through keyword searches, further actions can be taken to discover a set of less accessible and more interesting data sources. In particular, forums identified through keyword searches often contain numerous hyperlinks to

more covert Darknet communities that may contain higher-skilled participants and a greater volume of illicit content.

Snowball Identification

In many Darknet forum conversations, participants may reference or share hyperlinks to other cybercriminal communities or underground markets (Benjamin et al., 2015a). Such discussions can be exploited by researchers to discover new forums, including more private and secretive forums that do not appear indexed by major search engines. This method allows researchers to move away from surface-level Darknet forums onto more potentially secretive and interesting communities.

In order to operationalize snowball identification, text parsers can be written to automatically scan through cybercriminal postings collected from previously identified forums. A simple scenario would be to create a text parser that scans cybercriminal forum postings for the string "*http://*" in order to automatically identify and extract hyperlinks shared among Darknet forum participants; such hyperlinks may lead to other communities. More complex text parsers could be developed utilizing regular expressions, the process of analyzing text by searching for pre-defined patterns. For example, a regular expression could be crafted to scan cybercriminal forum postings for strings that resemble account numbers of stolen credit cards; assuming a credit card number consisted of 16 consecutive digits, a regular expression could be used to scan text for patterns of 16 consecutive digits. Cybercriminal posts found to contain such patterns may contain references to underground markets and carding communities. By utilizing a combination of different regular expression patterns, researchers can exploit their existing data collections to find brand new data sources of interest. Any URLs that are identified can be fed into a web

crawler for downloading, which will be detailed in our section on cybercriminal forum collection.

Deep Web Hidden Services

While many underground forums are accessible through the traditional Internet, there are some communities that exist in normally inaccessible networks (often referred to as the 'deep web'). In particular, much of the deep web exists as anonymized, peer-to-peer networks where network traffic is purposely obfuscated in an attempt to protect user identity and conceal activity patterns (Martin, 2013; Benjamin & Chen, 2014). Many users of such networks will often privately host "hidden services," or web services, for other network participants to use. Potential applications of hidden services include benign services such as anonymized web and e-mail hosting, as well as more nefarious services such as underground markets and Darknet discussion forums.

Darknet forums acting as hidden services may contain more advanced participants or more sensitive contents than more visible communities, explaining their need to be more secretive in nature (Martin, 2013). By extension, this means that such forums may be of great value to researchers. For example, in the cybersecurity context, hidden forums may be a source of more credible data for understanding emerging cyber threats or discovering potential targets of cybercrime. However, gaining access to such forums is nontrivial; a researcher generally requires special software or technical knowledge to connect to a deep web network and locate hidden content of interest. Thus, we will outline several steps researchers can take to connect to and identify cybercriminal forums within the Tor anonymity network[1], one of the most active

[1]The Tor anonymity network was developed in the mid-1990s by the United States Naval Research Laboratory and was later further advanced by DARPA. The Naval Research Laboratory released Tor under a free, open license in 2004. Since then, the network has grown in a variety of directions, including becoming home to a variety of illicit underground communities (Martin, 2013; Benjamin et al., 2015a).

networks at the time of this writing. The steps listed are also applicable to other similar anonymity networks.

Download a Network Client: To access a deep web anonymity network, a specialized software client must generally be used to establish a connection and communicate with the network. In the case of the Tor network, a public client can be downloaded from http://www.torproject.org. Options to download the Tor client in various forms exist, but perhaps the easiest to deploy is to download the "Tor Browser Bundle," where a Tor client is packaged as a plug-in into a stand-alone Mozilla Firebox browser. One can simply download the browser bundle and use the included browser to access and browse hidden services located within the Tor network. Additionally, since both Tor and Mozilla Firefox are open source projects, the browser bundle is cross-platform and available on a variety of operating systems. Other anonymity networks besides Tor may have their own custom software necessary for accessing the network.

It should be noted that since Tor is a peer-to-peer network, the Tor software operates by automatically connecting to Tor nodes hosted by volunteers globally. While this can help anonymize network traffic, it also can result in slow collection of Darknet forums due to the amount of network traffic routing that occurs before reaching the destination server. Further, the peer-to-peer nature of Tor and the fact that it is a gateway to the Darknet may present TOS issues for some cloud service providers. It is important to check for TOS violations before proceeding with Tor usage.

Identify Hidden Service Directories: Identifying hidden services within the Tor network is a nontrivial task. First, the web addresses belonging to Tor hidden services are generally sequences of random alphanumeric characters (e.g., https://pwoah7foa6au2pul.onion was the address for *Alphabay* market before its shutdown by law enforcement). The web addresses thus do not

indicate potential functionality or content of the hidden service they are assigned to. Second, Tor hidden services do not utilize traditional top-level domains such as '.com' or '.net.' Instead, Tor hidden services use the '.onion' nomenclature as a reference to the multi-layered network traffic encryption implemented in Tor. This multi-layered encryption is often conceptualized as layers of an onion, and thus is the reason that hidden services located within the Tor network are commonly referred to as "onion files."

These characteristics of Tor (and other similar anonymity networks) make it difficult for security researchers to identify relevant data sources. However, there are some hidden service directories that publicize themselves and can be discovered through keyword searches; simply querying "Tor hidden service directory" on a major search engine will yield lists of various hidden services. For example, this technique yields one of the most well-known hidden service directories, the Hidden Wiki (Figure 3). These directories are generally public, open source efforts that are created and maintained by community members. Known hidden services are typically categorized by their topical relevance or intended use. Further, the directories are not representative of all hidden services in existence, however, they may sometimes include web addresses of cybercriminal forums and other underground communities. After an initial set of cybercriminal forums is identified from hidden service directories, a snowball collection approach can be taken to find more data sources as described previously.

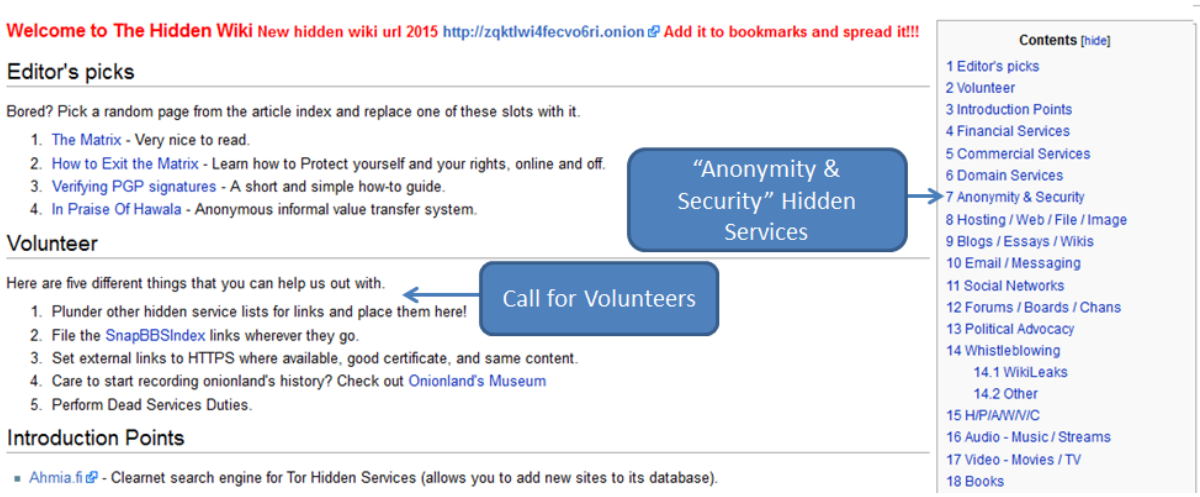

Figure 3 – The Hidden Wiki, http://zqktlwi4fecvo6ri.onion

Collection

After identifying Darknet forums suitable for research purposes, forum contents must be downloaded for offline analysis. Web crawlers can be used to automate the collection of websites and virtual communities, such as forums (Liu & Chen, 2013). To use a web crawler, one must specify a starting seed website. The crawler will automatically download webpages it encounters, while constantly discovering new webpages for collection by following encountered hyperlinks. Successfully downloaded cybercriminal data could then be processed and archived for long-term storage.

However, using web crawlers to collect Darknet forums presents many unique challenges not encountered when crawling more traditional virtual communities and websites. Many Darknet forums may employ various anti-crawling mechanisms that make automated data collection difficult (Benjamin et al., 2015a). The intention of such mechanisms is generally to prevent surveillance by law enforcement, obstruct security researchers from collecting data, and safeguard server resources from being abused by rival Darknet communities. Further, Darknet forums may include malware or other threats that can harm researchers (Cova et al., 2010).

Overall, the Darknet data collection process requires careful planning to safely gather and archive relevant contents.

Web Crawler Setup for Darknet Research

Web crawlers are a commonly used technology for various Internet-based data collection projects, especially in contexts where large-scale collection is required and manual effort is unfeasible. They operate on the principle of snowball collection procedures; an initial website or set of websites is targeted for collection, hyperlinks are identified and extracted from downloaded webpages, and the hyperlinks are then fed back into the crawler for collection. They make excellent candidates for operationalizing Darknet forum collection and may even be helpful for identifying previously unknown communities .

However, web crawler setup must consider forum heterogeneity. The exact procedures a web crawler utilizes to traverse a targeted forum may need to be custom tailored specifically for that forum. Some forums are pleasantly simple to collect, while others are frustratingly difficult due to technological challenges, or even impossible without exposing oneself to severe risks or ethical concerns. In particular, many Darknet forums will employ different levels of anti-crawling mechanisms, making collection difficult for researchers (Benjamin et al., 2015a). Many of these anti-crawling measures are easily circumvented with some effort, but will impact how web crawlers must be configured to successfully collect a given forum. Table 2 contains descriptions and recommended countermeasures to the most common anti-crawling mechanisms.

Anti-crawling Measure	Description	Countermeasure
CAPTCHA Images	CAPTCHA images are a type of test used by many web services to determine if the user is human or an automated bot. Their purpose is to prevent bots from accessing contents.	Solve the CAPTCHA manually and bind the generated server session cookie with web crawling software. Requires advanced web crawling software.
Distributed Denial of Service	The forum detects scripted behavior, such as web crawling, and	Researchers can alter crawling rates and introduce random intervals between web page requests in

(DDoS) Prevention	blocks the associated IP address. This is often done to prevent DDoS attacks against the forum.	order to mask crawling activity and avoid triggering DDoS prevention software.
IP Address Blacklists	The forum has blacklisted several IP addresses, including those of public proxy servers and Tor nodes that could be used for anonymity.	Set up a private, dedicated proxy server to reroute crawler network traffic. The proxy server can be deployed using cloud services so it is easy to spawn and destroy proxies for new IP addresses to use.
Paywalls	Forum content is locked behind a registration or access fee.	The only way to access forums with paywalls is to pay their fee. This carries a high risk as the researcher may be defrauded or encounter legal trouble. Recommend consulting law enforcement before pursuing.
User-agent Check	Forums verify HTTP requests come from common browser user-agents, and not web crawlers or other software.	Mimic accepted user-agents during crawling process. Many popular web crawlers possess this feature.
User/password Authentication	Forums require users to register and login before accessing the data.	Register an account with the forum. The registration process is generally completely automated and requires no interaction with forum participants.
Vouching	Gaining access to forum content requires receiving vouches from existing members.	Requires making connections within Darknet forums to receive vetting for access to private communities. Not recommended for security researchers due to potential ethical problems and biasing data due to researcher manipulation.

Table 2 – Cybercriminal Forum Anti-crawling Mechanisms

While most of the anti-crawling mechanisms listed in Table 2 can be circumvented, there are two that require extra consideration. Forums that require registration fees, and those that require vouching, carry great risks to researchers in the form of fraud, legal risks, ethical concerns, and biasing data due to researcher manipulation. For these reasons, it may be wise for security researchers to generally avoid such forums as the risks likely exceed the benefits. However, if there is strong interest in focusing on such communities, it would be wise to contact local law enforcement and university administrators to discuss how to best navigate around potential issues

For additional security, proxy servers and anonymity networks such as Tor can prove useful to researchers wishing to conceal their identity from Darknet participants and cybercriminals (Martin, 2013). Specifically, whenever an individual or web crawler accesses servers hosting Darknet forums, the server will generate log files revealing IP addresses that connected to the server. Thus, researchers' origin IP address of researchers, resulting in a significant security risk. However, proxy servers and anonymity networks can be utilized to re-route researcher web

traffic through external connections, effectively concealing the identity of researchers' machines from forum servers. Many popular web crawlers natively support proxy server usage, and researchers can simply search for public proxy servers in order to implement a web crawler. In the case of Tor, after a Tor client is installed on a computer, the web crawler can be bound to the client in order to communicate with the Tor network. To do this, the web crawler can be configured to forward traffic to a SOCKS proxy located at the local network port that the Tor client is listening to for network traffic (by default, this is generally http://127.0.0.1:9051). After this step, the Tor client will automatically handle network communication, and the researcher can continue operating the web crawler as normal.

Overall, researchers willing to take the necessary precautions can safely collect Darknet forum data while avoiding many technical risks and without exposing their identity. However, additional steps are needed to extract relevant data from raw webpages downloaded directly from forums.

Parsing Collected Webpages

As web crawlers traverse through Darknet forums, they download webpages that must be processed to extract information of interest. Text parser programs utilizing regular expressions (similar to those used in snowball collection) can be used to accomplish this task. For example, text parser programs can be written to automatically extract forum postings, author names, thread titles, and other information by identifying patterns of HTML code that correspond with data of interest. Specifically, forums generally follow HTML design templates that contain unique HTML code patterns for encapsulating each forum message and associated author data. Such repeated patterns can be manually identified by researchers and subsequently used within text parsers for automated information extraction across all webpages for a given forum.

Evaluation

After Darknet forum contents are collected and parsed into a database, researchers can begin to explore their data through various analytical techniques. This section contains a brief discussion of potential analytical directions that researchers can take to develop better understanding of Darknet forums. The intention here is not to provide an exhaustive listing of all potential analyses that can be taken, nor is it meant to be an in-depth discussion of the methodologies themselves. Rather, focus is placed on the analytical directions that are relevant for scrutinizing the most prominent types of data found within Darknet forums. Further, it is not recommended that researchers apply each of the described techniques for every Darknet study, nor is there a particular order for applying the techniques. Rather, researchers should select the most relevant techniques that lend themselves to exploring research questions of interest.

The types of data found in Darknet forums are largely comparable to those found in more traditional web forums, including mostly text content and social network features (e.g., thread reply-to structures). While the context of the data is very different, the similarity of data types between these two sources indicates that many analytical methods may crossover to Darknet research. In particular, content analysis and network analysis procedures used in the broader virtual community research stream appear to be natural candidates for analyzing Darknet contents.

Further, Darknet forums act as sources for new, unique data types that are largely unexplored in current literature. Cybercriminal assets that assist hackers in organizing and conducting attacks can be found in abundance (Yip et al., 2013). Such assets are core to many interesting research questions related to attribution of attacks, cybercriminal knowledge transfer,

how attacks are organized, and more. However, their unique nature necessitates a discussion as to how this class of data can be best handled.

Finally, Darknet forums often feature black markets that contain data on illicit trade and underground economies (Holt et al., 2012). To adequately extract meaningful knowledge from these Darknet markets, analytical methods can be borrowed from more traditional online market research. This includes analysis of many unique market data points that can yield knowledge regarding cybercriminal supply chains, valuation of hacking assets, exogenous shocks to market activity, and more.

Given the data types found in Darknet forums, a core set of computational methodologies can be emphasized for this research stream. A brief overview of each method is described in the following subsections. Adoption of the described techniques can help guide researchers in developing their own research questions and analyses for several unique studies on Darknet communities.

Content Analysis

Content analysis provides a powerful approach for gaining insights into the types of conversations that Darknet participants have amongst themselves. Computational approaches for analyzing text (i.e., text mining) are extremely useful and have seen widespread application in traditional virtual community research. Many attributes of virtual communities make them excellent candidates for text mining methodologies; for example, such attributes include high volumes of data, many participants, a multitude of distinct topics being discussed, and potential for multilingual data. In comparison, manual efforts to analyze virtual community data can often become limited by such issues.

Possible analytical directions enabled through text mining are numerous and commonly include techniques such as automated topic modeling, document classification, sentiment analysis, and language modeling. These methods are grounded in computational linguistics and statistical natural language processing (Hirschberg & Manning, 2015). They are popular approaches for conducting large-scale analysis of traditional virtual communities, and would serve as an excellent starting point for researchers wanting to explore Darknet contents through computational means.

For example, automated topic modeling can be used to detect the actively trending discussions within cybercriminal forums and to extract participant expertise based on discussed topics (Benjamin et al., 2015). Document classification is a similar approach; in the virtual community context, document classification would entail grouping participant messages together based on their topics or some other specified attributes. Sentiment analysis can be used to detect the reliability of underground market participants based on the feedback left by other market participants after transactions occur (Li & Chen, 2014). Such techniques can provide a rich depth of information about conversations that occur within the Darknet.

Network Analysis

Network analyses are a common technique used to understand the flow of information between participants of various types of virtual communities. Network analyses can be used to develop understanding of individual community participants, their relationships with other participants, and their location within the overall community structure. They have been used extensively in more traditional virtual community research to measure user influence, to aid in identification of key users, to assess how information travels within networks, why social ties develop, and more.

In the Darknet context, network analyses can be useful for revealing cybercriminal supply chains, how different Darknet and cybercriminal groups interact with one another, underground market activity, and more (Holt, 2013; Li & Chen, 2014; Benjamin et al., 2015). Additionally, network analyses would help researchers and practitioners better assess the credibility of threats emerging within Darknet forums. For example, literature suggests that there is a variation of knowledge proficiency among forum participants (Radianti, 2010). By using network analysis techniques to identify key actors within forums, the credibility of threats identified by researchers can be evaluated based on the associated participant or Darknet group.

Cybercriminal Asset Analysis

Darknet forums are rich in many unique contents that make interesting candidates for research. For example, in cybercriminal forums, hackers will often share malware, hacking tools, hacking tutorials, source code examples, and other illicit assets (Motoyama et al., 2011; Holt et al., 2012). These materials can be studied by researchers to understand more about Darknet trends, emerging cyber threats, potential targets or victims, cybercrime attribution, and more (Motoyama et al.,2011; Benjamin et al., 2015). Recent focus has been placed on developing new methods for automated categorization and analysis of such assets. Some data, such as hacking tutorials, may be treated as documents that can be analyzed through the aforementioned text mining techniques. Topic modeling and document classification are of particular relevance for conducting automated analyses.

However, analysis of source code and binary files (e.g., hacking tools) presents a more difficult challenge. There are limited recommendations possessing comprehensive scope that can be made. In general, advancements in computer science should be considered for this data type. Machine learning is promising for its capability of analyzing different data types that possess unique

characteristics (Samtani et al., 2016). Specifically, analysis of source code snippets shared between Darknet participants would likely require some manual annotation of data for the purpose of machine learning model training.

For example, this process could involve labeling potentially hundreds of source code files for their characteristics and subsequently using the trained model for analysis of newly encountered files. Source code would be annotated for its characteristics, but could then be used for automated classification of future source code. Further, binary files would present a similarly difficult challenge made even more complex by the fact that binary files are difficult to deconstruct. There is a great amount of literature in computer science and engineering that investigate this issue. Most techniques include capturing the run-time behaviors and opcode exhibited by binary files and performing analyses based on this captured data. Overall, referring to the state-of-the-art in computer science and engineering literature for more information regarding source code and binary file analysis is highly recommended. Recent advancements and the occasional release of public research tools resulting from such research has potential to benefit Darknet researchers.

Underground Economy Analysis

A number of Darknet forums possess underground markets where participants buy, sell, and trade assets and services (Benjamin et al., 2015a). There are several opportunities to analyze data from Darknet markets, including analysis of underground economy participants, pricing mechanisms, goods exchanged, cybercriminal supply chains, and more. In particular, there is a need to understand the currencies used by actors within underground markets, as they are a potential point of weakness that may be exploited by law enforcement agencies to disrupt financial transactions between buyers and sellers of malicious assets. Cryptocurrencies such as

Bitcoin are used because of their perceived security and anonymity, though the volume of transactions or how such transactions occur is unknown.

Additionally, there is a need for quantitative assessments of relationships between underground economy participants by considering the number, shape, and composition of networks in cybercriminal markets (Motoyama et al., 2011; Yip et al., 2013). Research on Darknet market network relationships is extremely limited and exploratory, generally using a single forum or small samples of data from multiple forums.

More effort in this area would increase our understanding of Darknet market dynamics. The use of econometric modeling techniques is of extreme relevance to this area. There exists a substantial and mature stream of literature that focuses on the analysis of traditional online markets. Many of the modeling approaches and conclusions drawn from this research can be applicable to the Darknet context.

Ethics

Conducting ethical research is a frequent concern when pursuing virtual community studies (Cotton, 2004; Buchanan & Ess, 2009; Hoser & Nitschke, 2010). While this issue has been discussed extensively in past works, there are specific issues regarding the Darknet context that warrant additional discussion. For example, many traditional virtual communities offer open access and free user registration to visitors. As a result, forum discussions are not considered private and are instead within the public domain. Typically, using these datasets in research is of low ethical concern (Chang & Chuang, 2011; Hoser & Nitschke, 2010; Liu & Chen, 2013).

However, in the Darknet context, many forums strive for secrecy and do not intend for internal discussions to be part of the public domain (Flick & Sandvik, 2013; Martin & Christin, 2016). The covert nature of the Darknet presents a unique challenge in understanding potential

ethics issues. To form a grounded basis for scrutinizing potential ethics issues in Darknet research, an extensive review of literature regarding Darknet research ethics was performed that yielded few relevant materials. Some recent works typically focus on singular issues and do not provide advice for operationalizing the kind of computationally driven and large-scale Darknet forum research described in this manuscript (Barratt & Maddox, 2016; Martin & Christin, 2016).

Thus, review of relevant research streams may provide some perspective. In particular, ethics-related work in traditional virtual community research and criminology research are of relevance. For example, as noted above, there exists a body of work investigating ethics violations in traditional virtual community work. Further, given the often-criminal nature of the Darknet, ethics issues in criminology research may provide useful insights (King & Wincup, 2008). While much criminology work focuses on real-world organizations rather than virtual communities, valuable perspectives may still be borrowed regarding cautions that researchers should be mindful of when dealing with criminal elements.

A constant challenge of technological advancement is that technology has often blurred the boundaries of ethical research. Such is the case with the Darknet. To the best of the authors' knowledge, there exists no comprehensive source of ethics guidelines to aid researchers and university administrators with navigating potential ethics pitfalls when conducting Darknet research projects. Thus, a discussion of potential ethics pitfalls and solutions would provide meaningful contribution for helping researchers navigate the Darknet.

Darknet research activities that generate potential ethical concerns are highlighted in Table 3. Each research activity is aligned with relevant literature that discusses the level of concern associated with the activity, reasoning for why the specific level of concern is assigned, and suggested action to remedy the concern. For some research activities, strong recommendations

could not be found in existing literature; such is the case regarding ethics issues with circumventing anti-crawling. To avoid ad hoc recommendations on how to handle such concerns, generalized guidelines are provided.

Research Activity	Ethics Concern?	Reasoning	Suggested Action
Research on Underground Web Communities	Low Concern (Barratt & Maddox, 2016; Martin & Christin, 2016)	-Observation without participation generally leaves a minimal footprint on communities -Similar observation performed in traditional virtual community research -Observation of criminal elements is the basis of many works in criminology	-Follow the DICE-E framework to safely identify and collect Darknet forums -When formulating research plans, consider major areas of IRB concern: privacy, data security, sensitivity of data, methods to elicit participant consent when necessary, confidentiality, and anonymity (Buchanan & Ess, 2009) -Generally of low concern unless researcher interacts with Darknet participants; check with your local IRB for specific issues
Circumventing Anti-crawling	Low Concern (Thelwall & Stuart, 2006; Martin & Christin, 2016)	-Employment of crawlers is critical for large-scale, data-driven virtual community research -Computational methods aid in reproducibility of research, including methods to automate data collection -Should be considered in a cost-benefit analysis mindful of context and importance of research	-Assess forum anti-crawling mechanisms and configure web crawlers as is required -Carefully document and report crawling procedure per forum for reproducibility
Identity Obfuscation	Low Concern (Flick & Sandvik 2013; Barratt & Maddox, 2016)	-Necessary procedure to maintain researcher safety -Hidden services are only accessible on anonymity networks where identity obfuscation is an inherent characteristic of the medium	-Maintain identity obfuscation to avoid network owners' exposure to security risks -Minimize risk of cyberattacks that could disrupt research progress by following DICE-E framework guidelines -Can be operationalized through anonymity networks such as Tor or through virtual private servers offered by cloud services
Researcher Interaction	Moderate Concern (Flick & Sandvik 2013; Barratt & Maddox, 2016)	-Some research questions are impossible to answer without active participation -For surveys and interviews, it is important to respect participants' anonymity; consent forms should be signed	-Requires thorough planning to operationalize research and mitigate risks -Requires discussion with local IRB office -Recommend notifying university administrators of research activities -For some research explorations, notify law enforcement

Table 3 – Darknet Forum Ethical Research Heuristics

Research on Underground Web Communities

An understanding of recent perspectives on the ethics of underground community research is critical for scholars wishing to conduct Darknet research. To explore this issue, it is worth first looking at ethics issues present in more traditional research. Ethical use of virtual community data for research is already prevalent across numerous domains and has been considered of low concern in many Institutional Review Board (IRB) decisions (Buchanan & Ess, 2009). In many cases, researchers will utilize data from public communities that provide unrestricted access and are thus considered as public domain. There is generally little to no expectation for privacy among users participating within such communities, resulting in low ethics concern.

Conversely, while some Darknet forums are public, the vast majority strive for privacy and secrecy. Research assumptions and perspectives that are prevalent in traditional research contexts may not hold true for the Darknet. Are there indeed ethical concerns regarding research on Darknet forums?

A local IRB office was asked whether approval or any special considerations were necessary for Darknet community research. It appears that Darknet community research would typically not qualify for the DHHS criteria of Human Subject Research because it would not involve interaction or interventions with an individual or contain identifiable private information. Further, Darknet community research would not meet any FDA definitions of Human Subjects Research because it does not involve a drug, device, or any other article regulated by the FDA. However, any research that involves interaction with Darknet participants would require IRB approval or exemption before the research is conducted.

Discussions found in recent literature can also be leveraged to help address this question. The consensus is that simply observing underground communities is generally of low ethical concern,

though several steps can be taken to ensure avoidance of ethics pitfalls (Barratt & Maddox, 2016; Martin & Christin, 2016). Specifically, it is important to consider major areas of IRB concern as a guide to avoid ethics pitfalls, including privacy of subjects, data security, sensitivity of data, confidentiality of subjects, and methods to elicit participant consent when necessary (Buchanan & Ess, 2009). Overall, a great amount of benefit can be derived from exploring Darknet communities without interacting with participants, as evidenced by literature focused on understanding Internet-enabled trafficking, fraud, terrorism, and other illicit behaviors (Leavitt, 2009; Martin, 2013).

Several actions during the operationalization of Darknet research projects can also help reduce concern. Proper utilization of the Darknet forum identification and collection methods outlined in the DICE-E framework can help reduce potential ethics violations. The described methods help protect researchers and scientific equipment from unnecessary exposure to risks that could result in ethics-related issues. For example, Darknet forums may contain ransomware, a type of malware that steals sensitive data and demands payment of ransoms for the return of the data. Researchers not following secure procedures could become afflicted, subsequently forcing them to decide whether to pay a ransom that will likely support future cybercrime. Such issues can be avoided by simply following the guidelines established by the DICE-E framework.

Circumventing Anti-crawling

Another research activity that raises potential ethical concerns is the legitimacy of circumventing anti-crawling mechanisms employed by Darknet forums. First, it should be noted that the act of web crawling generally does not present ethical concerns and is a commonly utilized technique in traditional virtual community research (Thelwall & Stuart, 2006). However, in the Darknet context, researchers must often explicitly customize their web crawlers to collect

data from forums wishing to remain private and undisturbed. Many of these forums also employ software measures to inhibit crawlers from successfully operating.

Anti-crawling mechanisms can be conceptualized as server-side, software-based barriers that will prevent web crawlers from performing their normal functions. To circumvent them, a web crawler must be customized by having its run-time behavior altered in ways that avoid triggering activation of these server-side barriers. As noted in the DICE-E framework, different forums may implement different types of anti-crawling mechanisms, and thus, crawling strategies should be tailored on a forum-basis.

To the best of the authors' knowledge, there is no formal literature regarding ethics concerns generated by circumventing anti-crawling measures. Conceptually, customization of web crawlers to bypass barriers can be regarded as a sort of deception practiced by researchers. Nevertheless, there is nothing to stop researchers from manually collecting forums that do feature anti-crawling. Many of the more private Darknet communities are smaller and trivial to collect by downloading webpages manually. This paradox makes an obscure issue even harder to deconstruct.

The lack of literature concerning the ethics of this practice makes it difficult to develop well-grounded arguments. Rather than making ad hoc statements directly addressing this matter, it is more worthwhile to consider a completely distinct perspective on this issue. Specifically, the legitimacy of research activities possessing undetermined ethics can be evaluated through a cost-benefit analysis (Martin & Christin, 2016). If the research will generate unique outcomes important to advance related fields of science, it appears more permissible to collect data and conduct exploration. Further, the degree of deviance from a known ethical research activity should be considered.

Literature also puts forth the notion that computationally based data collection methods are helpful for reproducibility of research and are preferably used when possible (Thelwall & Stuart, 2006). Anti-crawling circumvention contributes to the achievement of this goal, as manual techniques are not scalable to larger forums and may be difficult to replicate. In particular, different researchers and research groups possess unique capabilities that influence their manual collection capabilities. For example, international forums and language barriers can present a significant hurdle for/to manual efforts. Thus, it is recommended that researchers carefully document and report their anti-crawling circumvention procedures when applicable. Doing so will assist other researchers in replicating results and conducting their own Darknet studies.

Overall, circumvention of anti-crawling mechanisms should present little concern. The DICE-E framework presents a series of crawler strategies that can be employed. It is recommended that researchers utilize these strategies to ensure comprehensive collection. Reporting of implementation details would also be an asset to the greater Darknet research community.

Identity Obfuscation

Another potential concern regarding ethical conduct involves the use of proxy servers or anonymity networks to mask crawler network traffic and to protect researcher identity. Just as with anti-crawling circumvention, identity obfuscation may be seen as a method for researchers to deceive their subjects. However, recent works establish a clear consensus that researchers should take all necessary safeguards to ensure their own security (Flick & Sandvik, 2013; Barratt & Maddox, 2016).

This consensus was reached by evaluating a number of issues. For example, Darknet research performed on university networks can result in harm to the university if proper precautions are

not taken; it is conceptually possible for cybercriminals to unveil researcher identity and launch retaliatory attacks. Such attacks could disrupt or disable university network resources. Another realistic scenario would be for attackers to covertly gain access to research data and subsequently erase it or silently manipulate it. Researchers who do not take necessary precautions and suffer setbacks as a result are performing a disservice to themselves.

To prevent such issues, the DICE-E framework outlines potential actions that researchers can take to protect their identity and maximize their resiliency against cyberattacks. This includes the use of anonymity networks or relaying network traffic through virtual private servers to mask identity. Secure computing environments should be created for all forum identification and collection effort. Regarding the Tor anonymity network specifically, many Darknet forums require participants to utilize it upon connection, thus making it necessary (Barrett & Maddox, 2016). Such forums often have lists of known Tor nodes and will block connections from IP addresses not attributed to the Tor network.

Overall, researchers should not be concerned with ethics pitfalls regarding identity obfuscation, so long as appropriate procedures are taken. This research activity is for the benefit of various stakeholders involved, including universities that host Darknet research projects, and the greater Darknet research community.

Researcher Interaction

Before considering ethics, it is worth noting that researchers interacting directly with forum participants may inadvertently bias or manipulate the data they collect, presenting threats to research validity (Cook & Campbell, 1979). Forum participants may behave differently if they become suspicious due to actions enacted by researchers. Thus, researchers should always be mindful of the footprint they leave on Darknet forums they study.

Researcher interaction is a risky but sometimes unavoidable part of Darknet research. It primarily occurs for two reasons. First, some of the most interesting Darknet communities may only be accessed through payment of registration fees or referrals from existing participants (Benjamin et al., 2015a). This often subjects researchers to the risk of being defrauded, discovered, or otherwise compromised. Payment of registration fees could also directly support cybercrime, which may present ethical, if not legal, concerns. Both university administrators and law enforcement should be notified and consulted to stay within university policy and legal boundaries.

As with the previously discussed ethics concerns, there is no formal literature regarding what actions should be taken regarding this issue. Only a few generalizable recommendations can be made. First, a local IRB office should be consulted. Next, communities requiring referrals from existing participants should be considered off limits. Obtaining the referral would generally require some in-depth interaction directly with an existing Darknet participant, thus raising many ethics concerns. Lastly, for communities requiring payment, a judgement call should be made by the researcher and proper authorities when appropriate. For example, if an identified community contains data of great value to the Darknet research and cyber threat intelligence communities, it may be permissible to pay a registration fee. Note that researchers new to the Darknet context should avoid these issues altogether and instead focus on more accessible forums.

The second major reason that interaction with Darknet participants may be unavoidable is that some research methodologies require it. Action research, surveys, and interviews may all result in having scholars in direct contact with cybercriminals and other Internet miscreants (Flick & Sandvik, 2013; Barratt & Maddox, 2016). It is important to respect Darknet participants' anonymity in such cases, and subject consent forms should be signed when

appropriate. Careful planning is necessary to smoothly operationalize research and mitigate risks. It is also once again helpful to remain mindful of the aforementioned IRB concerns of privacy, data security, anonymity, etc. (Buchanan & Ess, 2009). For example, one strategy to protect subjects' identity is to censor their names in published works (Benjamin et al., 2015b). Coordination with local IRB offices is highly recommended.

Overall, there are moderate concerns generated by researcher interaction among Darknet forums. Most issues can be managed with caution and thorough planning. Researchers should feel comfortable reaching out to the proper authorities when necessary, as described in the aforementioned circumstances.

EMPIRICAL DEMONSTRATION OF DICE-E

The DICE-E framework is utilized to create an illustrative example that demonstrates the value of the research guidelines suggested in this manuscript. This example involves scrutiny of participant reputation within four distinct Darknet forums located in the United States, China, Russia, and Iran. The premise of this example is as such: prior work suggests that reputation plays a major role in Darknet communities (Motoyama et al. 2011; Holt et al. 2012). In general, possessing good reputation will lead to increased collaboration opportunities with peers and possible invitations to more private, highly-skilled communities. Many Darknet forums utilize internal reputation rating systems, as seen in Figure 4. These systems allow participants to rate the trustworthiness and contributions of others and assign negative reputation to scammers and other miscreants (Fallman et al., 2010; Benjamin & Chen, 2012). However, despite the importance of reputation within Darknet communities, there has been little work investigating the exact mechanism in which participants gain reputation among peers. Insights into these areas

would have value for security researchers and practitioners. To operationalize this study, we refer to the guidelines set forth in the DICE-E framework.

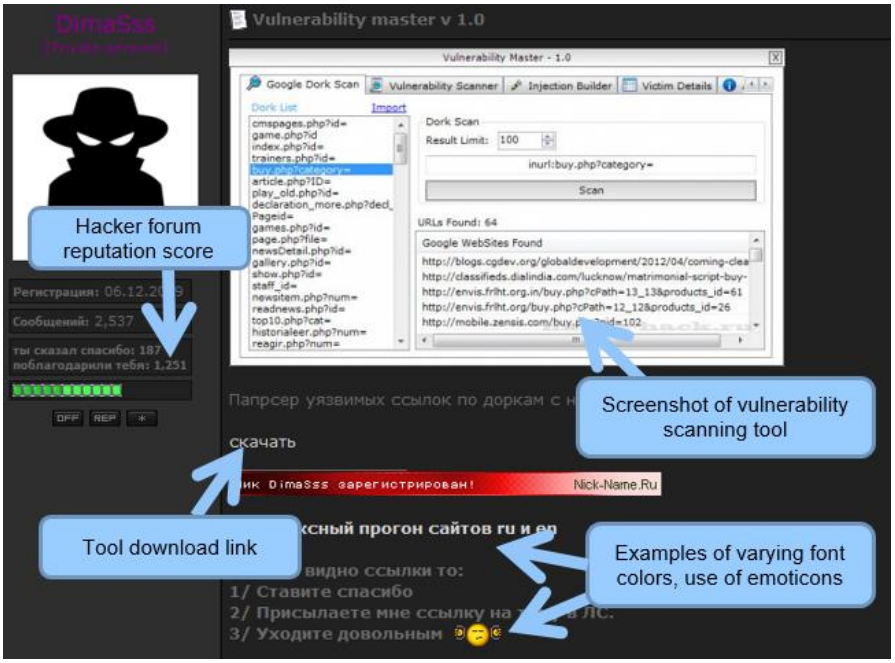

Figure 4 – Russian Darknet Forum Posting

Planning a Darknet Study

It may initially be unclear what steps must be taken to operationalize a given Darknet forum research project. In such cases, a process map that aligns with the DICE-E framework can be used to guide the development of a high-level research design (Figure 5). Steps to be taken are as follows. First, a secure research environment must be created, as recommended previously in DICE-E framework guidelines; doing so will minimize risks to the researcher. After establishing a secure environment, data identification, collection, and evaluation methods must be considered. Identified forums should be chosen for their activeness, population size, geopolitical origin, or other characteristics that match research objectives. Collection of the identified forums can be performed using web crawlers. Note that anti-crawling circumvention is not always necessary, as some Darknet forums do not implement technologies that impede crawling efforts. Finally, collected data should be evaluated using methods that can satisfactorily address research

questions. It is important to remain mindful of ethical concerns that should be evaluated at every stage of the research process. The process map can be used as a blueprint to operationalize the research design of many types of Darknet studies, such as the one described in this example.

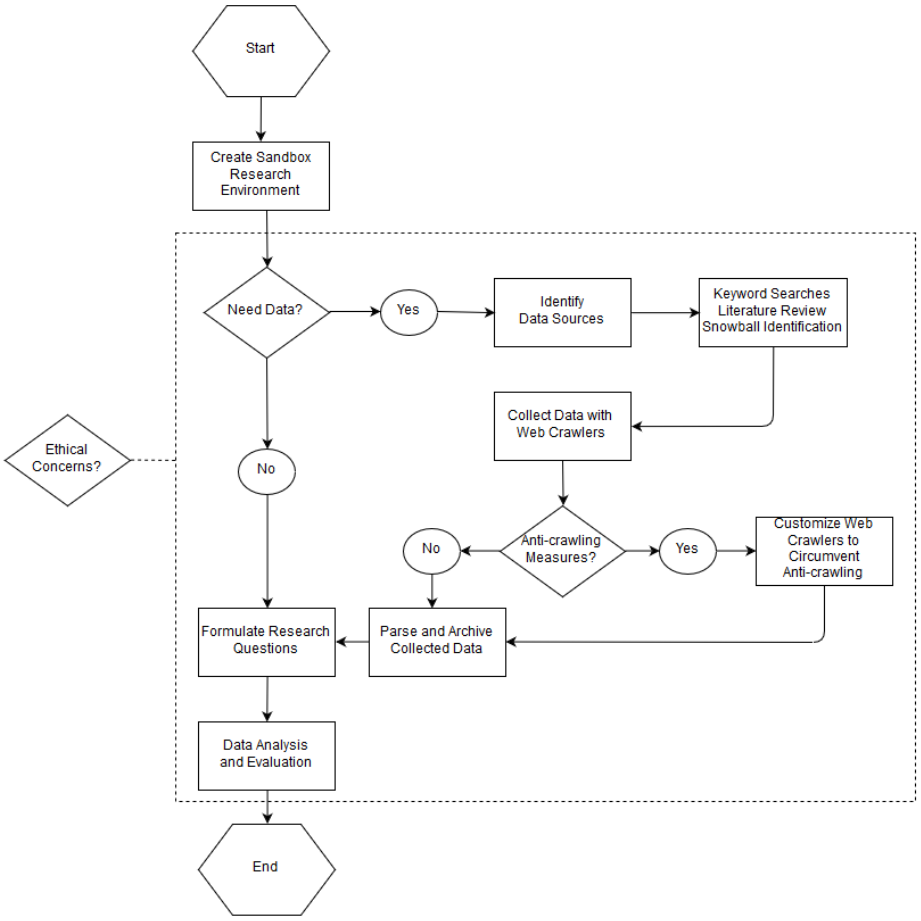

Figure 5 – Process Map for Darknet Research Projects

Identification

The selected Darknet forums are public cybercriminal forums that were discovered through a series of keyword searches as described previously in the DICE-E framework. Each one of the identified forums possesses a reputation system that forum participants use to evaluate peers, much like how reputation systems are used in more traditional virtual communities and markets (Ye et al., 2014). The reputation system provides a form of ground truth to help identify key

participants within each forum, allowing us to thus investigate the mechanisms of how such individuals become credible among peers.

A summary of identified forums can be viewed in Table 4. Note that because forums are natural archives of their own history, it is very common to visit a forum today and retrieve data from several years ago. This may result in large datasets, as demonstrated by this example study. Additionally, while some of the forums have aged considerably since they were first created, their functionality has remained constant. Users can use forums to transmit the same type of text, media, and cybercriminal assets now as they did several years ago, and thus techniques applicable to newer communities can be directly applied to many older ones as well. Collecting forums where data spans multiple years is generally non-problematic for these reasons. However, beyond forums, many other Darknet resources are generally shorter-lived and do not typically contain easily accessible archives. Researchers should expect less data from such sources unless historical records are explicitly made available by the platform.

Forum Name	Language	# of Messages	# of Users	Data Collection Start Date	Data Collection End Date
Antichat.ru	Russian	232,920	11,865	01-01-2003	1-2-2015
Ashiyane.org	Persian	12,903	2,922	08-26-2008	1-2-2015
HackHound.org	English	6,011	817	10-12-2012	1-2-2015
Unpack.cn	Chinese	521,101	18,840	11-12-2004	1-2-2015

Table 4 - Research Test Bed Summary

Collection

As described previously in DICE-E framework guidelines, web crawling programs were utilized to collect forum contents and mask researcher identity by routing crawler traffic through the Tor network. Using virtual private servers offered by cloud services was another possible identity obfuscation technique. After forum pages were collected, text parser programs utilizing regular expressions were written to extract relevant forum contents from downloaded webpages. Extracted data then was stored in a database for later retrieval and analysis.

Evaluation

After data collection and processing, exploration of research objectives can be performed. Recall that the goal of this example study is to scrutinize the mechanism by which Darknet participants can gain reputation among peers. Thus, literature was reviewed to identify a set of features that could be useful for predicting reputation. Two categories of features were developed based on the possible actions that can be taken by Darknet participants: message content features and more generalized forum usage features. Message content features encompass keywords that pertain to technical and cybercriminal-specific knowledge. High frequency of these keywords in a participant's messages may indicate expertise. Additionally, shared cybercriminal tools, source code, and other assets are also included as message content features, as cybercriminals can accumulate social capital by disseminating such assets. Conversely, forum usage features capture user behaviors and characteristics such as posting frequency, forum threads (i.e., conversations) started, seniority/tenure, message symbols supported by the web forum software, and so on. Table 5 contains a comprehensive list of all features used. The selected features coincide with the DICE-E framework's emphasis on computational and data-driven research methods.

Category	Feature	Description
Message Content Features	Attachment of Cybercriminal Assets to Forum Posts	Forum participants sometimes attach cybercriminal assets to their forum posts. These include written & video tutorials, hacking tools, cracked software, etc.
	Embedding of Source Code within Forum Posts	Cybercriminals sometimes share source code for tools, malware, etc., by embedding them directly into forum posts
	Discussion of Attack Vectors and Hacking Concepts	Demonstrates participant proficiency; Examples: *Rootkit, XSS, SQL Injection, DDoS, shellcode, PoC, drive-by*
	Discussion of Programming and other Technical Concepts	Demonstrates participant proficiency; Examples: *SQL, C++, ASM, .Net, XML*
	Reputation System Scores	Peer-evaluated cybercriminal reputation; reputation is not uniform across forums, i.e., a participant with good reputation in one forum may not necessarily have good reputation in another
Forum Usage Features	Message Symbol Diversity	Total usage of message symbols such as font color, font style, text bolding, italics, etc., at a per-message level
	Number of Threads Started	The number of threads a forum participant has started, normalized to the user's total number of posts
	Number of Posts Made	The number of posts a forum participant has made
	Seniority	The number of days a forum participant has belonged to a forum

Table 5 – Forum Content and Usage Features

In additional to the described content and usage features, four dummy variables are utilized, each corresponding to a forum's geopolitical region (China, Iran, Russia, and the United States). The use of such variables allows us to scrutinize any potential effects that geolocation may have on cybercriminal reputation.

To measure the impact of the extracted features on cybercriminal reputation, ordinary least squares (OLS) regression is utilized (Pohlmann & Leitner, 2003). The regression-based analysis is applied separately on each set of forum data. The results of the analysis are summarized in Table 6. Overall, it appears that contribution to the cognitive advancement of a community is directly related to reputation.

This example demonstrates the value of implementing the recommended guidelines to operationalize Darknet research. While there is much room for improved rigor and more advanced analytics in this example, the DICE-E framework can be closely followed to enable new Darknet researchers in successfully conducting studies in this space. An exciting potential for many works of importance and high societal impact is the resulting outcome.

	Antichat (Russia)		**Ashiyane (Iran)**		**Hackhound (U.S.)**		**Unpack (China)**	
	Estimate	*P-value*	*Estimate*	*P-value*	*Estimate*	*P-value*	*Estimate*	*P-value*
Attachments	0.1275	**0.0062****	0.0235	**0.0472***	0.0204	**0.0074****	0.0173	**0.0181***
Embedded	0.0194	**0.0150***	0.0149	**0.0342***	0.0686	**0.0456***	0.0151	**0.0021****
Tech_Terms	0.0106	0.5517	-0.0106	-0.2023	0.0032	0.1657	-0.0005	0.2822
Hack_Terms	-0.0149	0.4928	0.0049	-0.173	0.0045	0.1804	0.0064	0.533
Msg_Symbols	0.0323	**0.0291***	0.0224	**0.0321***	0.1105	0.1222	0.0018	0.1232
Threads_Started	-0.023	0.6956	-0.0049	-0.2549	-0.40201	0.3801	-0.0043	0.9054
Seniority	0.0056	0.8222	-0.0048	0.9976	-0.0903	0.8151	-0.0280	0.4190
Total_Posts	0.0145	**0.0024****	0.0249	**0.0082****	0.4864	**0.0023****	0.0343	**0.0026****
Geolocation	0.0032	0.8222	-0.0048	0.9976	-0.0143	0.8151	0.0218	0.4190
	Signif. Codes. ''<0.01, '*'<0.05**							
R^2	0.3184		0.4131		0.5299		0.3460	

Table 6 – OLS Regression Results

CONCLUSION

As cyber-based threats have become a significant factor that cause unforeseen, large-scale disruption of business operations and continuity, understanding of the Darknet is more critical

than ever. By studying Darknet communities and markets, business researchers can develop a multitude of new capabilities that will help advance this growing body of impactful work. There is a need for new techniques that support detection of emerging threats against businesses, identification of potential targets and victims, assignment of attack attribution, unveiling cybercriminal supply chains, and more.

Due to the challenges and limitations facing current work, a research framework for guiding Darknet community research is of immense value. In this manuscript, the DICE-E framework is described. The DICE-E framework contains a series of steps and recommendations that can help business researchers operationalize their own Darknet forum research within secured computing environments to minimize risks. Darknet identification, collection, and evaluation techniques are described. Additionally, a discussion of relevant ethics issues and recommendations is provided, as some research activities required for Darknet research may raise concerns.

Regarding Darknet identification strategies, a series of steps was outlined to help researchers locate their own data sources. Those who are just beginning Darknet explorations can start with keyword searches to identify public forums that serve as leads for more private communities. In particular, snowball identification procedures can help identify instances where participants of public forums share hyperlinks or other information that can be used to locate other hidden data sources.

Collection of Darknet forums is another nontrivial task. Web crawlers are excellent for automating the collection of web pages. However, many communities employ anti-crawling mechanisms that can inhibit crawler activity. In the DICE-E framework, several strategies are outlined that can help circumvent such anti-crawling mechanisms deployed by Darknet forums. Crawler strategies are discussed to ensure comprehensive collection of identified data sources.

DICE-E also strongly recommends researchers carefully document and report web crawler configurations in cases where anti-crawling is encountered, as doing so can help other researchers collect similar datasets for their own investigations.

The DICE-E framework recommends several computational and data-driven analytics that can be employed by researchers. Note that DICE-E is not intended to provide an exhaustive listing of all possible directions that can be taken, nor is it meant to be an in-depth discussion of the methodologies themselves. Rather, its purpose is to serve as an introductory point for researchers to begin thinking of research questions and analyses that could be possibly explored with Darknet data. Overall, research methods relevant to traditional virtual community research are typically applicable to Darknet forums. The DICE-E framework also provides recommendations on how researchers can handle cybercriminal assets unique to the Darknet.

Lastly, four activities related to Darknet research are identified as potential ethics concerns. These activities include research on underground communities, circumvention of anti-crawling mechanisms, identity obfuscation, and researcher interaction. Recommendations are made regarding possible ethics-related concerns for each research activity. In general, following the DICE-E framework will result in low concern about ethics pitfalls.

Overall, the Darknet possesses great relevance to business operations and continuity. This domain provides an opportunity for researchers to contribute high-impact works of great societal relevance. The Darknet and cybersecurity present opportunities for researchers to address critical challenges that make the rapidly changing business landscape even more complex. To the best of the authors' knowledge, no other formal literature exists that provides a comprehensive overview of these necessary components for large-scale, computationally driven Darknet research. The

DICE-E framework serves this purpose and is of great asset to business researchers who are interested or have decided to explore the Darknet for their own research.

REFERENCES

Abbasi, A. & Chen, H. 2008. "CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication," *MIS Quarterly* (32:4), pp. 811 – 837.

Adamic, L. A., Zhang, J., Bakshy, E., Ackerman, M. S., & Arbor, A. 2008. "Knowledge Sharing and Yahoo Answers: Everyone Knows Something," *Proceedings of the 17th International Conference on World Wide Web*. pp. 665–674.

Alfaro, L. & Kulshreshtha, A. 2011. "Reputation Systems for Open Collaboration," *Communications of the ACM* (54:8), pp. 81–87.

Anderson, C.L. & Agarwal, R., 2010. "Practicing safe computing: a multimedia empirical examination of home computer user security behavioral intentions," *MIS Quarterly*. (34:3), pp. 613-643.

Barratt, M.J. & Maddox, A., 2016. "Active engagement with stigmatised communities through digital ethnography," *Qualitative Research* (16:6), pp. 701-719.

Benjamin, V. & Chen, H. 2012. "Securing Cyberspace: Identifying Key Actors in Cybercriminal Communities," *IEEE Intelligence and Security Informatics*, pp. 24–29.

Benjamin, V., Chung, W., Abbasi, A., Chuang, J., Larson, C., & Chen, H. 2013. "Evaluating text visualization: An experiment in authorship analysis," *IEEE International Conference on Intelligence and Security Informatics*, pp. 16–20.

Benjamin, V. & Chen, H. 2015. "Developing Understanding of Cybercriminal Language through the use of Lexical Semantics," *IEEE Intelligence and Security Informatics*, pp. 79-84.

Benjamin, V., Li, W., Holt, T., & Chen H. 2015a. "Exploring Threats and Vulnerabilities in Cybercriminal Web Forums, IRC, and Carding Shops," *IEEE Intelligence and Security Informatics,* pp. 85-90.

Benjamin, V., Zhang, B., Nunamaker, J.F. & Chen, H. 2015b. "Examining Hacker Participation Length in Cybercriminal Internet-Relay-Chat Communities," *Journal of Management Information Systems* (33:2), pp. 482-510.

Buchanan, E.A. & Ess, C.M.. 2009. "Internet research ethics and the institutional review board: Current practices and issues," *ACM SIGCAS Computers and Society*, (39:3), pp. 43-49.

Carlson J. R. & Zmud, R. W. 1999. "Channel Expansion Theory and the Experiential Nature of Media Richness Perceptions," *Academy of Management Journal* (42:2), pp. 153-170.

Chen, H., Chiang, R. H. L., & Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165–1188.

Cook, T.D. & Campbell, D.T. 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin Company.

Cotton, A.H. 2004. "Ensnaring webs and nets: ethical issues in Internet-based research," *Contemporary nurse* (16:2), pp.114-123.

Cova, M., Kruegel, C., & Vigna, G. 2010. "Detection and analysis of drive-by-download attacks and malicious JavaScript code," *Proceedings of the 19th International Conference on World Wide Web*, pp. 281-290.

Daft, R. L., & Lengel, R. H. 1986. "Organizational Information Requirements, Media Richness and Structural Design," *Management Science* (32:5), pp. 554-571.

Décary-Hétu, D. & Laferrièr, D. 2015. "Discrediting Vendors in Online Criminal Markets," *Disrupting Criminal Networks: Network Analysis in Crime Prevention,* pp. 129-152.

Décary-Hétu, D. & Leppänen, A. 2016. "Criminals and Signals: An Assessment of Criminal Performance in the Carding Underworld." *Security Journal* (29:3), pp. 442-460.

Dennis, A. R., Fuller, R. M., & Valacich, J. S. 2008. "Media, Tasks, and Communication Processes: A Theory of Media Synchronicty," *MIS Quarterly* (32:3), pp. 575–600.

Dholakia, U. M., Bagozzi, R. P., & Pearo, L. K. 2004. "A social influence model of consumer participation in network- and small-group-based virtual communities," *International Journal of Research in Marketing* (21:3), pp. 241–263.

Fallman, H., Wondracek, G., & Platzer, C. 2010. "Covertly Probing Underground Economy Marketplaces," *Proceedings of the 7th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA),* pp. 101–110.

Flick, C. & Sandvik, R.A. 2013. "Tor and the darknet: researching the world of hidden services," *The possibilities of ethical ICT*, pp. 150-157.

George, J.F., Carlson, J., & Valacich, J.S. 2013. "Media Selection as a Strategic Component of Deceptive Communication," *MIS Quarterly* (37:4), pp.1233-1251.

Graham, L. 2017. Cybercrime costs the global economy $450 billion: CEO. *CNBC.* http://www.cnbc.com/2017/02/07/cybercrime-costs-the-global-economy-450-billion-ceo.html

Hirschberg, J., & Manning, C.D. 2015. "Advances in natural language processing," *Science* (349: 6245), pp. 261-266.

Holt, T. J., & Kilger, M. 2012. "Know Your Enemy: The Social Dynamics of Hacking," *The Honeynet Project*, pp. 1–17.

Holt, T. J., Strumsky, D., Smirnova, O., & Kilger, M. 2012. "Examining the Social Networks of Malware Writers and Cybercriminals," *International Journal of Cyber Criminology* (6:1), pp. 891–903.

Holt, T.J., Freilich, J.D. & Chermak, S.M. 2017. "Exploring the Subculture of Ideologically Motivated Cyber-Attackers," *Journal of Contemporary Criminal Justice* (33:3). pp. 212-233.

Hoser, B. & Nitschke, T. 2010. "Questions on ethics for research in the virtually connected world," *Social Networks* (32:3), pp. 180-186.

Hutchings, A. & Clayton, R. 2017. "Configuring Zeus: A case study of online crime target selection and knowledge transmission," *2017 APWG Symposium on Electric Crime Research (eCrime)*, pp. 33-40.

Jenkins, J., Anderson, B., Vance, A., Kirwan, B., & Eargle, D. 2016. "More Harm than Good? How Security Messages that Intterupt Make Us Vulnerable," *Information Systems Research* (27:4), pp. 880-896.

Kim, H.S., & Sundar, S. 2011. "Using interface cues in online health community boards to change impressions and encourage user contribution," *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, pp. 599–608.

King, R. and Wincup, E. 2008. *Doing research on crime and justice*. Oxford University Press.

Leavitt, N. 2009. "Anonymization Technology Takes a High Profile," *IEEE Computer Society*, pp. 15–18.

Li, W., and Chen, H. 2014. "Identifying top sellers in underground economy using deep learning-based sentiment analysis," *IEEE Joint Intelligence and Security Informatics Conference*, pp. 64-67.

Liu, L., and Munro, M. 2012. "Systematic analysis of centralized online reputation systems," *Decision Support Systems* (52:2), pp. 438–449.

Liu, X., and Chen, H. 2013. "AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events," *Smart Health*, pp. 134-150.

Mahmood, M. A., Siponen, M., Straub, D., Rao, H. R., & Raghu, T. S. 2010. "Moving Toward Black Hat Research in Information Systems Security: An Editorial Introduction to the Special Issue," *MIS Quarterly* (34:3), pp. 431–433.

Macdonald, M. & Frank, R. 2017. "Shuffle Up and Deal: Use of a Capture-Recapture Method to Estimate the Size of Stolen Data Markets," *American Behavioral Scientist*, pp. 1-28.

Martin, J. 2013. "Lost on the Silk Road: Online drug distribution and the 'cryptomarket'," *Criminology and Criminal Justice*, pp. 351-367.

Martin, J. and Christin, N., 2016. "Ethics in Cryptomarket Research," *International Journal of Drug Policy* (35), pp. 84-91.

Moore, T. & Clayton, R. 2009. "Evil Searching : Compromise and Recompromise of Internet Hosts for Phishing," *Financial Cryptography and Data Security*, pp. 256–272.

Motoyama, M., McCoy, D., Levchenko, K., Savage, S., & Voelker, G. M. 2011. "An analysis of underground forums," *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, pp. 71-79.

Nahapiet, J. & Ghoshal, S. 1998. "Social Capital, Intellectual Capital, and the Organizational Advantage," *Academy of Management Review* (23:2), pp. 242–266.

National Science and Technology Council (2011). *Trustworthy Cyberspace: Strategic Plan for the Federal Cybersecurity Research and Development Program*.

Pohlman, J. T. & Leitner, D. W. 2003. "Comparison of Ordinary Least Squares and Logistic Regression," *The Ohio Journal of Science* (103:5), pp. 118–125.

Qassrawi, M. T. & Zhang, H. 2010. "Client Honeypots: Approaches and Challenges," *4th International Conference on New Trends in Information Science and Service Science (NISS)*, pp. 19–25.

Radianti, J. 2010. "A Study of a Social Behavior inside the Online Black Markets," *2010 Fourth International Conference on Emerging Security Information, Systems and Technologies*, pp. 88–92.

Samtani, S., Chinn, K., Larson, C. & Chen, H. 2016. "AZSecure Hacker Assets Portal: Cyber threat intelligence and malware analysis," *IEEE Intelligence and Security Informatics*, pp. 19-24.

Sutanto, J., Palme, E., Tan, C.H. & Phang, C.W., 2013. "Addressing the Personalization-Privacy Paradox: An Empirical Assessment from a Field Experiment on Smartphone Users," *MIS Quarterly* (37:4), pp.1141-1164.

Thelwall, M., & Stuart, D. 2006. Web crawling ethics revisited: Cost, privacy, and denial of service. Journal of the American Society for Information Science and Technology, 57(13), 1771–1779.

Van Hardeveld, G.J., & O'Hara, K. 2017. "Deviating from the Cybercriminal Script: Exploring Tools of Anonymity (Mis)Used by Carders on Cryptomarkets. American Behavioral Scientist. 1-23.

Wasko, M. & Faraj, S. 2005. "Why should I share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice," *MIS Quarterly* (29:1), pp. 39–57.

Ye, S., Gao, G. & Viswanathan, S. 2014. "Strategic behavior in Online Reputation Systems:

Evidence from Revoking on eBay," *MIS Quarterly* (38:4), pp. 1033-1056.

Yip, M., Shadbolt, N., & Webber, C. 2013. "Why Forums? An Empirical Analysis into the Facilitating Factors of Carding Forums," *ACM Web Science,* pp. 453-462.

Zhuge, J., Holz, T., Song, C., Guo, J., & Han, X. 2008. "Studying Malicious Websites and the Underground Economy on the Chinese Web," *Workshop on the Economics of Information Security*, pp. 225–244.