Tor, what is it good for? How crime predicts domain failure on the darkweb

by

Bryan Monk
B.A., Simon Fraser University, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
School of Criminology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Summer 2017

Approval

Name:	**Bryan Monk**
Degree:	**Master of Arts**
Title:	**Tor what is it good for? How crime predicts domain failure on the darkweb**
Examining Committee:	**Chair: Bryan Kinney** Associate Professor
	Richard Frank Senior Supervisor Assistant Professor
	Martin Bouchard Supervisor Professor
	Garth Davies Supervisor Associate Professor
	Thomas Holt External Examiner Professor School of Criminal Justice Michigan State University

Date Defended/Approved: July 14th, 2017

Abstract

Content analysis of the darkweb shows the volume of illicit domains which are speculated to facilitate criminal activity. Describing Tor serves a valuable purpose, however does not allow for broader speculations about the criminogenic nature of the environment and the dismantling of the hidden services. Examining how the criminal content leads to domain failures is a small step towards providing insight into any casual mechanisms on Tor. The current study analyzes how 774 categorized domains explain website failure using a Cox repeated events regression while controlling for structure, popularity and size. Tor domain failure was found to be a function of popularity and size rather than criminality. Some criminally focused domains, however did survive longer on average than legal websites. The visibility of the domains may lead to increased costs financially as well as socially. The lack of infrastructure paired with law enforcement interventions may explain domain failures on Tor.

Keywords: Tor; darkweb; survival analysis; cox regression; crime severity index

Acknowledgements

The author would like to thank Dr. Richard Frank and Dr. Martin Bouchard for their patience and guidance during the thesis project. Dr. Garth Davies also made significant contributions, lending his knowledge of statistics in both the analysis and interpretation of results. The author would like to acknowledge additional support from by Dr. Evan McCuish and Jeffrey Mathesius on testing interrater reliability. Numerous graduate students within the Criminology department also deserve thanks for their support during the project: Mitchell Macdonald, Evan Thomas, Monica Ly, Krysta Dawson, and Melissa Gregg. Lastly I would like to thank my family for all of the stability and support they provided me during the last two years.

Table of Contents

List of Tables

List of Figures

List of Acronyms

CSI	Crime Severity Index
DNS	Domain Name System
FBI	Federal Bureau of Investigation
IANA	Internet Assigned Numbers Authority
ICANN	International Corporation of Assigned Names and Numbers
IoT	Internet of Things
ITU	International Telecommunications Union
SNA	Social Network Analysis
SFU	Simon Fraser University
TLD	Top Level Domains
TDC	The Dark Crawler
Tor	The onion router
URL	Uniform Resource Locator

Chapter 1.

Introduction

The Internet permeates nearly all spheres of life and provides an abundance of opportunities to connect users across the globe. Depending on knowledge and expertise, users can engage in fast communications and access boundless amounts of information and applications by delving into the depths of the Internet. It is understood that the vast majority of global Internet users only access a fraction of the web, known as "the surface web" (Lu, 2015). The "darkweb" is a concealed and difficult to access part of the Internet, consisting of unindexed web pages (Lu, 2015). The darkweb delves much deeper into the Internet and can only be accessed through specialized darkweb browsers or technologies (Mansfield-Devine, 2009). Darkwebs are the version of darknets which contain domains or websites. There are numerous darknets which exist strictly as peer-to-peer networking services or information/messaging exchanges such as Freenet. Numerous darweb softwares have been created and are available open source to the public (Misata, 2012). Mansfield-Devine (2009) identifies three central characteristics of a true darknet: 1) it is decentralized, often through the use of peer-to-peer technology; 2) it utilizes the infrastructure of the Internet; and 3) it operates through non-standard protocols and ports (p. 4).

A variety of individuals and groups around the world deploy darknets for a multitude of objectives. Darknets are primarily used to keep Internet activities and identity anonymous, evade censorship, and communicate sensitive information securely. Recent revelations into the extent of pervasive government surveillance and a growing public concern over privacy have undoubtedly increased the popularity of these virtual anonymity-protecting tools (Dredge, 2013). Darknet technology is also employed by corporations, military agencies, and law enforcement to survey or investigate online activity (The Tor Project, n.d.). At the same time, darknets also increase opportunities for individuals to engage in malicious activities and access illicit content or goods (Dredge, 2013). There are growing concerns over the extent to which darknets may be facilitating and fostering serious criminal activity, as well as assisting the operations of terrorists and violent extremist groups (Guitton, 2013).

As the popularity of darknets grow, enhancements to the technological capabilities of darkweb browsers have also progressed. Although law enforcement operations have shut down numerous darknets over the years, many have persisted and evolved. Several well-

known darkweb browsers have managed to sustain their operations and maintain an extensive user following despite increased monitoring by governments and law enforcement agencies. Darkweb browsers allow users to access an encrypted and contained part of the internet specific to each network. Although the objectives for using darknet browsers may vary, it is likely that the use of these sophisticated anonymity tools will persist. One of the most popular and enduring of the darweb's is known as Tor (The onion router). Tor contains websites or domains known as hidden services. These hidden services can only be accessed using a darkweb browser. The browser is a modified version of Firefox preconfigured for ease of use. Given its prominence, Tor will serve as the focus of this current study.

Tor's ascension to the premiere darkweb can be attributed to three factors: The first being its early creation allowing ample time for adoption from the general public. Second, Tor not only serves as a way to provide users with additional anonymity and security while browsing the internet it also contains its own community of websites known as hidden services. Lastly, the notorious hidden service known as The Silk Road helped Tor gain much of its infamy during its run from June, 2011 until it was shut down by the Federal Bureau of Investigation (FBI) in October, 2013 (Barratt, 2012; Dolliver, 2015). When discussing the cost benefit that Tor provides it is the hidden services which elicits a majority of the focus from researchers, law enforcement and governments (Guitton, 2013; Moore, & Rid, 2016; POST, 2015). The hidden services on Tor encompass a variety of topics which are discussed in depth further on in the paper, however it is this speculation which forms the basis for the current research. Numerous studies have been conducted in regards to the content which exists on Tor through the hidden services but fail to move beyond a descriptive level (Guitton, 2013; Moore, & Rid, 2016; Misata, 2012; Dredge, 2013). The problem lies within the general conclusions drawn from these studies which is that Tor is a facilitator of criminal activity and that the hidden services should be removed.

Drawing parallels to an offline environment, these assertions would be akin to counting all of the crimes present within a neighbourhood or community, deciding that the amount of illicit activity outweighs the legal uses and calling for its destruction. While the content analyses highlight a portion of the darkweb and allow researchers insight into some of its uses they do not provide any casual mechanisms for how Tor may be facilitating criminal activity. Compounding on this idea, to call for the removal of the hidden services (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014; Dredge, 2013; Guitton, 2013, Misata, 2012) without attempting to provide empirical support is disingenuous.

The requirements needed to fulfill all of the elements necessary to prove that Tor is inherently criminogenic are arguably vast, however this study seeks to provide one pillar which can form part of the foundation for a darkweb crime rate. Unlike traditional crime rates formed using arrest data and population statistics derived from the census (Allen, 2015) cybercrime does not share in the same luxuries of government funded programs. The data available to cybercrime researchers does not present itself in straight forward formats because of the global nature of cybercrime. Defining the space where a crime takes place within the internet is contentious. If a person in Vancouver is the victim of identity theft where their credit card information was stolen from a server in Atlanta by a hacker from Moscow, which jurisdiction is responsible for enforcing the law? The answer could be all of them, or it could be none given that the local policing agencies lack the resources and technology to pursue offenders. Generating arrest data to use within a crime rate then is implausible, or at the very least a tall task and one that would be incredibly time consuming. It would be necessary to aggregate available statistics from as many policing jurisdictions as exist to generate this metric. The dark figure of crime provides another challenge for reliable statistics as this phenomenon is likely exacerbated online (Tcherni, Davies, Lopes, & Lizotte, 2015). Instead of looking at arrest data specifically then it may be necessary to include additional sources including victimization surveys, or less traditional sources tailored to address cybercrime specifically such as the capture and re-capture of populations within illicit domains (Macdonald, & Frank, 2017) or conducting a content analysis of websites to determine the proportion of criminal activity.

Population data for the internet too is sparse and hard to collect where previously it may have been feasible to collect user numbers based upon assigned Internet Protocol (IP) addresses this is likely no longer possible. With the transition globally to smart devices such as microwaves, TVs, fridges, etc. connecting to the internet this number would likely now overestimate the population. The Internet of Things or IoT represents additional challenges to calculating crime rates as these devices acting autonomously can be utilized by hackers to commit crimes on their behalf. The IoT was used to attack Dyn, a domain name system (DNS) hosting company which disrupted service to millions of people in the east coast of the United States in 2016 (Newman, 2016). Additionally, not all IP addresses are assigned through public channels such as the Internet Assigned Numbers Authority (IANA) or through accredited agencies like the Internet Corporation for Assigned Names and Numbers (ICANN). Together these present challenges to calculating an accurate internet usage population although, the International Telecommunication Union (ITU) a United Nations agency has started to conduct

surveys to address this issue. In 2016, the ITU estimated that 3.5 billion people were using the internet based upon information derived from household surveys that were conducted. Obtaining the appropriate arrest data then serves as the last remaining hurdle to possibly generating a cybercrime rate.

Given the size and scope of generating a reliable and valid cybercrime rate for the internet at large, resources would likely be better spent focusing on a smaller sub population. The darkweb arguably provides an ideal place to develop potential methods to estimate a cybercrime rate although, this would require a conceptually different way of thinking about a crime rate metric which would have to include an amalgamation of different sources. While victimization surveys can be useful at uncovering some aspects of the dark figure of crime, for cybercrime, a significant portion of victims aren't even aware that they have had a crime committed against them (Tcherni, Davies, Lopes, & Lizotte, 2015). The central focus then for a cybercrime rate should be derived from observable phenomena which can be obtained and updated with regular consistency. This lack of access to an offender population means that the criminal aspect must be derived from non-traditional sources as mentioned previously. This requires transitioning from assessing the crime rate from only an individual level to one integrated with the group level. At the individual level it may be possible to estimate the number of offenders involved within a particular domain as uncovered by capture-recapture methods. At the group level where these populations are not present the domain itself may need to be treated as a criminal offender as found by a content analysis. By using a tertiary metric such as traffic analysis or transaction analysis triangulation may be achieved to determine if one-hundred sellers of stolen identities generates more income or traffic than a single domain soliciting the same information.

Tor unlike the regular internet has significantly more overt criminal activity present (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014; Dredge, 2013; Guitton, 2013, Misata, 2012). In a sample of 1 million domains on the regular internet it may be possible to include not a single instance of criminal activity where on Tor ~40% of hidden services are illicit in nature (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014; Guitton, 2013). The openness of the criminal activity on Tor means that this data is accessible and consistently available. Tor also captures the amount of known users, and the traffic of Tor on a daily basis (Tor metrics, 2017) due to it being a closed environment. Previous studies have examined the amount of traffic individually for domains which could be used to estimate criminal populations (Biryukov, Pustogarov, Thill, & Weinmann, 2014). The Tor metrics

combined with the proportion of criminal activity makes Tor an ideal candidate to generate a crime rate. While the generation of this rate on a much smaller sub-population of the internet is arguably more feasible, it is still a significant undertaking. The focus of this paper then is limited to assessing the implications of the domain content analysis. Previous studies have conducted content analysis in a static process which may misappropriate the impact of illicit domains. Before moving forward on a cybercrime rate, each element must be examined in greater detail.

The content analysis of Tor domains serves as a way to measure the proportion of illicit websites that exist at a particular time period within the network. The previous research designs, however treat domains as fixed entities which provide limitless access to illegal content while simultaneously acknowledging the transient nature of the domains (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014; Guitton, 2013). The inaccessibility of the content then diminishes the true impact of both the criminality of the domain itself and the ability of domains to allow users to co-offend. Legal domains too which link users to criminal domains with variable uptimes will reduce user access and subsequent potential co-offending. One way to assess domain stability is through the use of survival analysis. By implementing a longitudinal research design, domains can be captured over waves of similar lengths to determine how many domains are active after each time point. A cox regression can then be used to analyze what variables may predict domain failure. Assessing domain failures has been conducted previously with success using this methodological framework (Westlake, & Bouchard, 2016).

The previous study by Westlake and Bouchard focused on child exploitation (CE) domains on the Clearnet and compared failures to control networks of legal pornography and sports domains (2016). The results indicated that CE domains did not fail significantly more often than that of the comparison groups (Westlake, & Bouchard, 2016). Some variables, however did impact survival rates such as the volume of illegal CE images (Westlake, & Bouchard, 2016). If these results are comparable on Tor there should be no significant differences between domain failures of either legal or illegal content. Variables which measure the scope of the criminality on Tor such as the crime severity index should similarly to the volume of CE images, indicate more domain failures. Observations which differ from these findings may support the notions put forward by previous researchers that Tor facilitates criminal activity (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014; Dredge,

2013; Guitton, 2013, Misata, 2012). Due to the vastly different infrastructure landscape between Tor and the Clearnet, results which do not mirror the previous research seem plausible. The implications of Tor fostering a criminal environment, however is subject to the degree to which the network suffers from instability. Criminal activity may be more readily accepted on Tor and law enforcement are unable to intervene and shutdown domains in the same manner as on the Clearnet but, if domains lack stability, users are not able to exploit these advantages

Chapter 2.

What is Tor?

The onion router, known simply as “Tor,” is a peer-to-peer network that operates by connecting users through specialized software designed to anonymize and encrypt data sent between users. The network does this by routing traffic through other users running the software, peeling off a layer of encryption at each step along its path. Despite the existence of other darkweb browsers, the Tor network remains one of the most well-known and frequently operated sub-networks used to access and operate on the darkweb (Lu, 2015). The Tor network allows users to access the darkweb through a specially developed web browser. Since its inception in 2002, the Tor network has become one of the most universally used darkweb technologies due to its anonymity features and its ability to efficiently yield darkweb content (Misata, 2013). The Tor Project was originally developed in the 1990s by the US Navy and was primarily funded by the State Department and Department of Defense (Dredge, 2013). When the network became operational in 2003, the objective of Tor was for the US Navy to engage in open source intelligence gathering and conduct research on how the network could be improved (Dredge, 2013). The Tor Project, a non-profit organization based in the US, has maintained the Tor network since 2006. A diverse group of sponsors fund the Tor Project, ranging from private donors, human rights groups, and various international sciences and digital rights foundations (The Tor Project, n.d.). In 2014, the organization received funding from agencies such as the Bureau of Democracy, the Defense Advanced Research Project (DARPA) of the US Department of Defense, and the NSF (The Tor Project, n.d.).

The Tor network has maintained a steady user following due largely to its anonymizing architecture and easy accessibility. Tor is specifically designed to protect users’ identities and avoid traffic analysis of their online activities. Members of the Tor Project maintain that the network’s primary mission is to provide users the technology to evade intrusive surveillance and data collection by governments and corporations, while simultaneously fostering innovative research in online anonymity and privacy. The Tor software package is available to all Internet users, free to download, in a reasonably user-friendly format (The Tor Project, n.d.). In 2014, it was estimated that Tor was attracting 2.5 million users daily (POST, 2014).

Tor has elicited both admiration and condemnation for its anonymizing features. The developers of Tor state that the network's primary mission is to empower users by providing a way for users to have control over their security and privacy (Misata, 2013). The Tor Project (n.d.) outlines many positive applications of the network, such as allowing for uncensored communication and fostering valuable research on privacy, anonymity, and sensitive topics. Individuals living in countries that censor Internet browsing are able to use these safe channels in order to access or communicate sensitive and important information. Its developers maintain that Tor is critical in an era where reports of national and international private data breaches have become commonplace (Misata, 2013). While Tor has many positive capabilities, its anonymizing features and open-source availability also attract malicious users. The rapid evolvement and growing capabilities of the Tor network have been the subject of much debate surrounding the increasing prevalence of crime in the virtual realm. Critics of the network have focused on the role of Tor in facilitating the growing online presence of criminal and terrorist organizations (Misata, 2013; Dredge, 2013).

2.1. Tor architecture

The Tor network utilizes a layered encryption system, known as onion routing, in order to enhance the privacy of its users. This system operates by relaying communications between users through servers known as Onion Routers, which are all linked through Transport Layer Security (TLS) connections (Conrad & Shirazi, 2014). The Onion Routers contain two types of keys, long-term keys and short-term keys, each of which is related to a specific set of tasks. The long-term keys are used to sign the TLS certificates and contain both the router descriptors and their directories (Conrad & Shirazi, 2014). The short-term keys are used for decryption and to establish circuits which are the ordered set of linked Onion Routers a user's request travels through (Conrad & Shirazi, 2014).

When the user joins the Tor network, a virtual circuit is constructed out of a random selection of Tor nodes, and each piece of data sent into Tor from the source is encrypted as many times as there are Tor nodes in the virtual circuit. This ensures that as data is passed through this circuit, and a layer of encryption is removed from the original data at each step, the data still remains encrypted (The Tor Project, n.d.). The

virtual circuit is used for approximately ten minutes for sending and receiving data, before it is rebuilt randomly once again using a set of new encryption keys.

A virtual circuit comprises of three different types of nodes within the Tor network, and each of these nodes is operating an Onion Router. These three types of nodes are: entry nodes, intermediary nodes and exit nodes.

In order to first connect to the Tor network one must first establish a connection with an entry node. The entry nodes are found by identifying directory servers which list all known Onion Routers at the time of connection (Conrad & Shirazi, 2014). The directory servers are public and commonly known, and thus allows anyone to join the network. Once the user joins the Tor network, a set of encryption keys are used to encrypt each request to the Tor network. Thus, the communications between the source computer and entry node is encrypted multiple times, once for each node along the virtual circuit and is a shown in Figure 1. If a user's connection is compromised prior to establishing contact with the network, outgoing data requests made through the entry node can be intercepted.

The Tor software, known as an Onion Proxy, maintains a list of three entry nodes that are refreshed every 30 days (Conrad & Shirazi, 2014). If bandwidth exceeds the capacity of the three entry nodes then other nodes are chosen at random using a path selection algorithm contained within the Onion Routing Protocol.

The majority of nodes in the network are intermediary nodes, which serve as the links between the entry nodes and the exit nodes. If the source computer is requesting resources (website, for example) that is located in the Clearnet (outside of Tor), then the intermediary node simply passes the data through the established circuit to allow it to reach the destination outside of Tor. As the data is travelling through the virtual circuit, a layer of encryption is removed by each intermediary node, until it reaches an exit node where the final remaining encryption layer is removed. This is shown by path B in Figure 1.

Alternatively, a resource could be requested within Tor itself, in which case an intermediary node can be the destination for the original request, called a Tor hidden service (such as a website in the case of Tor an onion address). When a request is made for a hidden service the intermediary node becomes what is known as a rendezvous point. This node acts as a mid-point between the computer making the request and the server hosting the destination URL. This prevents either computer from knowing the true destination of the information in order to maintain anonymity. In this

instance, Tor traffic never leaves the Tor system. An example is shown by path A in Figure 1.

The last type of node in the Tor network is known as the exit node, which is the final point where a request from the user is passed out of the Tor network into the Clearnet. A request made by a user from the exit node to the Clearnet is no longer encrypted and does not contain the same protections as it did while traveling within Tor itself. This is the primary attack angle that researchers and cybersecurity experts have used to de-anonymize aspects of Tor in order to find individuals or hidden services operating within the Darknet (Hardesty, 2015). These attacks are possible due to the nature of how information travels through the established virtual circuits as governed by the Onion Routing.

2.2. Flow of Data

On the Tor network data travels through two types of cells instead of packets: control cells and relay cells (Conrad & Shirazi, 2014). The relay cells follow a fixed algorithm in an attempt to reduce latency, which allows attackers to trace paths and the timing of cells from when a message was sent to when a message is received. Control cells are in charge of maintaining and designing the circuits for the optimal paths that data travels through along the network. In the case of a user making a request using Tor to obtain information on the Clearnet, Tor can more easily be conceptualized and understood as a three-step process. First, a user launches the software, which provides a connection to the known Onion Routers, and a virtual circuit is established. The information request, or relay cell, is then sent from the entry node to the intermediary node, where a layer of encryption is removed by the short-term key contained within that node's Onion Router. The relay node then moves on to the exit node where the last layer of encryption is removed and the cell's payload is interpreted and the request is sent out into the Clearnet (Conrad & Shirazi, 2014). The message or information from the Clearnet is sent back to the exit node, which is the only known return address, where it is encrypted once again. The process then happens in reverse, with an additional layer of encryption added at each intermediary node, until finally the User's computer receives the heavily encrypted data and proceeds to decrypt it.

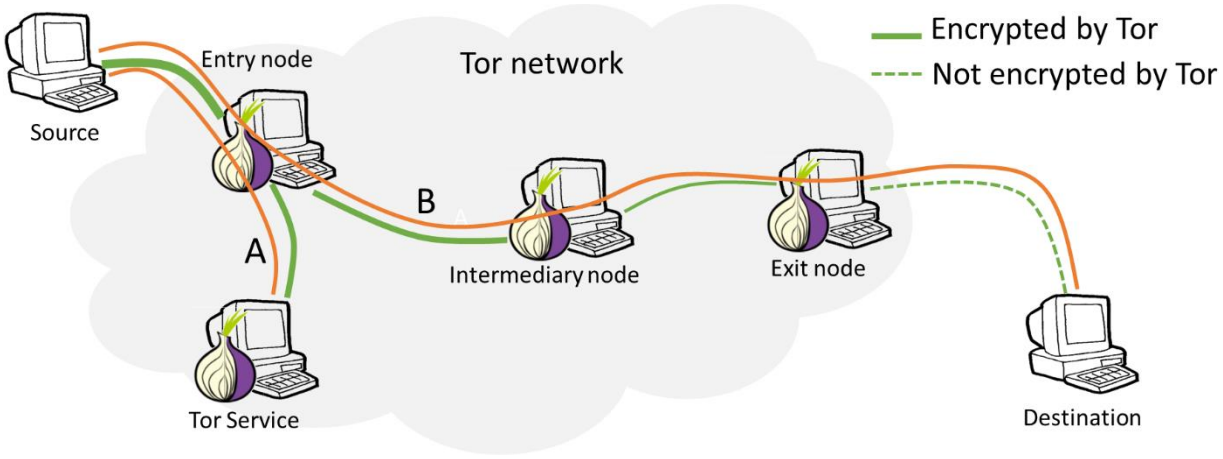

Figure 1. Flow of Data on Tor

Image: The Tor Project, Inc. *How Tor works,* 2017. Reproduced under Creative Commons Attribution 3.0 United States License.

2.3. Public key encryption

Public key infrastructure is a type of asymmetric cryptography which can be used to establish a secure connection between two users (Ellis, 1970; Diffie, & Hellman, 1976). Prior to the inception of the public key concept, cryptography relied primarily on the private key model which conversely utilized a symmetric model. Private key cryptography works by encoding data using an algorithm where both the sender of the data and the receiver are aware of the encoding and decryption keys (Ellis, 1970, Diffie, & Hellman, 1976). This works effectively if there is a secure channel in which the decryption key can be distributed safely to the person who needs to decipher the message. If no such channel exists then the encryption is virtually ineffective as anyone can intercept the message and decode it. This concept is known as the key distribution problem and formed the basis for the creation of public key cryptography (Ellis, 1970, Diffie, & Hellman, 1976).

As happens with scientific research on occasion two groups were simultaneously working to solve the key distribution problem at the same time (Ellis, 1970, Diffie, & Hellman, 1976). Both sets of researchers arrived at similar conclusions suggesting that a way to solve the problem was to involve the receiver of the data in the encryption process (Ellis, 1970, Diffie, & Hellman, 1976). The solution that was landed upon involved multiplying extremely large prime numbers to arrive at a product. This operates conceptually as follows, a user sends a request to another user or node which generates

an encryption key as well as a decryption key which are mathematically linked using these prime numbers (Ellis, 1970, Diffie, & Hellman, 1976). The encryption key in this case one of the prime numbers is sent to the original sender of the information who uses it to encode their message or data. The data can then be sent to the recipient without any fear of the message being intercepted and decoded as only the recipient has the decryption key which is the other half of the equation (Ellis, 1970, Diffie, & Hellman, 1976).

This is done in Tor to establish the original connection between the user and three recipient nodes. Each node has its own decryption key and peels away one layer of the encryption until the data reaches its final destination where the information is decoded (Biryukov, Pustogarov, & Weinmann, 2014). This circuit is re-established every ten minutes adding an additional layer of security. Nodes are only aware of the directly preceding and following nodes within the network. This is the standard functionality of Tor when a user leaves the network to examine websites on the Clearnet. When a user wants to find a hidden service within the Tor network this process is altered slightly. An additional node known as a rendezvous point serves as a bridge between the hidden service node and the client as mentioned previously. A hidden service domain name which may appear like gibberish (http://22222222ay7mhtbs.onion) is a sixteen character, hashed version of the public encryption key (Biryukov, Pustogarov, & Weinmann, 2014). Tor relays within the network maintain a list of the hidden services which are online and operational and are known as hidden service directories. The hidden service directory contains the descriptor provided by a hidden service and helps establish the circuit between the user, the website and the rendezvous point (Biryukov, Pustogarov, & Weinmann, 2014). These directories serve the same functionality as a DNS server on the Clearnet which provide the location (in cyberspace) of the website. This process is critical to understanding the fundamental differences between hidden services on Tor and domains on the Clearnet.

2.4. Hidden Services versus Domains

Domains on the Clearnet are a representation of two concepts which operate independently of one another. The first is known as Top Level Domains, or TLDs, and most people are familiar with which are the ending to a URL, for example .com. The

ending is the most critical piece as it directs the request of the URL to the correct location within cyberspace. At the time of writing, the TLDs are undergoing a massive expansion which will vastly change the landscape of the internet. The traditional TLDs represent three institutions: commercial, .com, .org, .net, government (United States), .edu, .gov and countries .ca, .uk, .br, etc. Each of these TLDs are managed by companies or organizations known as registries (Verisign owns the .com TLD) who allow secondary companies like GoDaddy known as registrars to sell specific domain names such as Google.com. ICANN is the organization responsible for overseeing this entire process and is currently allowing applications for new registries to own new TLDs such as .hotel, and .beer.

The TLDs, while overseen by corporations, are still exploited by hackers who engage in phishing scams through emails, Facebook, Twitter etc. where hyperlinks that look legitimate are in fact nefarious. The goal is to re-direct unsuspecting users to fake websites in order to gain their credentials for things like banking, or corporate information or use the fraudulent domains to inject malware directly into a person's computer known as a man-in-the-middle attack. As an example of this consider these two URLs: www.criminology.sfu.ca and www.sfu.criminology.ca. While they appear similar, one directs a user to a webspace on the sfu.ca domain, while the other to SFU's page on the criminology.ca domain which does not exist. This highlights the difference between domains on the Clearnet which have some oversight and those on Tor which do not. Tor domains as mentioned previously are simply manifestations of the hashed version of the public encryption key, they do not have any regulatory agencies monitoring the domains or corporations involved in the selling or distribution of them.

The differences in domain name creation between Tor and the Clearnet create an inverse relationship between the two spaces. By decentralizing authority over domain names as with Tor, security and anonymity are increased by taking nomenclature out of the hands of the registrar. Exploiting the lack of knowledge about domains through social engineering is less of a concern on Tor as no descriptor information is provided in the URL itself. Looking at a URL like the example provided above both appear to contain the same information while directing a user to different places. This also increases the anonymity of domains but decreases accessibility as the specific sixteen-character address must be known to the user ahead of time on Tor. The trade-off to deregulation is

that no entity is responsible for managing any content on the Tor network while, on the Clearnet Verisign works with law enforcement to remove specific illegal domains (ICE, 2015). This juxtaposition forms the basis for the arguments made by researchers that Tor is inherently criminogenic (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014; Dredge, 2013; Guitton, 2013, Misata, 2012).

2.5. Tor's Uses

Tor continues to be one of the leading anonymized darkweb technologies, utilized by a highly diverse community of individuals, groups, and organizations. Tor relies on these users and a daily flow of traffic for its peer-to-peer system to operate effectively (Misata, 2013). Tor also requires a large group of operators to hide users and act as relays to manage traffic on the network (Misata, 2013). Since its public release, Tor operators and users have emerged from a multitude of nations across the globe to communicate and interact online. In addition to the countries already noted (United States, Germany, and China), other top countries where users connect to Tor include Italy, Spain, France, Ukraine, and Russia (Misata, 2013). Users appear to have varying degrees of experience operating on the darkweb and use the network for a variety of purposes (The Tor project, n.d.). While some connect to Tor in order to safeguard their online privacy, other users are attracted to the anonymized network to engage in nefarious or illegal activities on the darkweb. Consequently, applications of Tor appear to encompass both legitimate and illegitimate activities to varying degrees. Vigorous debate continues to focus on users of online privacy-protection tools, and whether the privacy capabilities the network provides to users outweigh the dangers posed by criminal exploitation. It is important to investigate the Tor community in order for users and regulators to understand and consider the negative and positive aspects of the network.

2.5.1. Legitimate Uses

For many, Tor is a reputable system for engaging in unrestricted and borderless Internet use, where online activities are protected against surveillance or control. Tor users have been known to connect to the network to protect their privacy from corporations, protect the privacy of their children's online activities, or access information

that is blocked on the surface web (The Tor Project, n.d.). Corporations also use the Tor network to securely analyze competitors without such activities and traffic patterns being scrutinized in return (The Tor Project, n.d.). Perhaps the most proclaimed benefits of the Tor network are its ability to afford individuals and groups the privacy to access information and communicate information without restrictions or reprisal by governments. Many citizens, academics, activists, and journalists living under oppressive government regimes or highly censored nations use the Tor network to engage in free speech. The use of Tor has increased in some countries during times of burgeoning political turmoil or extreme government censorship. Most notably, the use of Tor greatly spiked in the Middle East during the Arab Spring demonstrations, when governments heightened levels of Internet censorship (POST, 2015). Through the Tor network, whistleblowers and human rights activist are also able to exchange and disseminate sensitive information to the public and media. Anonymized and uncensored communication permits individuals and groups to engage in open dialogue and preserve civil liberties without fear of persecution or interference from government or law enforcement (The Tor Project, n.d.; Misata, 2013).

Government and law enforcement also utilize the Tor network to monitor and intervene in illegitimate online activities. The US military has extensively deployed and advanced darknet technologies and practices. Military personnel use the network to securely gather and communicate sensitive information. The network also allows military personnel to conduct surveillance and hide their geographical location, which can be particularly useful when stationed in foreign countries. Law enforcement agencies have also been known to use the Tor network to monitor, detect, and remove illegal content by issuing takedown notices to illegitimate Tor users (McCoy et al., 2008). Agents are also able to conduct online investigations undercover and set up anonymous tip lines (The Tor Project, n.d.).

A branch of the NSA is specifically dedicated to exploring ways to de-anonymize suspicious Tor users on a wide scale. NSA representatives maintain that efforts to undermine the Tor network are focused solely on de-anonymizing cyber-criminals and terrorists. However, many have raised concerns that NSA will likely target and surveil citizens and activists as well (Ball, Schneier, & Greenwald, 2013). Potential debates of governmental intrusiveness notwithstanding, law enforcement agencies and the NSA have admitted that de-anonymizing Tor users poses significant challenges. Their

investigations have only been able to identify a very small group of users during these operations (Dredge, 2013). The most notable investigation by law enforcement to date was the October 2013 shutdown of the Silk Road by the FBI (Templeton, 2014). The agency was able to identify the IP address of the Silk Road operator and lay charges. Although the Silk Road 2.0 was launched a month after the shutdown of the original, this was subsequently shut down in 2014 and the operators were prosecuted. This shutdown was attributed to an international operation called Onymous. This operation involved the coordination of law enforcement agencies from the US and various European countries to develop and advance an attack on Tor in order to identify the Silk Road 2.0 server and other illicit markets (Templeton, 2014). While these types of law enforcement operations appear to shake up Tor criminal communities for a period of time, criminal behaviour and illegal markets are often quick to return (POST, 2015). Moreover, although law enforcement agencies and researchers have been able to infiltrate the Tor system and de-anonymize users, it is recognized by these groups that such attacks require a highly sophisticated level of computer expertise and a considerable expenditure of time and resources. Often once an attack vector is identified by research the Tor project patches the software and that avenue is closed (Tor project, n.d.).

According to the Tor project only ~3%-6% of the total network traffic on Tor is used to access the hidden services, this amounts to approximately 875Mbits/s (Tor project, n.d.). The remaining 94% of traffic is being used for other purposes like regular internet browsing. While some of the traffic is undoubtedly being used for peer-to-peer file transferring, and well documented Botnets communicating with command and control servers on Tor (Biryukov, Pustogarov, & Weinmann, 2013) not all traffic should be inferred as malicious. Tor's usage beyond the hidden services accounts for almost 100GBits/s of traffic (Tor project, n.d.). This leaves a significant portion of the Tor traffic unaccounted for and, without empirical evidence, can only be speculated upon, however it should be acknowledged that there is a significant discrepancy between the hidden services and Tor as a security tool.

2.5.2. Illegitimate Uses

While Tor has fostered an anonymous and safe communication platform for many individuals and groups, it has simultaneously attracted and enabled secure

interaction for users with malicious objectives. Critics of Tor contend that the network allows criminals and terrorists to run rampant and use its anonymizing capabilities for nefarious and illegitimate purposes. Past research demonstrates a varying degree of illegal activity on the network. McCoy et al. (2008) found virtual interaction and file sharing to be the most prominent illegitimate uses of darknets, including Tor. File sharing includes sharing copyright-infringing files, such as music, movies, and TV (McCoy et al., 2008). Cyber-criminals also have been known to utilize the network in order to exchange information and transfer data for hacking, identity fraud, and for buying and selling illegal goods.

Tor and similar darknet technologies allow users to access websites that sell illegal goods and services, similar to an eBay type of site. At one point, the most prominent black marketplace on the Tor Network was the Silk Road. The Silk Road provided a medium to sell and buy drugs, guns, identities, pornography, and other illegal commodities. While the Silk Road and the Silk Road 2.0 were shut down in 2013 and 2014, respectively, other black markets were quick to take their place and carry the same functions (Mansfield-Devine, 2014). As noted earlier, the third version of this site, known as the Silk Road Reloaded has emerged on other darknets, such as I2P, and provides a platform to buy and sell illicit material (Cox, 2015).

Black markets carry a variety of illegal commodities and appear to cater to users looking for specific materials. Most of these markets require Bitcoin as virtual trading currency. Bitcoin is an increasingly popular peer-to-peer decentralized payment system (Christin, 2012). The commodities found on these markets range from assassins for hire, child pornography, illegal drugs, stolen social security numbers, and other fraudulent identity information (Christin, 2012). The identification of various types of malware have also been identified on the Tor network. Such malware presents a highly dangerous threat to individuals and has involved ransomware that encrypts individual's files and prohibits access to these files until a payment is provided (Mansfield-Devine, 2014). This type of criminal activity functions as a particularly dangerous threat to other users and can be extremely challenging for law enforcement to track and charge those users responsible (Mansfield-Devine, 2014). It has been speculated that terrorist organizations operate on these markets to finance their activities through the buying and selling of illegal weapons and drugs.

The transactions themselves may also be shifting on Tor as noted by Aldridge and Décary-Hétu in a 2014 study, which found that the sales of drugs on a darkweb market more closely represented a larger distribution model. Individual to individual sales were still present but a large majority of transactions appeared to be between a whole sale distributor and a local distributor (Aldridge, & Décary-Hétu, 2014). This finding has implications about how large scale the markets on Tor have become and is supported by additional research which found on average markets are generating $300,000 to $500,000 USD per day in sales (Soska, & Christin, 2015). In contrast at its peak during 2012 The Silk Road was operating closer to 1.22million USD per month or ~$40,000 USD per day (Christin, 2012). Compounding on the transactional data between users, the operators of the domains themselves actually take a portion of the value from the sales (Christin, 2012).

Commissions generated from drug markets emphasize the complicit nature domains have in fostering crime on Tor. Barring any other illegal activity on Tor, the sheer volume of income earned through these commissions warrants consideration. Trust between users looking to make transactions on the internet is a fundamental concept (Jarvenpaa, Tractinsky, & Saarinen, 1999) which is only worsened by a highly anonymous encrypted network like Tor. This issue was resolved by operators on drug markets by reducing the trust that is needed between users for a successful transaction to take place. The drug markets instead asked users to trust in the website itself as their online reputation would suffer if they failed to deliver on their promises. Escrow allows the domains to be directly involved in each transaction that takes place within its walls. The buyer in each transaction sends bitcoins/money to the domain directly which holds onto the currency until the buyer verifies that the seller has sent the promised goods (Christin, 2012). At any point if either the buyer or seller is acting in a duplicitous manner the transaction can be cancelled and the funds returned to the buyer. For providing this service often domains charge a commission on each sale, the original Silk Road for example modeled the pricing schedule from eBay (Christin, 2012).

In January 2012, the Silk Road modified their pricing structure from a flat 6.23% to a tiered system which increased commissions 1.82 times (Christin, 2012). Near the end of the Silk Road's run researchers estimated the commissioners were earning $92,000 USDs per month (Christin, 2012). This result is likely affirmed after the arrest of the operator of the Silk Road, Ross Ulbricht in 2013 the FBI entered into evidence

documents which showed his earnings of roughly $1.1 million USD per year (Soska, & Christin, 2015). After the shutdown of the Silk Road decentralization of the drug scene occurred and numerous markets sprung up trying to capture the user base (Soska, & Christin, 2015). Soska and Christin identified 35 active drug markets between 2013 and 2015 and informed their study regarding the current state of the volume of sales generated on darkweb markets (2015). While not all drug markets take a commission, in a sample of 24 established Tor markets, 19 domains had a percentage fee, 1 domain operated by donation, 1 domain had no listed value, and 3 domains were unknown (Deepdotweb, 2017). The average commission across all 24 domains including the unknowns counting as 0s is 2.47%. Taking this conservative value and applying it to the $300,000 USD in daily sales identified by Soska and Christin from 2015 equates to $7,410 USD per day in commissions by drug markets. This translates to roughly ~$2.4 million USD per year generated by drug market domains alone. While drug markets are prominent within Tor they are still only a portion of the hidden services offered.

2.6. Content on Tor

In addition to usage patterns, a key research interested has been the content available on the Tor network. Drug domains are certainly prominent within Tor and are focused upon in many studies, (Soska, & Christin, 2015; Aldridge, & Décary-Hétu, 2014) however they only represent a portion of the darkweb. A content analysis by Guitton (2013) revealed that an extremely high volume of unethical content was generated through the network. Unethical content was defined as negative behaviour, and included: “anti-social behavior, drugs, weapons, hacking, cannibalism, bomb making, hit man services, black markets, and child pornography” (Guitton, 2013). In total, this illicit content accounted for 45% of all the hidden services found (Guitton, 2013). Drugs, black market goods, racial discrimination, and child pornography were found to be the most prominent topics found within forums functioning on the Tor network. Guitton (2013) concluded that the prominence and pervasiveness of this type of unethical content on Tor far outweighs the benefits of the service. However, these findings were based exclusively on content found on forum discussions of three main databases and are potentially limited in their generalizability. Guitton (2013) was also unable to assess whether the uncovered Tor content significantly differed from content found on surface web forums. Further, Guitton (2013) acknowledges the study lacked the ability to

accurately assess the jurisdiction and related legality of Tor content due to the network's ability to conceal the content author's location.

An additional content analysis study was conducted by researchers using a machine learning classification algorithm the results displayed in. (Moore, & Rid, 2016). The algorithm found that out of 5,205 .onion domains 1,547 or 29.72% contained illicit content (Moore, & Rid, 2016). Those results are similar to the ones found by Guitton (2013) and reflect that in the three years between the studies the amount of illicit activity remained relatively stable. The difference observed between the two studies is likely due to classification differences rather than major content changes. As shown in Figure 2, a large portion of the data in the study was not classified. A third study from 2014 found that illicit content may be as high as 44% similar to the results found previously (Biryukov, Pustogarov, Thill & Weinmann, 2014; Guitton, 2013). All three studies shared overlapping methodologies using machine learning classification to analyze the content (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill & Weinmann, 2014; Guitton, 2013) with two using a webcrawler design to capture a large dataset (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill & Weinmann, 2014).

The machine learning classifications were conducted using various programs, however none of the studies reported their test dataset validation numbers or most importantly their unclassified dataset validation. One study did say manual verification was performed with "good results" but failed to quantify that assertion (More, & Rid, 2016). Although manual analysis is far more time consuming in this endeavour, small, seemingly innocuous details that may provide insight into the content of a domain are better able to be assessed. The current study utilizes this manual validation in order to include sites that may not be entirely in English, and can be cross-validated with additional resources such as deepdotweb.com.

The previous research provides a snapshot into the Tor network at various times over the past few years, however the general conclusions that are drawn from these studies may be overreaching. Webcrawlers are valuable tools that assist in the datacollection process and have been used successfully by both industry and academia. As websites are collected for analysis the webcrawler makes an attempt to visit each domain and capture the data. If the domain is not online at the time of the attempted visit the domain is labeled as offline or the content is listed as none (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill & Weinmann, 2014). Additional attempts can be made to visit

the page however this can lead to a significant loss of data to the sample. In the Moore and Rid 2016 study, 3,503 of the 5,205 domains collected were listed under a “None” or “Other” category and were not classified. While the webcrawlers are able to assist researchers in gathering data, they must be supported by better research designs that mirror this dynamic environment.

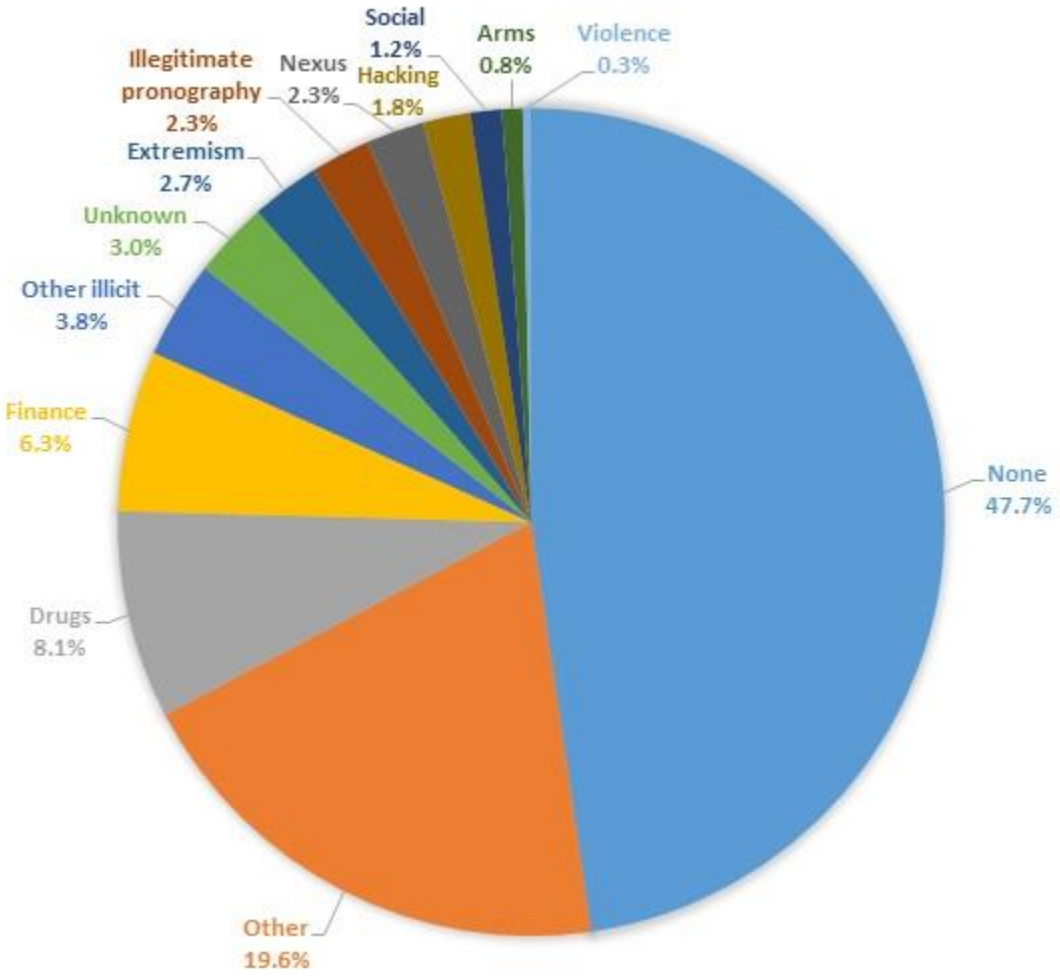

Figure 2. Content Classification (Moore, & Rid, 2016)

2.7. Domains as offenders

Websites allow users to access information, services, data, and products in the style and format conceptualized by the content creator. Domains themselves are inanimate objects, merely pieces of code constructed in such a way to facilitate the wishes of the author. The underlying architects for websites then are humans (for the most part), which is important for discussing the content on Tor. While studying the people using Tor, creating domains, and browsing content would be ideal, accessing that population, however would be incredibly difficult. As a network that operates with heightened anonymity it may even be dangerous for some users to participate in such a

study. In addition to the risks of accidentally outing individuals who may need Tor to conceal their identity, typical criminological variables used to measure concepts like co-offending aren't even available (Westlake, & Bouchard, 2016). Hidden services on Tor provide the window into this network and the people contained within. While websites themselves are not people, the designer of the domains and the communities which exist within each reflect the values of that space. Some domains are a single webpage and reflect the views of an individual while others provide space for thousands of users to share their voices. Each of these domains independently should be thought of conceptually as individuals with their own trajectories (Westlake, & Bouchard, 2016). The users likewise would represent potential co-offenders, as they access illicit domains they become accomplices in the criminal activity. Treating the illicit domains as offenders forms the basis for analysis to occur, moving studies beyond content descriptions on Tor in the hopes of ultimately providing the building blocks for a darkweb crime rate.

The framework for analyzing domains is also inherently dependent upon the linkages formed between the websites. One way to study the relationships formed between individuals in an offline context is done through social network analysis (SNA). The application of network theory to the social sciences makes an important assumption about the nature of relationships between individuals (Borgatti, Mehra, Brass, & Labianca, 2009). This assumption is that relationships matter and that individuals are deeply embedded within social networks, which forms the basis of SNA (Borgatti, Mehra, Brass, & Labianca, 2009). Two significant features of SNA are related to the types of relationships between individuals, also known as ties, and the overall structure that the network takes (Borgatti, Mehra, Brass, & Labianca, 2009). The types of relationships can vary between loose connections to intimate partnerships, and also lack reciprocity. A person who consider two others as friends may not be nominated back by those same individuals at the same level. The strength of the relationship can be coded as a binary process or it can be allowed to vary based upon different criteria. The position of an individual embedded in a network is also integral to making inferences about the flow of information (Borgatti, Mehra, Brass, & Labianca, 2009). The structures can take on many shapes which drastically change outcomes in processes like diffusion. In the context for the current study the hyperlinks would serve as a proxy measure for these relationship ties which has been used previously with effectiveness (Westlake, &

Bouchard, 2016). The users then are able to leverage these hyperlinks to expand their criminal networks and find new domains where they can co-offend.

These two fundamental transformations arguably form the basis for analyzing content in an online setting and specifically in this instance on Tor. Domains on Tor are either connected through hyperlinkages or exist primarily in isolation. The domains which contain hyperlinks between Tor sites, known as onions, provide characteristics similar to those found in offline social networks. As an example: consider this scenario, a darkweb directory site such as The Hidden Wiki may choose not to provide links to child exploitation sites, however those sites may provide links to The Hidden Wiki. This is similar to the example provided earlier explaining friendship ties but, in this case shows that although child exploitation material may exist on Tor it is not necessarily accepted as a social norm. Although the domains are not individuals specifically, the users' actions of the site do have implications about the environment to which they are embedded. In a similar vein, domains which hold strategic structural positions within Tor may control the flow of information in the network as a whole. This structure has additional importance when considering that on Tor, not only does a domain control the flow of information it can potentially act as a gatekeeper for users to traverse the network and access content. Network traversal allows users to jump from one domain to the next without having to have prior knowledge of the hashed public key used for the URL. The connectivity of the domains then will have a significant impact on how users are able to access criminal content.

2.8. On again off again

Tor domain generation as previously discussed in section 2.4 does not follow the standard procedure for the Clearnet. Absent of the regulatory structure provided by three levels of oversight, Tor additionally suffers from a lack of infrastructure. Hosting companies on the Clearnet readily provide data storage and server space for domains at little cost (Rackspace, 2017; GoDaddy, 2017; Deephost, 2017, Kowloon, 2017). Figure 1 provides comparable pricing for domain hosting between Tor and the Clearnet. The first noticeable difference is the lack of information provided by the hosting services on Tor, neither of the two largest services mentioned some key specifications for the servers. The storage capacity for the Tor hosting services are also severely lacking, as an

extreme example it would take 1000 Kowloon servers to equal one server provided by GoDaddy. Although Deephost appears to have a more reasonable pricing structure it is likely an attempt to regain market share rather than a standardized value. The domain for Deephost is redirected from the former hosting service known as Freedom Hosting II which was hacked by the vigilante group known as anonymous in February 2017 (Burgess, 2017).

Table 1. Clearnet vs Tor domain hosting

Location	Clearnet		Tor	
Company	Rackspace	GoDaddy	Deephost	Kowloon
Processing	2.5GHz/12cores	3.1GHz/4cores	Not Listed	Not Listed
RAM	128GB	32GB	Not Listed	Not Listed
Storage	5 x 600GB	2TB	30GB	2 GB
Bandwidth	2TB/m	Unlimited	Unlimited	Not Listed
Cost (monthly)	$749 USD	$507.97	$20.17 USD	$303.60 USD

The take down and exposure of Freedom Hosting II domains on Tor by anonymous forced the company to rebrand in an attempt to regain the trust of users. It is also possible that the site is capitalizing on the previous' hosts success and is a front for a fraudulent domain attempting to deceive clients either by law enforcement or offenders.

The acceptance of crypto currencies such as Bitcoin on Tor are likely due to the additional anonymity provided to transactions (Nakamoto, 2008; Aldridge, & DécaryHétu,

2014). This disconnection between the buyer and seller creates an element of mistrust as discussed previously which when coupled with inadequate infrastructure leads users to either host their own content or have to incur significant costs. The individual cost of hosting a server requires the additional technological knowledge necessary to do so which creates yet another barrier. These combined challenges make for a network which has been previously described as unstable (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014; Guitton, 2013). That is, the domains on Tor are not always available and online at any given moment. There are frequent interruptions to service where domains can be offline for days at a time. This is hypothesized to disrupt the network which has the effect of denying users access to content when domains acting as gatekeepers are offline. Domains will then appear back online days later and some linkages are restored between websites. The on again, off again nature of this process makes Tor difficult to assess.

Websites that do come back online could possibly do so under a new URL which poses another challenge to data collection. Should a domain captured in wave 1 that shows up in wave 6 under a new name but the same content be counted separately or as the same domain? For this study domains which entered the data collection later in the study were counted as new entities and not duplicates. This was done because of the substantial time investment required to build trust in these environments (Holt, Smirnova, Hutchings, 2016). Changing a domain name frequently would disrupt the process of building trust and limit the ability of the domain to create a userbase. The traffic to the domain too has a snowballing effect which brings in more users who verify and trust the domain which leads to more exposure and hyperlinkages. In order to validate the authenticity of some communications domains can implement PGP encryption to show that they are in fact the author of the new domains (Zimmermann, P.R., 1995).

Previous research has employed a static model of data collection where webcrawlers attempt to access a domain once and subsequently conclude that the domain is offline (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014). The same research will then concede that Tor is an unstable environment which frequently has these service interruptions without providing empirical support (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014; Guitton, 2013). The research

designs then are implicitly flawed and a more appropriate design would employ a longitudinal model for data collection. Ideally attempting each day to crawl a website would lead to more consistent results, however the sample of data collected would likely be limited due to the length of time required to crawl each domain. A weekly data collection would likely capture more significant service interruptions while allowing sites to have some downtime without being excluded from the study. Ideally no domains would be left to a none, or other category as reported in previous research (Moore, & Rid, 2016).

2.9. Domain persistence

The success of a website is not only contingent upon the content it provides to users which of course is subjective, there are additional structural elements which contribute to its longevity (Huizingh, 2000). Certain design features attract and retain more users and are targeted with a specific audience in mind. In a 2000 study on corporate domains researchers found that the primary aim was to appeal to investors (Esrock, & Leichty, 2000). A more recent study in 2003, found some subtle cultural differences between domains which used English versus other languages but, a relatively universal framework for web design (Robbins, & Stylianou, 2003). On the Clearnet website design can be divided into categories which are dependent upon the organizations' motivations. A corporate domain, as mentioned above, may focus on appealing to potential investors first while focusing on customers second (Esrock, & Leichty, 2000). The success of the website then is a product of the success of the company or brand. Domains which are not linked to corporate entities do not have these same underlying motivations, and may be driven by other factors such as advertisement revenue generated from user visits. Illicit domain success may also be linked more closely to customer interests and the evasion of law enforcement (Westlake, & Bouchard, 2016). The websites on Tor more similarly model the illicit domain focus with few corporations creating a hidden service and the lack of an advertisement revenue model.

The previous studies on domain success for corporate entities measured both the content and structure of the websites (Huizingh, 2000; Esrock, & Leichty, 2000,

Robbins, & Stlianou, 2003). One study did follow a longitudinal design conducted in two waves which found some domains offline in the second capture period (Esrock, & Leichty, 2000). In the original sample of 100 identified companies, 90 had websites and one year later only 88 corporations had maintained their sites (Esrock, & Leichty, 2000). Some of the companies originally identified had added domains, while some diminished due to mergers (Esrock, & Leichty, 2000). In this corporate setting domain survival then is more closely associated to the success of the company, and not vice versa, as zero domains were dropped due to the lack of usage of the website.

Non-corporate domains follow a slightly different model where shareholder and investors are not the primary target and instead are customer facing. The structural features of these domains also differed depending upon the content provided (Huizingh, 2000). Websites dedicated to information sharing differed significantly from financial domains in utilizing a search functionality (Huizingh, 2000). Search functions in modern domains are fairly common place and financial domains such as Goldman and Sachs have integrated them into their websites. Unlike financial domains, however the goals of the websites are to promote information or present ideas. Since the domains aren't always backed by a Fortune 500 company, survival may be linked to the ability of the author to solicit donations or generate income through advertisements. The costs of hosting domains on the Clearnet, while less expensive than Tor, as shown in Table 1, are still considerable. The domain then needs to implement both a structure which is easy to use, and fast (Robbins, & Stylianou, 2003) as well as producing content which appeals to potential users.

The advertisement model follows an opposite process to that of the content model where the goal is to get users to the domain but not necessarily retain them. This concept has given rise recently to the term known as click-bait. The content on the domain is substandard and the structure is designed to trap users into traversing as many pages as possible thereby generating more advertising income. The survival of these domains are linked to catchy headlines and trendy articles rather than producing quality content which brings users back to the domain. The infrastructure of Tor inhibits the ability for banner-based advertising to occur and thus domain survival is more closely linked to a content model.

The content model of domain survival is the idea that the consumers come back to a domain because the content or service provided matches the desires of the user (Huizingh, 2000). Domains may transition to meet the needs of the user base to accommodate new interests or to exclude others. The original Silk Road provides an example of this growth. As the drug market aspect of the site gained popularity, the small community which used the site for CE material came under spotlight. The Silk Road, through community feedback, closed down the CE part of the forums and implemented new policies prohibiting the sharing of those materials (Christin, 2012). The content on the domain was a direct reflection of the community values that the site was interested in maintaining. In a community of CE domains where more severe criminal elements are widely accepted, the content adaptation takes a different form. Domains which host certain types of content are not likely to be closely connected to similar others (Westlake, & Bouchard, 2016). The CE domains position themselves within the network structure to fill the niche of that particular community instead of attempting to provide everything all at once (Westlake, & Bouchard, 2016). Persistence within these communities then is driven by the content provided to the users and between the users themselves.

The naming convention of domains on the Clearnet as mentioned in section 2.4, differ drastically from those on Tor. Westlake and Bouchard posit that offenders will seek to retain a pseudonym online despite the perceived risks associated with doing so (2016). While the anonymity provided by Tor may reduce trust between users, the associated risks of using a particular username are also eliminated. Domains, like users on Tor, are subject to this enhanced security, so even though domain name generation varies drastically between both the Clearnet and Tor, the goals are the same. The trust gained by maintaining the same URL may be further emphasized in an environment like Tor where trust is difficult to attain (Soska, & Christin, 2015). Domains, like users, then have a continued online presence which have an observable life span, as noted by Westlake and Bouchard (2016).

The life span of an illicit domain may be characterized as its criminal career (Westlake, & Bouchard, 2016). The criminal career for offline offenders is a product of three factors: offending frequency, crime-type mix and co-offending (Blumstein, Cohen, Roth, & Visher, 1986). Westlake and Bouchard note that this framework needs to be

adapted for the online environment (2016). Offending frequency is difficult to measure in the online context as criminal events are not one-off occurrences and instead an illicit domain signals an perpetually on-going crime (Westlake, & Bouchard, 2016). Each time the event is measured it could be counted as a separate crime, however Westlake and Bouchard suggest that the volume of illicit materials on a domain better serves as this measure (2016). The continuation of the criminal event allows new materials to be added which signals a furtherance of criminal activity (Westlake, & Bouchard, 2016). This conceptually follows when measuring CE domains, however in the current context with the inclusion of many different types of illicit domains needs further modification. The criminal events on Tor are varied across a range of subjects, from simple drug possession to facilitating terrorist activities. The volume of illicit content then varies accordingly, how many marijuana transactions are equal to one CE image? The volume of the content then is subjective to the context of the domain in question. Instead of attempting to compare dissimilar content volumes, simply comparing the types of content in conjunction with structural variables serves this same purpose. As an example, a drug domain with a standard website structure is limited by the creator's ability to supply all subsequent orders for distribution while a drug market allows many users to buy and sell from each other. The volume of illicit content is related to the relationship between the content and structure variables.

The crime severity index variable serves as the proxy measure for crime-type mix in the current study and is similar to the one postulated by Westlake and Bouchard (2016). The severity of criminal offences, however, differs widely across many crime types and a standardized metric was utilized to account for these differences. There is no linear progression of crime severity within a crime type as discussed in the instance of CE materials (Westlake, & Bouchard, 2016). Domains which have multiple instances of crime are bound to the most serious offences found within that website. The co-offending opportunities provided by the domains are measured using social network analysis variables and are similar to those used previously (Westlake, & Bouchard, 2016). The degree to which domains are hyperlinked to others provide the mechanisms necessary for users to co-offend. Linking to non-criminal domains too provides benefits such as exposure and further co-offending opportunities (Holt, 2007). The impact of both crime severity and co-offending on offending frequency forms the basis for the current analysis. The hyperlinks themselves between non-criminal domains, although seemingly

inert, provide these crucial pathways which may link offenders to criminal domains. This is an indirect way to assess the ability of users looking to offend with the domains themselves. A legal domain which directs users to an illicit hidden service bridges this relationship and is arguably part of the co-offending network.

The results of the current study should reflect the unregulated environment which Tor provides, which should differ from the findings of Westlake and Bouchard (2016). The proportionality of criminal domains on Tor suggests that the community of users are complacent with the status quo and that serious criminal offences are not concerning. This environment would suggest that crime severity of domains should have no impact on domain failure and possibly may even contribute to its success or survival. Additionally, with inherent built-in anonymity and security provided by the network the popularity and size of the domains should also not contribute to domain failures. The hidden infrastructure costs, however, may account for discrepancies found within the social network metrics.

2.10. Current study

The darkweb encapsulates a centralized closed environment with overt criminal activity which creates an opportunity to move cybercrime research in a new direction. The early stages of a framework for the creation of a darkweb crime rate are discussed in previous sections. The current study aims to target one component of this by examining the stability of domains on Tor as speculated by researchers (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014; Guitton, 2013). More specifically the impact of the illicit domains on Tor will be assessed over time to determine if crime severity predicts a reduction in domain stability while controlling for the volume of illicit content and the indirect co-offending opportunities provided by the hyperlinks. If illicit domains are more likely to go offline for extended periods of time than legal websites, co-offending opportunities are indirectly reduced. This dampened effect would suggest that when determining the potential impact of illicit material on Tor the accessibility of the content should influence the overall score. A content analysis which reveals illegal domains comprise 45% of the sample have a significantly lower impact if a week later more than half of those domains are offline and inaccessible to potential co-offenders. If crime severity predicts a reduction in offending frequency the relationship should be

thought of as self-defeating and likely does not support notion of Tor as a criminogenic environment.

Chapter 3.

Data and Methods

The webcrawler project, at the time called *Child Exploitation Network Extractor* (CENE), started with a small research project focusing on mapping online child exploitation websites (Frank et al, 2010), with the focus soon shifting to the best ways of disrupting those networks (Joffres et al, 2011; Westlake et al, 2012). Research eventually shifted into the domain of violent extremism through various research projects and grants (Bouchard et al, 2014) mainly focusing on recruitment (Davies et al, 2015) and other uses of the Internet by extremists (Frank et al, 2015). The *Terrorism and Extremism Network Extractor* (TENE) was developed through the evolution of these projects.

Although CENE and TENE are in two different domains, and have two distinct names, the actual programming code-base is the same; the two are simply used to study different types of websites. Throughout each of the research projects, the code that makes up the webcrawler continued to evolve, and will continue to do so for the foreseeable future. Most recently, the focus has been on adding geolocation capabilities and WhoIS analysis (Allsup et al, 2015; Monk et al, 2015) through a grant from the Canadian Internet Registration Authority (CIRA). In a concurrent research project, the feasibility of sentiment analysis was used to try to guide the webcrawler by automatically differentiating various types of webpages from each other (Mei et al, 2015).

The current study was able to successfully develop and employ The Dark Crawler (TDC) for mass site collection, use keyword searches to automatically sift through and identify sites hosting specific content, examine network usage, and evaluate the data through social network and content analysis to gain an understanding of content prevalent on the Tor network.

3.1. Data Collection

The domains were captured using TDC developed at the ICCRC (TDC is explained in more details in a previous paper by Zulkarnine, Frank, Monk, Mitchell, & Davies, 2016). In short, the process is as follows. Seed websites are universal resource locators (URLs), which can be manually selected, or in this case, used from a different data collection period. Starting with the seed websites, the crawler downloads each top-level domain and adds any found hyperlinks containing an onion address to its internal queue. It recursively follows these links from the queue until either no onion domains are found or it reaches the termination criteria.

The data collection period consisted of a seed data collection period which took place in April, 2016, and the longitudinal data collection, which took place between December, 2016 and April, 2017. The seed data collection period was selected to generate enough onion domains to conduct the current longitudinal study. The termination criteria in the current study was 8 days, at which point the data collection was stopped and started over again for the next capture period. This meant that the seed websites from the original crawl were always used to start the data collection so the seed domain pool did not increase in each capture period. The final results of the longitudinal data collection period yielded 11,014 unique onion addresses. To facilitate manual classification, the sample used in this study consisted of 774 randomly selected websites.

3.2. Data coding

Domains were classified using six variables: the crime severity index (CSI), content, structure, degree centrality, page ranking and survival time. The first three variables are manual classifications done by three researchers. The next two variables are social networking measures that served as proxies and were used to assess the popularity and size of the domains. *Survival time* measures the persistence of each domain across the 11 waves of activity, while excluding the seed data collection (as this would have skewed the results). An additional grouping variable (group number) was created after the coding process had taken place, during which it was determined that many onion domains shared similar characteristics. The grouping variable was used to

reduce the number of duplicates which appeared in the sample; due to the possibility of domains changing URLs some duplicate websites (but not domains) may have ended up in the sample. Although the text on the domain appeared similar, not all of the bitcoin wallets were identical, perhaps indicating that one was a fraudulent version of the domain and not a duplicate of the owner. Many domains also contained variable survival times which suggests that although the content may be a duplicate, the underlying infrastructure that supports the domain differs. The domains could be hosted on different servers or some may be generating more traffic than others. Three additional subcategories were created: description, BitcoinWallets, and email addresses.

Although these three subcategories were mainly used to help assess the grouping variable, they may also provide insight into the ability of users to generate income through darkweb activity.

The manually classified data was established to help determine the types of content that comprise the Tor network. This assists the network measures in determining if the popularity and content of websites drives the survival time of a domain. An intraclass correlation (ICC) was conducted to determine researcher agreement for the manually coded variables. Due to the manner in which the data was coded by researchers, a two-way mixed model was selected to calculate the ICC. The results of the ICC showed strong researcher agreement of 0.82 across the variables. All of the collected domains were stored in a private online database that was only available to researchers at the ICCRC. In order to code the data a researcher would login to the database and select one of the crawls used in the analysis. Selecting a crawl populates a database with all of the collected domains stored as HTML code. This process removes images and some formatting that may be esthetic choices for the domains, leaving only the text. The text was then used to assess the crime severity, the content and the structure of the domain.

If there wasn't enough information provided by the text, researchers would search additional pages found within each website for more information, or visit the onion domain to corroborate information. If a domain was found that was not in English, the page was translated using Google translate prior to classification. Additionally, if no information was provided from the HTML code, the domain was searched using deepdotweb.com, which provides a repository of information for some onion addresses.

3.3. Variable descriptions

The first variable, *legality*, is binary in its coding: 0 for legal content and 1 for illegal content. Illegality was determined to be any content found within the site which violated a provision of the *Canadian Criminal Code* (1985). If a website discussed a safe place to promote Child Exploitation it would be considered legal; however, if it featured any discussions about how to carry out the crime or evade law enforcement after the commission of an offence it was classified as illegal. This binary classification posed a potential problem with the content variable, as the two could be seen as tautological. The legality variable then was recoded into a Crime Severity Index as determined by the Uniform Crime Reporting survey (UCR, 2016). Each crime within the UCR is weighted accordingly by the previous five years to develop the CSI based upon the proportion of sentences of incarceration for that particular offence and the mean incarceration time in days (Babyak, Alavi, Collins, Halladay, & Tapper, 2009; Appendix A). However, the CSI used in the study was established for crimes within the province of BC and will reflect severities established within this jurisdiction. Although a potential problem arises due to sentencing disparities, it does not vary significantly with some of the aggregated CSI weights across Canada (Appendix A). A domain then was classified by the most serious level of identified content found within the site. This allowed for variability within the domains as sites could have a certain type of content but differ within their severity. The CSI variable was characterized by kurtosis and was naturally logged to account for this problem.

The content variable represents one of the measures used to assess the volume of illicit activity and was coded into 11 categories: child exploitation, counterfeit (fraud and money laundering), privacy, down (offline), drugs, e-currency, filesharing (movies, music, books, pornography), hacking, hosting, directory, and weapons (firearms, bomb making, terrorism, hitman services). This differs from previous studies, which primarily focused their content on illicit activity with large samples of unclassified data (Moore, & Rid, 2016). Coupled with the CSI the content now does not directly implicate a certain status of the website. The *weapons* category as an example can be weighted depending on the severity of the content on the webpage as described in Appendix A. An onion domain which features the selling of firearms would be identified as weapons trafficking,

while a different domain featuring hitman services would be conspiring to commit murder.

Website structure was the final manually coded variable in the sample and served as the second metric used to determine the volume of content. Conceptually, websites, like neighbourhoods, have distinct characteristics that help to define the designated space. A website, or domain, that features a forum structure could be analogous to a neighbourhood or community that is tight-knit, where most people know each other and discussions happen frequently. *Market* domains are similar to neighbourhoods where land-use is primarily industrial spaces such as malls, or shopping centers, while websites categorized by *shops* could be conceptualized as neighbourhoods with small businesses or self-employment. The *structure* variable was divided into 5 categories: Blogs, social, financial, search engines (SE), and websites.

Blogs were personalized webspaces where authors sought to post opinion pieces often political in nature. These sites are structurally different from a traditional website in that they have chronological posts made by the author which can be observed over time. A traditional website does not usually have an identifiable author, content is not captured in a post format and changes can be made at random with no indication of time ordering of events. Social domains were an amalgamation of social networking sites, chatroom sites and webforums. Like blogs, social domains use different formatting, however the primary purpose is the interaction of users which makes these social sites take a much different shape. Financial sites were any domains which had as a central focus a way for users to buy or sell material goods, this contained both markets and shops. Search engines are structurally different from websites because they take on a fundamentally different role within the network. Users have direct input into search engines which direct them to content albeit with some difficulty on the darkweb. Search engines also require constant maintenance and a significant amount of resources to operate. Websites were deemed to be any other domain found which did not fall into the other categories.

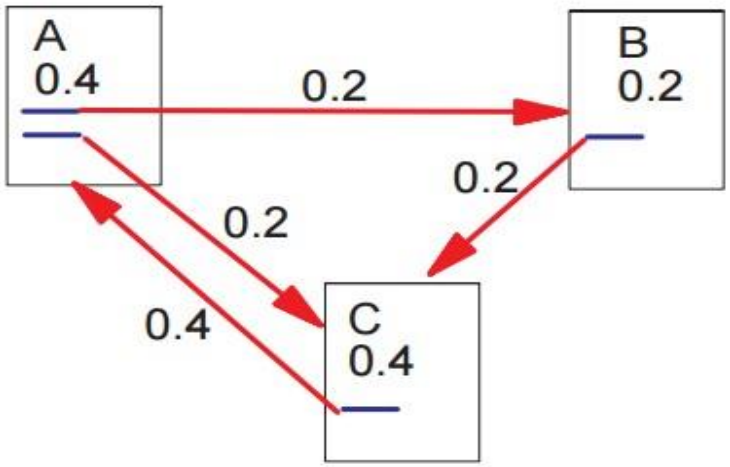

Figure 3. Page Rank Example

Image: Page, L., Brin, S., Motwani, R., & Winograd, T., *The PageRank citation ranking: Bringing order to the web*, 1999.

The network measures were calculated using Gephi 0.9.1 (Bastian, Heymann, & Jacomy, 2009). Degree centrality is an aggregate of the total number of incoming and outgoing ties a person has to other individuals. Degree centrality is used in this study to measure the number of incoming and outgoing hyperlinks a domain has, which serves as a proxy for the relationships between individuals. This concept has been used previously to identify the connectivity of child exploitation networks in an online setting (Westlake, & Bouchard, 2016). Building upon the idea of degree centrality, page rank is a measurement of the popularity of a node within the network (Page, Brin, Motwani, & Winograd, 1999). Both degree centrality and page rank scores were standardized within each network capture period to account for the variability in network sizes across time points.

The original Google page rank algorithm was a metric developed to assess the popularity of webpages to guide users to content. When a search was conducted on a keyword or phrase the page rank algorithm found the webpage with that specific keyword which contained the most backlinks (incoming hyperlinks) (Page, Brin, Motwani, & Winograd, 1999). Essentially, the webpage with the contained keyword that had the most pages linking to it was deemed to be the most popular and assigned the top ranking or the first result of the search as shown in Figure 3 (Page, Brin, Motwani, &

Winograd, 1999). Node A has links from both node C and B in a closed loop: this gives it a value of 0.4. Node C has links from both A and B, so it receives the same score, while node B only has one incoming link, for a relative importance of 0.2. This is a simple calculation done by summing all of the hyperlinks found as Google crawls each of those pages. This made searches incredibly fast but also systematic. The summation of the hyperlinks creates a relative probability that a random user would happen upon that domain only by following the backlinks listed on each website.

3.4. Survival analysis

Survival analysis is a statistical method designed to measure the time it takes for an event to happen from a starting point until an end point or failure (Garson, 2013). This technique is used in medical studies to track the efficacy of treatment options, as when patients start a drug treatment regime until they experience death or the end of the tracking period for the study (Garson, 2013). In survival analysis, the possibility that event (such as death) might occur after the end of the tracking period is referred to as censoring. The event of interest is usually coded as a binary process; in this study, the "event" of interest is whether the domain is online or offline. This was coded as 0 for online and 1 for offline to indicate that the domain had experienced the event of being inaccessible. Censorship in a sample happens when the event of interest happens outside of the data collection period. The situation above, when the event may occur after the end of the data collection period, is referred to as right censoring. Conversely, left censoring occurs when the event occurred prior to inception of the study (Garson, 2013). In the current study, we are not able to identify the true start date of each domain before it is included in the sample. The one-year lag period between the seed websites and the new domains does provide some context for this and will be discussed further. In contracts, domains which are online at the end of the survival period are characterized has having right censoring. They may go offline at some point in the future, but at the time of final data collection, they were still online.

Survival analysis has been used successfully in previous studies on websites. For example, Westlake & Bouchard (2015) found that child exploitation (CE) websites did not fail or go offline at a higher frequency than their control networks on sexuality and sports. This suggests that even on the surface web illegal websites were able to persist

with little to no interference from law enforcement. Given that illegal content is more prevalent on Tor than the surface web it is likely that illegal content too will not suffer from a reduced amount of time online. The survival rates for the CE content found on the surface web were an astounding 0.95, despite containing known pre-categorized CE images (Westlake, & Bouchard, 2015). The transitivity and instability of Tor has been widely speculated (Moore, & Rid, 2016); however no known studies have been conducted on the time Tor domains are online and offline. This is an important piece of research that will shed light on the accessibility of Tor as users are attempting to traverse the network to find content.

Popular websites with high connectivity may render themselves ineffective if they spend a majority of time offline to users. This is an important distinction from cross sectional analysis, which would show only the popular websites at one specific time period. In a subsequent wave if that website is offline its ability to disseminate information or bridge users to content is effectively halved. It is crucial then to distinguish not only the most connected and popular websites, but also ones which remain online and available for the longest periods of time. Stable websites provide the pathway for new users to find their way beyond directory sites like the Hidden Wiki. This research design means that a website is not considered failed if it is offline for one collection period because it has the ability to come back online at any moment. The analysis was conducted in two steps first to determine if Crime Severity predicted website survival using a Poisson regression in IBM SPSS 23 (2014). Survival was a count of the time periods in which a website was online at the time of data capture. The network variables degree centrality and page rank varied by time point so they were averaged across all 11 capture periods. This arguably overemphasizes the value of each domain, but this is corrected for in the survival analysis. The instability of the network was tested using a repeated events survival analysis in Stata 14. Repeated events allows for the fact that domains can fail more than once across the capture period this differs from traditional survival analysis which calculates the time to first failure. In a repeated events model time varying co-variates can also be used which eliminates the issues of using the standardized average social networking scores.

Chapter 4.

Results

4.1. Descriptives

The content breakdown of the Tor sample is shown in Table 2. The most commonly found sites were related to advertising the hosting of content (20.2%) and bitcoins/cryptocurrencies (19.5%). The next most common sites belonged to privacy and drugs, both of which represented 15.5% of the content found. Counterfeiting sites, which consisted primarily of credit card dumps, were the next most prevalent with 8%. Child exploitation content (3.0%), hacking (1.9%) and weapons (0.5%), all examples of more severe criminal content, were found in smaller quantities.

These totals are consistent with those found by previous researchers (Moore & Rid, 2016; Biryukov, Pustogarov, Thill & Weinmann, 2014) although not using machine learning lead to slightly different outcomes. In the current sample criminal content was found on at least 52.91% domains, or 408 onions total.

The structure variable has no known comparison to other research on Tor and is a new addition to website analysis built conceptually from work done on child exploitation sites (Westlake, & Bouchard, 2016) and corporate domain construction (Huizingh, 2000; Esrock, & Leichty, 2000, Robbins, & Stlianou, 2003). The most commonly found structure was that of a standard website (62.2%), with less than 5 webpages and content readily available without login. Financial based domains were the next most common, with structures that represented either a market where numerous individuals could post content for sale or shops where an individual user could sell materials (13.3%). The search engine structure (11.1%) was dominated by Grams, which allows users to search for products across the major drug markets. Only a few search engines were dedicated to the darkweb, with most being Tor versions of Clearnet search engines such as DuckDuckgo. Socially structured domains (8.3%) focused on enabling users to interact with each other in some format, which included forums, chats, and social networking sites. Finally, blogs where users would actively voice an opinion about certain subject matters made up 5% of the sample.

Table 2. Descriptives of Tor domains

	n(%)	Mean	SD
Content			
CE	23(3.0%)	-	-
Counterfeit	62(8.0%)	-	-
Privacy	120(15.5%)	-	-
Down	30(3.9%)	-	-
Drugs	120(15.5%)	-	-
eCurrency	151(19.5%)	-	-
FileSharing	38(4.9%)	-	-
Hacking	15(1.9%)	-	-
Hosting	156(20.2%)	-	-
Directory	54(7.0%)	-	-
Weapons	4(0.5%)	-	-
Structure		-	-
Blog	39(5.0%)	-	-
Social	64(8.3%)	-	-
Financial	103(13.3%)	-	-
SE	86(11.1%)	-	-
Website	481(62.2%)	-	-
AvgZPR	-	.043	.928
AvgZD	-	.051	.443
CSI_L	-	2.370	2.279
Survival	-	4.80	3.453

The social networking variables used in the study showed some overdispersion, which was caused by zero-inflated values when domains would be offline during each time point. The zero-inflated values unfortunately are unavoidable and do reflect the true nature of the data, thus were used for analysis despite this problem. The crime severity variable was also overdispersed, but in this case the overdispersion was corrected for by naturally logging the variable. The survival variable indicated the average amount of time points each domain was online for during the 11 waves. On average, domains were online for 4.8 of the 11 crawls, or 43.63% of the time. This number highlights the proposed instability of Tor domains, but on its own does not elude to any explanations for website failures.

4.2. Bivariate Correlations

The results of the bivariate correlations conducted to test for relationships between the independent variables and website survival are shown in Table 3. Although the crime severity variable was not indicated as a predictor for website survival, numerous content variables were found to be significantly related. Child exploitation had a weak positive association with website survival at $r = .08$ ($p < .05$). Hosting and eCurrency both had a weak negative association with website survival at $r = -.11$ ($p < .01$) and $r = -.11$ ($p < .01$) respectively. The structure variable was also weakly negatively associated with survival at $r = -.09$ ($p < .05$). The presence of significant effects in the bivariate analysis between the independent variables and the study variable suggests that further investigation is appropriate. The impact of each variable is explored further in the Poisson regression and the Repeated events analysis. Degree centrality and page rank both demonstrated a weak positive relationship with website survival $r = .16$ ($p < .01$) and $r = .13$ ($p < .01$), indicating that survival was related to both the size and popularity of a domain. The bivariate correlation matrix also indicated that there were no issues with multicollinearity in the data.

Table 3. Bivariate correlations

	CE	Counterfeit	Privacy	Down	Drugs	eCurrency	FileSharing	Hacking	Hosting	Directory	Weapons	Structure	AvgZD	AvgZPR	CSI_Log	Survival
CE	-	-0.05	-0.08*	-0.04	-0.08*	-0.09*	-0.04	-0.03	-0.09*	-0.05	-0.01	-0.14**	-0.03	-0.01	0.22**	0.08*
Counterfeit		-	-0.13**	-0.06	-0.13**	-0.15**	-0.07	-0.04	-0.15**	-0.08*	-0.02	-0.16**	-0.05	-0.01	0.27**	0.06
Privacy			-	-0.09*	-0.18**	-0.21**	-0.1**	-0.06	-0.21**	-0.12**	-0.03	-0.43**	0.06	0.04	-0.37**	0.04
Down				-	-0.09*	-0.1**	-0.05	-0.03	-0.1**	-0.06	-0.01	-0.25**	-0.04	-0.05	-0.21**	0.05
Drugs					-	-0.21**	-0.1**	-0.06	-0.22**	-0.12**	-0.03	-0.08*	-0.04	-0.01	0.52**	0.04
eCurrency						-	-0.21**	-0.07	-0.25**	-0.13**	-0.03	0.33**	-0.08*	-0.02	0.29**	-0.11**
FileSharing							-	-0.03	-0.11**	-0.06	-0.01	0.05	-0.03	-0.01	0.11**	0.01
Hacking								-	-0.07*	-0.04	-0.01	-0.07*	0.01	-0.02	0.12**	0
Hosting									-	-0.14**	-0.03	0.31**	-0.03	0.04	-0.52**	-0.11
Directory										-	-0.02	0.16**	0.29*	-0.01	-0.28**	0.07
Weapons											-	-0.04	0.01	0.01	0.09*	0.05
Structure												-	0.01	-0.03	-0.03	-0.09*
AvgZD													-	0.32**	-0.15**	0.16**
AvgZPR														-	-0.06	0.13**
CSI_Log															-	0.02
Survival																-

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$ & $^{+}p < .10$

4.3 Repeated events analysis

The last analysis used in the study was a repeated events cox regression. This analytic strategy deals with time in a manner that is different from how it was addressed in the Poisson regression. The repeated event analysis takes multiple failures into account and allows each domain to vary in relation to time dependent covariates such as degree centrality and page rank. The dependent variable in this analysis was not survival but instead failure, or a domain being offline. This was reversed to account for the possibility that failures may help to predict future failure. The time to each failure is also addressed in this model which will change the impact of the independent variables. Assuming that domains are following an unstable pattern as suggested by the results in the Poisson regression the current analysis better models the data. The results of the Repeated events cox regression are displayed in Table 4 and were run in three models.

The first model shows the effects of only the content variables on predicting domain failures. The log pseudolikelihood for the model was -4346.66 which indicates an appropriate fit for the data. The largest predictors unsurprisingly were down-domains which indicated they were already offline and only had placeholder text on the website (Exp(B) = 1.60, $p < .001$). Privacy focused domains and directory domains were the next largest predictors of failure at Exp(B) = 1.31 ($p < .001$) and Exp(B) = 1.39 ($p < .001$) respectively. FileSharing domains which contained more criminal based domains predicted domain failure at Exp(B) = 1.24 ($p < .01$). The next largest predictors of failure contained two subsets of content with primarily criminal motivations, counterfeiting and drugs. Counterfeiting predicted more domain failure at Exp(B) = 1.22 ($p < .001$) and drug focused domains also saw this increase at Exp(B) = 1.13 ($p < .001$). Hosting domains contained the least significant result and induced domain failure at Exp(B) = 1.06 ($p < .001$) indicating a small effect. Hacking, CE and eCurrency focused domains all saw marginal effects at predicting domain failures with hacking showing the largest effect size at Exp(B) = 1.23 ($p < .10$), CE next with Exp(B) = 1.14 ($p < .10$) and eCurrency with a small indicator at Exp(B) = 1.02 ($p < .10$).

The structure variables were run separately in Model 2 in order to determine the impacts of how website construction may impact domain failure (log pseudo likelihood = -4348.30). Domains with a financial based structure i.e. markets and shops predicted the

largest domain failure at Exp(B) 1.20 ($p < .001$) as compared to the standard website structure. Socially focused domains such as forums, chats, and social networking sites also significantly predicted domain failures at Exp(B) = 1.15 ($p < .05$). Search engines saw an inverse effect and predicted less domain failures at Exp(B) = 0.93 ($p < .01$).

Privacy domains significantly increase the hazard of going offline by 28% (Exp(B) = 1.28, $p < .001$). These domains were usually focused on anti-United States government sentiments and on helping other users be more secure and comfortable by using all of the features of Tor or how-to guides to create even more secure darkweb environments. The motivation for users to maintain these domains and keep them online may be more altruistic than financial, which could help to explain its failure. The effect was also reduced from Model 1 indicating that the structure variables and social networking variables are impacting the content. Websites that were offline were also more likely to fail Exp(B) = 1.57 ($p < .001$) than others. In the repeated events regression the results make more sense given that the offline domains do not host content and simply reflect the current status of the domain. A website that states that the previous content is no longer displayed at this URL does not have a vested interest in remaining online. Directory domains also predict failure Exp(B) = 1.28 ($p < .01$) at a higher rate than others. Filesharing domains also predicted a higher chance of failure at Exp(B) 1.22 ($p < .05$) along with drug focused domains Exp(B) = 1.31 ($p < .05$).

The structure variables had only one marginally significant result where search engines predicted a reduced chance of failure (Exp(B) = 0.81, $p < .10$). The standardized page rank significantly induced failure when it was allowed to vary (Exp(B) = 1.27, $p < .001$). Standardized degree centrality also predicted a significant increase in domain failure Exp(B) = 1.07 ($p < .001$) suggesting that maintaining a large website on Tor may lead to its failure. This result is congruent with that of directory domains predicting failure where popular and visible domains such as the hidden wiki may be at an increased risk. This visibility leads to an increased risk of failure as domains risk detection from law enforcement or incur greater infrastructure costs (Baker, & Faulkner, 1993). More successful criminal enterprises in the offline world too benefited from less visibility and notoriety, as individuals in these enterprises spent less time incarcerated and increased their monetary gains (Morselli, & Tremblay, 2004).

Table 4. Repeated Events Regression

	Model 1		Model 2		Model 3	
	B(SE)	Exp(B)	B(SE)	Exp(B)	B(SE)	Exp(B)
Content						
CE	0.13(.07)	1.14			0.15(.11)	1.17
Counterfeit	0.20(.06)	1.22***			0.12(.09)	1.13
Privacy	0.27(.05)	1.31***			0.25(.07)	1.28***
Down	0.47(.13)	1.60***			0.45(.14)	1.57***
Drugs	0.12(.03)	1.13***			0.27(.13)	1.31*
eCurrency	0.02(.01)	1.02^{+}			0.03(.04)	1.03
FileSharing	0.22(.08)	1.24**			0.20(.08)	1.22*
Hacking	0.21(.12)	1.23			0.22(.14)	1.24
Hosting	0.06(.02)	1.06***			0.02(.04)	1.02
Directory	0.33(.08)	1.39***			0.25(.09)	1.28**
Weapons	0.37(.36)	1.45			0.29(.37)	1.34
Structure						
Blog			0.05(.06)	1.05	-0.10(.08)	0.90
Social			0.14(.06)	1.15*	0.01(.08)	1.01
Financial			0.18(.06)	1.20***	0.10(.09)	1.11
SE			-0.07(.02)	0.93**	-0.21(.12)	0.81
ZD					0.07(.01)	1.07***
ZPR					0.24(.03)	1.27***
CSI_L					-0.01(.02)	0.99

**p < .05, **p < .01, ***p < .001 &*

The results of Model 1 indicate that most of the content variables do increase the likelihood of domain failure. The largest predictors of domain failure are domains already offline or down and directory domains. Neither of these two content types explicitly contain criminal activity which suggests the volume of content provided whether legal or illegal may contribute to the domain failures. Domains that indicate the hidden service no longer exists at that particular URL logically fail more often as that webpage is likely a place holder to inform users of its demise. There is no vested interest from the domain owner to continue to support the website beyond a small window of time. If infrastructure services additionally are not paid the termination of the domain will occur shortly after the payment period ends. The failure of directory domains differs from the previous content result as these hidden services provide hundreds of links to both legal and illegal domains. The directory domains are often well-known spaces within Tor, with The Hidden Wiki arguably the most popular. The content provided is simply hyperlinks, however both legal and illegal domains are linked too. The volume of content then, while legal, serves more-so as an indicator for co-offending opportunities. The hidden wiki, while not committing an offence, provides an unique connection between the user and the criminal domain. Without this hyperlink a user would have no pathway to the illicit domain without having the direct URL. The failure of these domains then suggests that exposure to a larger audience may contribute to their instability which is supported by the positive correlation found between directories and the average standardized page rank. These findings are partially supported further by the idea that users in stolen data markets actively take steps to avoid risks (Holt, Smirnova, Chua, & Copes, 2015). The users are aware of potential law enforcement interventions and forum administrators actively help to eliminate those threats (Holt, Smirnova, Chua, & Copes, 2015). As domains increase their exposure both on the darkweb and more generally they increase their risk of becoming a target despite the perceived benefits of operating on Tor.

The second Model also indicated that most domain structures contributed to their failure. Markets and shops predicted the most domain failure despite these enterprises having a financial interest in remaining online. In contrast to blog structures where the domain serves as a soap box for the authors voice the interest in remaining online is purely informational. A financial motivation for domains like the Silk Road to earn commission and the non-criminal benefits of gaining reputation and trust would arguably

incentivise domain survival over a personalized blog. This result like the one found previously with directory domains may indicate that the popular sites such as the drug markets on Tor are failing because of the increases in popularity and exposure of the domain. Search engines failing less often is supported by the arguments for infrastructure costs. Legitimate Clearnet domains such as DuckDuckGo have Tor based hidden services, these groups are able to support the financial costs associated with hosting a stable domain in this environment.

In the final combined model with the introduction of the degree and page rank variables variables the individual content effects saw a reduced effect. However, the model fit did improve slightly (log pseudo likelihood = -4323.90) Some content predictors were no longer significant and the structure variables too had a dampened effect. Drugs, Privacy and FileSharing domains retained significance from the first model and indicated more domain failure. Drug domains, especially drug markets, do have a large volume of illicit content and may implicate domain failure as a result. FileSharing domains too were often found to contain illicit material such as stolen books, and movies which supports the findings of the drug domains. Privacy domains contained both legal and illegal components, however the volume of content on these domains was substantially lower than those found on Drugs or FileSharing domains. These results arguably run contrary to the idea that Tor facilitates criminality. If the volume of illicit content increases and domains fail more often that either indicates the lack of community support for these hidden services or an additional explanation such as a lack of infrastructure to support the growing size.

The structure variables in Model 3 saw a reduction in some effects while Search Engines retained their inverse effect, although it became marginal. The social networking variables in this analysis were no longer averages for each domain and instead varied for each wave of the data. The standardized page rank significantly indicated an increase in domain failure suggesting that the most popular domains were the most likely to fail. The significant result of degree centrality additionally supports this idea that the most connected and the most popular domains saw increases in domain failures. These results, like the content effects shown above, suggest that domain failure may be attributed to the costs of exposure. Crime severity was also not found to be a

significant contributor to domain failures which at the very least indicates that the community on Tor does not actively suppress more extreme content and they may simply be indifferent. Whether indifference to severe criminal activity supports the notion that Tor facilitates crime, or opposes it, could be its own study.

The results of the repeated events Cox regression are visually represented as a Kaplan Meier curve in Figure 4 which show how each category of content remains after each wave of the crawl. Weapons and CE content have the highest cumulative survival rates, although the rate for CE is significantly lower than the results found in offline CE networks (Westlake, & Bouchard, 2015). Domains which provide hosting information had the lowest cumulative survival, followed closely by eCurrency and Drugs. Specifically, 12 of the 23 child exploitation sites (52.7%) failed, with a mean survival time of 6.57 periods. Two of the four weapons sites failed, leaving half of them online, with a mean survival time of 7.25 crawls. Directory sites faired slightly worse, with 31 of the 57 domains failing (57.41%) and a mean survival of 5.83 periods. Domains that indicated they were not online had a failure rate of 60% where 18 of the 30 were offline during data collection and they survived for 5.73 crawls on average. Onions dedicated to Filesharing failed at a 60.52% rate with 23 of 38 going dark and survived for 5.18 crawls. Privacy domains were offline in 62.5% of cases and 75 of the 120 total were offline with a mean survival time of 5.43. The last set of content in the middle cluster belonged to Counterfeiting domains and 64.52% failed during the collection period (40 of 62) with a mean survival of 5.69.

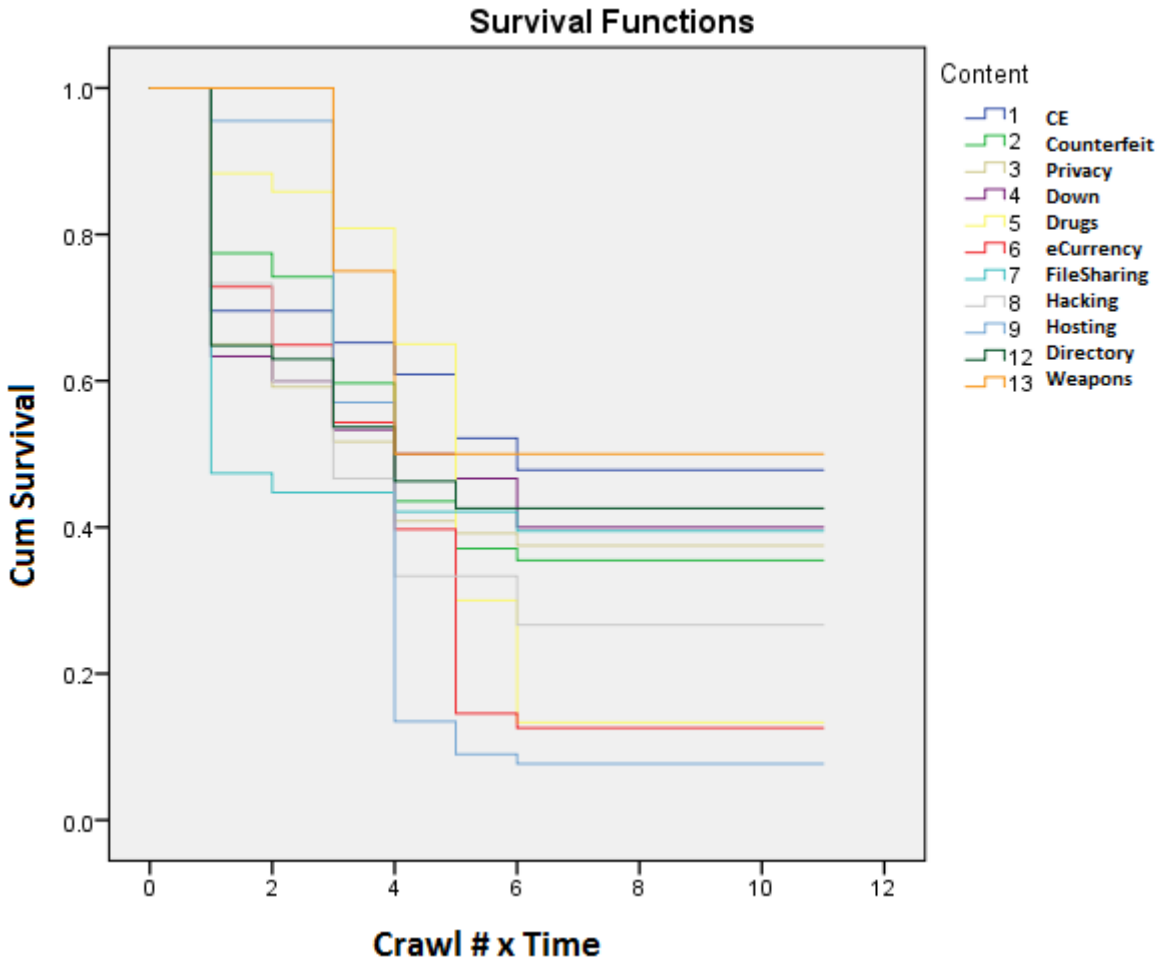

Figure 4. **Kaplan Meier curves for content**

There was a large drop in survival from counterfeiting to hacking domains which failed in 73.33% of cases and only four survived all the waves of data collection. The mean survival time dropped almost an entire crawl wave to 4.8 waves. Despite the assertions that the darkweb is full of drug domains, they also failed at a significant rate in 86.67% of cases: a total of 104 of the 120 collected, failed. The mean survival for drug domains was 5.17 crawls. Websites that pertained to eCurrency failed at a slightly higher rate than that of drugs at 87.42% with 132 of the 151 domains going offline and the mean survival time also dropped from 5.17 crawls to 4.09.. Lastly, domains which focused on hosting content failed at the greatest rate at 92.31%, where 144 of the 156 domains failed during the 3 months of data collection. The mean survival time for hosting domains was consistent with the eCurrency domains at 4.09.

Chapter 5.

Conclusion

5.1. Discussion of findings

In the Tor sample presented in this study of domain failure was primarily driven by visibility and exposure. Selected criminal domains containing more severe content were also found to have longer survival times when treated as fixed events. The lack of finding similar results in the repeated events cox regression demonstrates the importance of allowing time to vary, and speaks to the instability of the network. Looking at the Tor network on any given day does not present a representative sample of the network, especially for predicting stability. There are many plausible explanations for why domains with higher visibility fail at an increased pace, with three primary theories presented below: the first is that law enforcement is able to disrupt network traffic for these domains and is based upon findings from offline criminal networks (Morselli, & Tremblay, 2004; Baker, & Faulker, 1993). The perceived criminal content distribution on Tor drives the second explanation, which includes disruption from online vigilante individuals or groups such as Anonymous. The effects of disruption for Tor domain stability by Anonymous is shown in Table 5. During Crawl 176, or wave 8 of the data collection, Anonymous publicly breached the hosting service known as Freedom Hosting II, at which point it knocked ~15-20% of all onion domains offline (Burgess, 2017). This had a significant impact on TDC's ability to traverse the network and collect domains, as shown in Table 6.

Crawl 176 saw a massive reduction in the domains found compared to the twotime points directly preceding. The disruption caused by Anonymous did have an effect, more than halving the available domains. The fact that the network has slowly recovered speaks to the resiliency of Tor; however it has not yet fully recovered to the original levels of Crawl 50, which found more than 6,200 domains. There was also a large increase of new domains found in Crawl 180, which likely drove the recovery of the network to the pre-attack levels. This suggests that attacking the network and trying to

disrupt content may have a limited effect in reducing the accessibility of Tor. The resiliency of the network may be tested by examining the effects of either random or focused network attacks, which has been used in simulation for clearnet CE networks (Joffres et al, 2011).

The desistance of offline offenders described by Sampson and Laub (1993) considers the social influences on a person's life course of criminal activity. These social influences may affect users in much the same way, but don't necessarily apply to domains. Long-term domain desistance appears to be factored around exposure, traffic and finances while short-term or temporary desistance may be linked to infrastructure. This consideration is less present with offline offenders, and is similar to temporary desistance seen when individuals are incarcerated (Warr, 1998). The absence of offending then is not a product of rational choice by the offender or domain, but instead is a consequence of being involuntarily blocked from being able to do so. The degree to which domains experience temporary desistance however is more extreme than typical incarceration for offline offenders. Daily or weekly interruptions to domain stability may more closely mirror how some offenders make frequent contact with law enforcement and are released quickly thereafter. The pattern of desistance described above does not necessarily indicate desistance, as described in its original form, as the cessation of offending (Sampson & Laub, 1993). It is also dissimilar to revised explanations suggesting reductions in criminal offending may signal desistance (Laub & Sampson, 2001). A reduction in the volume of illicit materials on a domain is likely a better indicator of the desistance process. Some darkweb markets may fall into this category as they gain notoriety and an expanding user base often will prohibit child exploitation materials. The reduction in crime severity could be considered a component of the desistance process rather than as a binary concept. The frequency of offending or stability of the domain has a substantially lower social cost than its traditional counterpart. An offender doesn't sit inside of a stolen car for 23 hours per day and only exsts the vehicle due to external factors.

Table 5. New Tor domains found in each webcrawl

Crawl ID	50 (1)	157 (2)	162 (3)	171 (4)	173 (5)	174 (6)	175 (7)	176 (8)	177 (9)	179 (10)	180 (11)	181 (12)
Domains	6,282	1,105	4,459	2,393	2,303	5,271	5,496	2,255	3,912	3,533	4,857	4,269
New	0	300	1186	91	55	335	521	262	217	523	1,333	121

The final explanation relates to the infrastructure costs associated with hosting a Tor website. As mentioned above, few hosting services exist to provide users with storage and bandwidth to host onions. Freedom Hosting II is the second iteration of itself, which was also shutdown previously and has restarted again under the Deephost domain. This lack of hosting options means that providing content is up to the individual user, who must have both the financial resources and technical skills to setup a hidden service. Lacking external assistance provided by for profit companies on the Clearnet, Tor suffers from a lack of resources, which, as addressed earlier, may be offset by globally reduced infrastructure costs. Tor's greatest weakness then may be its own success, as visibility and notoriety increases the chances of interventions from law enforcement and vigilante groups alike. Further, the additional costs associated with the lack of infrastructure may strain the ability of Tor to expand. The development of more hosting companies on Tor may hold the key to the expansion of the hidden services and should be examined in greater detail.

The results of the repeated events model show that crime severity does not significantly predict domain failure on Tor. The implications of this in a broader context suggests that crime severity does not diminish co-offending opportunities. If crime severity predicted website failure it would imply a cyclical model. Severe criminal domains, such as weapons trafficking, would provide less frequent criminal opportunities for co-offending further inhibiting the impact of that website. Conversely if crime severity predicted an increase in domain survival it would suggest that not only do sites containing content like CE material stay online longer, they also provide more co-offending opportunities to users. That result would support the notion of crime facilitation put forth by researchers but according to the current results are not supported (Moore, & Rid, 2016; Biryukov, Pustogarov, Thill, & Weinmann, 2014; Dredge, 2013; Guitton, 2013, Misata, 2012).

While the current study does not support the assertion that Tor facilitates criminal activity necessarily, the results do support the idea that the darkweb is unstable. The contribution of content analysis to the developmental framework for a darkweb crime rate then is contextual. Domains cannot be treated as static entities which represent both a criminal offence and a co-offending opportunity. A domain advertising hitman services with a 10% uptime, negligible bandwidth traffic and 0 transactions associated

with a cryptocurrency wallet should not be used to inform policy. The presence of the domain in some cases is the entirety of that criminal element while in others is only a small piece. A drug market such as the Silk Road where the domain administrators earn 92,000 $USD per month in commissions is only a fraction of the 1.22m $USD generated by the site (Christin, 2012). The co-offending opportunities provided by the domain arguably dwarf the instance of the crime being committed specifically by the website. As an example a website hosting CE materials where known abuse images (not any less deplorable) were presented to attract co-offenders is less severe than if the users themselves bring new images or perpetuate new offences to present on the website. . The co-offenders have a substantially larger impact on the crime commission process than the domains themselves but could not commit the crime without the space presented by the website to do so. The domain instability then impacts to possibly a much greater degree of co-offending opportunities for users rather than a single instance of a crime. As a comparative example, two Filesharing domains where one operates structurally as a webforum and the other as a standard website will have different impacts on crime facilitation. A standard website may have the domain author provide ten free movies available for download while a webforum may have ten different people provide the same movies creating significantly more opportunities for co-offending.

The results from the current study can be used to influence policy, but should be considered conservatively. Crime severity does not appear to deteriorate domain stability indicating indifference towards these more serious websites. This does not, however indicate that more criminal domains necessarily are garnering more traffic, growth, or use on Tor. The content analyses conducted previously fail to address this issue and like the current study should only serve as an informative research article rather than a policy driving piece for governments. Law enforcement may use the current research to develop a strategy which can help identify domains with high levels of stability on Tor. These domains may serve as the lynch pin for bridging users to criminal domains and indirectly facilitating co-offending. Law enforcement attacks could aim to disrupt these domains at the ISP level, hosting level, or through compromising users.

5.2. Limitations and future research

Conducting studies using novel technology on unexplored data sources leads to some important limitations. Tor domains that were used in the sample only represented ones that could be found, raising questions about generalizability. This is partially corrected for by the way TDC traverses links in recursive function. Seeding the crawler with domains found from outside sources, as was done with the original sample, helps to eliminate this potential bias. The absence of comparable studies beyond content analyses similarly limits the generalizability of the study involving network stability. Additional studies are needed to measure longer trends, such as month to month variability within the network.

The crime severity variable is subject to sentencing discrepancies within the jurisdiction from where it was calculated. This has the potential to be problematic; however, in the current study the importance of the variable was the rank ordering of the crimes allowing the more severe domains to have a larger impact than the less severe hidden services. Further the CSI does not take into account the size of the domain in question. The structure variables acting as proxies for the volume of illicit content, may, mitigate the lack of a control for the size of the domain, but by no means is a perfect measure. An interaction effect may be necessary to create intermediary variables to account for the discrepancies in crime severity between small and large domains. The coding for the CSI was done by associating the most severe content found on the domain with a criminal offence. It is not a multiplicative or additive process and only counts the highest ranking severe offence. Not all domains had an immediately identifiable counterpart listed for the criminal offences so the best available offence was found. The coding was more conservative if it was difficult to locate an offence. The best example of such is the case of the domains listed as bitcoin gambling sites. The sites suggest that by paying a small amount of bitcoins they will “double your bitcoins in 24 hours.” It sounds too good to be true, and it is, instead of simply adding bitcoins from a reserve pool, thereby doubling the volume, the site washes the bitcoins as part of a money laundering scheme. Your bitcoins are traded for illegally obtained bitcoins through payments in order to obfuscate the trail of money. On the surface, this site is a gambling site, coded very low on the CSI, upon further investigation it is revealed to be a money laundering scheme, which is significantly higher. There were no listed offences

directly for money laundering however, so the aggregate offence listed as “Offences relating to currency” was chosen to represent all bitcoin washing domains. Multiple instances of the same offenses were not counted directly by the CSI but instead coupled together using the structure and content variables.

The results do not directly answer the question about the criminogenic nature of Tor, but attempt to make a contribution to assist other studies which may be conducted. Calling for the shutdown of the hidden services on Tor without further research is analogous to counting all the crimes in a neighbourhood and suggesting that the community be isolated and removed from society. The knowledge base on Tor thus far is lacking even something as simple as a crime rate. Without knowledge of the proportion of illegality versus the legal implementation and uses of Tor, it is disingenuous at best to suggest its cessation. All of the elements needed to calculate a crime rate are present and could be collected in a properly conducted research design and should be the focus of future studies.

The impact of the co-offending metrics on domain failure suggests that the network structure has a significant role in criminal activity. A content analysis falls short by not including any network variables. Future studies on Tor could explore additional avenues for co-offending. Examining the communities of domains by using a clustering analysis could reveal potential effects of domain traversal on co-offending. If a smaller community that is highly linked together provides criminal opportunities across a wide range of crime types the proportion of illicit domains in the entire network becomes less important. Each community could then be mapped and a traffic analysis using the hidden service directory could determine which communities are using the most bandwidth. Together coupled with the volume of illicit content localized community crime rates could be calculated. Amalgamated these rates could shed light onto the scope of criminality on the darkweb.

5.3. Concluding remarks

The mean survival time for domains on Tor of 4.94 crawls is deceptively low. Hidden service failures can be accounted for in two ways. First, analyses could employ a traditional cox regression model, which counts failures as the first time a domain goes

offline. This vastly over-estimates domain failure, as onions can be brought back online in the next crawl period. To account for this re-emergence, time to each failure must be taken into account. As well, covariates that change from crawl to crawl, such as social networking variables, must also be allowed to vary. Network stability can then be predicted using this model, which shows that visible and popular domains primarily drive failure on Tor. There may be hidden costs to hosting content both financial and social that help to explain this result. The results also suggest that calling Tor instable may be misleading. Domains failure is neither simply binary nor necessarily permanent: the majority that fail once also come back online and may fail again. The current sample is right-censored, so this pattern may repeat on a much larger scale or cycle beyond what this paper is able to speculate on. If Tor had been unable to rebound when domains failed, such as during the attack by Anonymous, then declaring the network instable might be more appropriate. The addition of new domains and the gradual climb of the total sample size in the waves after the attack show Tor's resiliency. While the instability may prevent the widespread adoption of Tor by the general public, or at least inhibit rapid growth of Tor users, it does not seem to deter the core base of users. If the prohibitive infrastructure costs turn out to be financial rather than social then the global increases in computational power and bandwidth speeds may eventually offset these limitations.

References

Aldridge, J., & Décary-Hétu, D. (2014). Not an 'Ebay for drugs': the Cryptomarket 'Silk Road' as a paradigm shifting criminal innovation. Available at SSRN 2436643.

Allen, M. (2015). “Police-reported crime statistics in Canada, 2015”. *Juristat.* Statistics Canada Catalogue no. 85-002-X. (accessed, May 01, 2017).

Allsup, R., Monk, B. & Frank, R. (2015). Geo-Mapping and Social Network Analysis. IEEE/ACM ASONAM 2015 Conference, Paris.

Babyak, C., Alavi, A., Collins, K., Halladay, A., and Tapper, D. (2009) The methodology of the police-reported crime severity index. *Proceedings of the Survey Methods Section.* SCC Annual Meeting, June 2009.

Babyak, C., Campbell, A., Evra, R., & Franklin, S. (2013). Updating the police-reported crime severity index weights: refinements to the methodology. *Household Survey Methods Division.* Methodology Branch – HSMD-2013-005E/F.

Baker, W. E., & Faulkner, R. R. (1993). The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. *American sociological review*, 837-860.

Ball, J., Schneier, B., & Greenwald, G. (2013, October 4). NSA and GCHQ target tor network that protects anonymity of web users. *The Guardian*. Retrieved from http://www.theguardian.com/world/2013/oct/04/nsa-gchq-attack-tor-networkencryption

Barratt, M. J. (2012). Silk Road: eBay for drugs. *Addiction*, *107*(3), 683-683.

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Biryukov, A., Pustogarov, I., Thill, F., & Weinmann, R. P. (2014, June). Content and popularity analysis of Tor hidden services. In *Distributed Computing Systems Workshops (ICDCSW), 2014 IEEE 34th International Conference on* (pp. 188193). IEEE.

Biryukov, A., Pustogarov, I., & Weinmann, R. P. (2013, May). Trawling for tor hidden services: Detection, measurement, deanonymization. In *Security and Privacy (SP), 2013 IEEE Symposium on* (pp. 80-94). IEEE.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *science*, *323*(5916), 892-895.

Bouchard, M., Joffres, K., & Frank, R. (2014). Preliminary Analytical Considerations in Designing a Terrorism and Extremism Online Network Extractor. *Computational Models of Complex Systems, Intelligent Systems Reference Library, 53*,171-184

Burgess, M. (2017). Hackers took more than 10,000 darkweb sites offline. *Wired Magazine*. Retrieved from http://www.wired.co.uk/article/freedom-hosting-ii-hackdarkweb-offline

Christin, N. (2012). Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace (Working Paper). Retrieved from http://arxiv.org/abs/1207.7139

Conrad, B., & Shirazi, F. (2014). A survey on tor and I2P. Proceedings from ICIMP 2014: The Ninth International Conference on Internet Monitoring and Protection.

Cox, J. (2015). 'Silk road reloaded' just launched on a network more secret than Tor. Retrieved from: http://motherboard.vice.com/read/silk-road-reloaded-i2p

Davies, G., Bouchard, M., Wu, E., Joffres, K., & Frank, R. (2015). Terrorist and extremist organizations' use of the Internet for recruitment. *Social Networks, Terrorism and Counter-terrorism: Radical and Connected*, *105*.

Deepdotweb (2017). Dark net markets comparison chart. Retrieved from https://www.deepdotweb.com/dark-net-market-comparison-chart/

Diffie, W., & Hellman, M. (1976). New directions in cryptography. *IEEE transactions on Information Theory*, *22*(6), 644-654.

Dolliver, D. S. (2015). Evaluating drug trafficking on the Tor Network: Silk Road 2, the sequel. *International Journal of Drug Policy*, *26*(11), 1113-1123.

Dredge, S. (2013, November 5). What is tor? A beginner's guide to the privacy tool. *The Guardian*. Retrieved from http://www.theguardian.com/technology/2013/nov/05/tor-beginners-guide-nsabrowser

Ellis, J. H. (1970). The possibility of non-secret encryption. In *British CommunicationsElectronics Security Group (CESG)*.

Esrock, S. L., & Leichty, G. B. (2000). Organization of corporate web pages: Publics and functions. *Public Relations Review*, *26*(3), 327-344.

Frank, R., Bouchard, M., Davies, G., & Mei, J. (2015). Spreading the Message Digitally:

A Look into Extremist Organizations' Use of the Internet. In *Cybercrime Risks and Responses* (pp. 130-145). Palgrave Macmillan UK.

Frank, R., Westlake, B., & Bouchard, M. (2010). The Structure and Content of Online Child Exploitation Networks. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (KDD).*

Garson, D. (2013). Cox regression. Statistical Associates Blue Book Series. G. David Garson and Statistical Associates Publishing, NC, USA.

GoDaddy (2017). Windows dedicated server hosting. Retrieved from https://ca.godaddy.com/hosting/windows-dedicated-server

Guitton, C. (2013). A review of the available content on tor hidden services: The case against further development. *Computers in Human Behavior, 29*(6), 2805-2815. doi:10.1016/j.chb.2013.07.031

Hardesty, L. (2015, July 28). Shoring up tor. *MIT News*. Retrieved from https://news.mit.edu/2015/tor-vulnerability-0729

Holt, T. J. (2007). Subcultural evolution? Examining the influence of on and offline experiences on deviant subcultures. *Deviant Behavior*, *28*, 171-198.

Holt, T. J., Smirnova, O., Chua, Y. T., & Copes, H. (2015). Examining the risk reduction strategies of actors in online criminal markets. *Global Crime*, *16*(2), 81-103.

Holt, T. J., Smirnova, O., & Hutchings, A. (2016). Examining signals of trust in criminal markets online. *Journal of Cybersecurity*, *2*(2), 137-145.

Huizingh, E. K. (2000). The content and design of web sites: an empirical study. *Information & Management*, *37*(3), 123-134.

IBM Corp. (2014). IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.

Immigration and Customs Enforcement. (2015). Illegal websites seized in global operation. Retrieved from https://www.ice.gov/news/releases/illegal-websitesseized-global-operation

Internet Telecommunications Union. (2016). Global ICT developments [Excel File]. Retrieved from https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2016/Stat_page_all_charts_2016.xls

Jarvenpaa, S. L., Tractinsky, N., & Saarinen, L. (1999). Consumer trust in an internet store: a cross-cultural validation. *Journal of Computer-Mediated Communication*, *5*(2).

Joffres, K., Bouchard, M., Frank, R. & Westlake, B. (2011) Strategies to Disrupt Online Child Pornography Networks. In *Proceedings of the 2011 European Intelligence and Security Informatics Conference (EISIC).*

Laub, J. H., & Sampson, R. J. (2001). Understanding desistance from crime. *Crime and justice*, *28*, 1-69.

Lu, S. (2015, Aug 19). What is the darkweb and who uses it? *The Globe and Mail.* Retrieved from http://www.theglobeandmail.com/technology/tech-news/what-isthe-darkweb-and-who-uses-it/article26026082/

Macdonald, M., & Frank, R. (2017). Shuffle up and deal: An application of capturerecapture methods to estimate the size of stolen data markets. Manuscript submitted for publication.

Mansfield, Devine, S. (2009). Darknets. *Computer Fraud & Security, 12* 4-6. doi.org/10.1016/S1361-3723(09)70150-2

Mansfield-Devine, S. (2014). Tor under attack. *Computer Fraud & Security, 8,* 15-18. doi.org/10.1016/S1361-3723(14)70523-8

McCoy, D., Bauer, K., Grunwald, D., Kohno, T., & Sicker, D. (2008). Shining light in dark places: Understanding the tor network. In N. Borisov & I. Goldberg (Eds.). *Privacy enhancing technologies.* Berlin Heidelberg: Springer.

Mei, J., & Frank, R. (2015, August). Sentiment crawling: Extremist content collection through a sentiment analysis guided web-crawler. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on* (pp. 1024-1027). IEEE.

Monk, B., Allsup, R. & Frank, R. (2015). LECENing places to hide: Geo-mapping child exploitation material. IEEE ISI 2015 Conference, Baltimore MD. Nominated for best paper.

Moore, D., & Rid, T. (2016). Cryptopolitik and the Darknet. *Survival*, *58*(1), 7-38.

Morselli, C., & Tremblay, P. (2004). Criminal achievement, offender networks and the benefits of low self-control. *Criminology*, *42*(3), 773-804.

Misata, K. (2013). The tor project: An inside view. *XRDS: Crossroads, 20*(1), 45-47. doi:10.1145/2510125

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

Newman, L.H. (2016). What we know about Friday's east coast internet outage. *Wired Magazine.* Retrieved from: https://www.wired.com/2016/10/internet-outage-ddosdns-dyn/

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Parliamentary Office of Science and Technology. (2015). The darknet and online anonymity. *POSTnote, 488,* 1-4.

Rackspace. (2017). Fully managed dedicated server + firewall configurations. Retrieved from https://www.rackspace.com/dedicated-servers.

Robbins, S. S., & Stylianou, A. C. (2003). Global corporate web sites: an empirical investigation of content and design. *Information & Management*, *40*(3), 205-212.

Sampson, R. J., & Laub, J. H. (1992). Crime and deviance in the life course. *Annual Review of Sociology*, *18*(1), 63-84.

StataCorp. 2015. *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.

Soska, K., & Christin, N. (2015, August). Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. In *USENIX Security*, *15*.

Tcherni, M., Davies, A., Lopes, G., & Lizotte, A. (2016). The dark figure of online property crime: is cyberspace hiding a crime wave?. *Justice Quarterly*, *33*(5), 890-911.

Templeton, G. (2014, November 8). Dark market massacre: FBI shuts down Silk Road 2.0 and dozens more tor websites. *Extreme Tech,* 1-2. Retrieved from http://www.extremetech.com/extreme/193821-dark-market-massacre-fbi-shutsdown-silk-road-2-0-and-400-other-tor-websites

The Tor Project. (n.d.) *About tor.* Retrieved from https://www.torproject.org

Warr, M. (1998). Life-course transitions and desistance from crime*Criminology*, *36*(2), 183-216.

Westlake, B., & Bouchard, M. (2016). Criminal Careers in Cyberspace: Examining Website Failure within Child Exploitation Networks. *Justice Quarterly*, *33*(7), 1154-1181.

Westlake, B. G., & Bouchard, M. (2016). Liking and hyperlinking: Community detection in online child sexual exploitation networks. *Social Science Research*, *59*, 23-36.

Westlake, B., Bouchard, M., & Frank, R. (2012). Comparing Methods for Detecting Child Exploitation Content Online. In *Proceedings of the 2012 European Intelligence and Security Informatics Conference (EISIC).*

Zimmermann, P. R. (1995). *The official PGP user's guide*. MIT press.

Zulkarnine, A., Frank, R., Monk, B., Mitchell, J., and Davies, G. (Forthcoming, 2016). Surfacing Collaborated Networks in Darkweb to Find Illicit and Criminal Content. In *Proceedings of the 2016 IEEE International Conference on Intelligence and Security Informatics (ISI)*.

Appendix A.

Crime Severity Weights

Name	Canada CSI Weight - 2013	BC CSI Weight - 2006-2014
Murder 1st and 2nd Degree	7554.94	7041.75
Manslaughter	1781.68	1821.56
Importation and Exportation: Heroin	1760.56	92.86
Attempted Murder	1733.14	1047.22
Kidnapping (effective 2010-01-08)	1,410.30	477.42
Incest (effective 2008-04-01)	881.4	678.35
Anal Intercourse (effective 2008-04-01)	718.6	210.98
Sexual Assault, level 2	678	210.98
Robbery	523.33	583.32
Sexual Exploitation (effective 2008-04-01)	486.3	210.98
Trafficking in Persons (effective 2005-11-01)	423.4	1278.01
Invitation To Sexual Touching (effective 2008-04-01)	380.8	210.98
Sexual Exploitation of a Person with a Disability (effective 2008-05-01)	376.5	210.98
Luring a Child via a Computer (effective 2008-04-01)	368.7	404.88
Forcible Confinement (effective 2010-01-08)	356.2	70.36