The Criminal Career Evolution of Child Exploitation Websites: Identification, Survival, and Community.

by

Bryce Garreth Westlake

M.A (Criminology), Simon Fraser University, 2011
B.A. (Psychology/Sociology), University of British Columbia, 2007

Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Criminology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Summer 2015

Approval

Name:	**Bryce Garreth Westlake**
Degree:	**Doctor of Philosophy (Criminology)**
Title:	***The Criminal Career Evolution of Child Exploitation Websites: Identification, Survival, and Community.***
Examining Committee:	**Chair:** William Glackman Associate Professor

Martin Bouchard, Ph.D.
Senior Supervisor
Associate Professor _____

Eric Beauregard, Ph.D.
Supervisor
Professor _____

Martin Andresen, Ph.D.
Supervisor
Associate Professor _____

Deborah Connolly
Internal Examiner
Associate Professor
Department of Psychology _____

Michael Seto
External Examiner
Associate Professor
Department of Psychiatry
University of Toronto _____

Date Defended/Approved: August 4th, 2015

Abstract

The distribution of child sexual exploitation (CE) material has been transformed by the emergence of the Internet. Efforts to combat distribution have been hindered by the prevalence and graphic nature of the material. One way to aid combating is to use automated data collection techniques to scan websites for CE-related criteria. Another is to contribute to proactive combat strategies by developing a theoretical framework to explain the evolution of CE distribution. Within this dissertation I develop a custom-designed webcrawler to collect data on hyperlinked networks containing CE websites and compare them to non-CE website networks. I then begin to develop a theoretical framework based on the criminal career paradigm and social network analysis to explain the evolution of website entities. Through the first study, I assess the effectiveness of a police CE-images database and 82 CE-related keywords at distinguishing websites within 10 CE-based networks from 10 sexuality and 10 sports networks. In the second study, I use a repeated measures design to compare baseline survival rates across the 30 collected networks. I then conduct Cox regression models, using criminal career dimensions adapted to website characteristics, to predict failure in CE-based networks. In the third study, I use the faction analysis to explore the formation of communities within CE-seeded networks and the characteristics that bind those communities. Results show that a) automated data collection tools can be effective, provided that the appropriate inclusion criteria is selected; b) a modified criminal career framework can be applied to CE websites, and their surrounding networks, to explain their evolution; c) individual-based criminal career dimensions can be transitioned to entity-based offenders (websites); d) websites within CE-seeded networks differ from non-CE-seeded networks in composition, survival, and network structure. The findings in this dissertation have implications for law enforcement strategies, private data-hosting services, CE researchers, and criminologists. Future research will refine inclusion criteria, expand to the Deep Web, and continue to develop an online criminal career framework.

Keywords: Child sexual exploitation; criminal career; social network analysis; webcrawler; cybercrime; child pornography

To my blood and extended family who supported and believed in me, pushing me to succeed even when I doubted myself.

To Brianne, Bubba, Caron, Cheryl, Cyd, Graham, little one, RougeOgre, Schatzi, Shane, and Shannon. I would not be in the place I am today without your love and hard work in making me the person I am today.

Acknowledgements

The completion of this dissertation and degree is the culmination of many years of hard work by myself, but especially by those in my life. Without the support of many individuals over the last 20 years, I would not be where I am today. First, I want to thank my senior supervisor Dr. Martin Bouchard for his support and guidance throughout my graduate career. From the first time I met him, in Pro-Sem, until today Martin has continually provided me with the knowledge and opportunities necessary for a successful academic career.

Second, I would like to thank all of the professors who saw the potential in me and gave me the opportunity to reach that potential by providing opportunities for growth. Specifically, my undergraduate supervisor Dr. Del Paulhus, who gave me plenty of opportunities to conduct research and publish, and Dr. Becki Ross, who gave me my first opportunity to design and teach a course on sexual violence. I also want to thank Dr. Eric Beauregard for his professional and academic support throughout my graduate career, and Dr. Martin Andresen for his insight and assistance with classes, publications, references, and career plans during my doctoral degree. Finally, I want to thank Drs. Michael Seto and Deborah Connolly for agreeing to serve on my committee. I had originally applied to the University of Toronto, for my Masters, in hopes of working with Dr. Seto. It is nice to come full circle and have Dr. Seto as part of my Ph.D. committee.

Third, I want to thank Dr. Richard Frank (webcrawler designer) and Ashleigh Girodat (research assistant). Dr. Frank and I were colleagues during the first year of my Masters and without his interest in collaboration and background in computing science, this entire project would not have been possible. To Ashleigh Girodat, thank you for putting up with me for several years! You helped me grow as a supervisor. Your hard work and effort can be found throughout this dissertation and the corresponding publications.

Finally, I would like to thank my family for their love and support throughout my life. To my mother (Cheryl) and sister (Brianne) who always stood by me and believed in me, and to my Aunt (Caron) for the opportunities she provided to me when I was young that allowed me to grow and challenge myself.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

The proliferation of material depicting the sexual exploitation of children has long been a part of society. Historically, those who produced and disseminated material operated in small, covert, networks (Beech, Elliott, Birgden, & Findlater, 2008; Tremblay, 2006). Material produced, such as images, was typically expensive and of poor quality (Wortley & Smallbone, 2012). The rise of an accessible global marketplace, via the Internet, and the technological advancements of image and video capturing devices, transformed the crime of child sexual exploitation (CE) from a local issue to an international epidemic. By 2000, 77 percent of CE cases were Internet related (Hughes, 2002).

In addition to the World Wide Web (WWW) transitioning components of CE to cyberspace, it modified the processes involved in the crime. The WWW afforded like-minded offenders the opportunity to connect across long distances, changing a historically solitary crime into a global social network (Quayle & Taylor, 2011; Taylor & Quayle, 2003; Tremblay, 2006). Offenders also received support for their interests while limiting their risk of apprehension, due to the anonymity provided (O'Halloran & Quayle, 2010). The number of ways in which material could be acquired also increased (Callanan, Gercke, DeMarco, & Dries-Ziekenheiner, 2009; Faber, Mostert, Faber, & Vrolijk, 2010; van Wijk, Nieuwenhuis, & Smeltink, 2009). Finally, CE material became a commodity for exchange rather than one for sale (Beech, Elliott, Birgden, & Findlater, 2008; Estes, 2001), aiding in the facilitation of free flowing material in large social networks.

Growth in the online distribution of CE material has resulted in an increased number of researchers examining the phenomenon. This research has examined international law (Akdeniz, 2013; Gillespie, 2011), offender characteristics (Babchishin, Hanson, & Van Zuylen, 2015; Elliott, Beech, & Mandeville-Norden, 2013; Webb,

Craissati, & Keen, 2005; Wolak, Finkelhor, & Mitchell, 2005), typologies (Alexy, Burgess & Baker, 2005; Krone, 2004; Young, 2005), risk of victimization (Lewandowski, 2003; Mitchell, Finkelhor, & Wolak, 2003; Mitchell, Wolak, & Finkelhor, 2008; Palfrey, Boyd, & Sacco, 2010), detection tools (de la Cruz, Aller, Garcia, & Gallardo, 2010; Frank, Westlake, & Bouchard, 2011; Peersman, Schulze, Rashid, Brennan, & Fischer, 2014), strategies (Hurley et al., 2013; Joffres, Bouchard, Frank, & Westlake, 2011; Pandit, Kulkami, & Dhore, 2009; Westlake, Bouchard, & Frank, 2011; Wolak, Finkelhor, & Mitchell, 2012), and prevalence (Wolak, Liberatore, & Levine, 2014). Research into distribution has focused on quantifying the amount of material available (Steel, 2009), the type of material distributed (Taylor, Holland, & Quayle, 2001), as well as where and how it is distributed (Carr, 2004; Latapy, Magnien, & Fournier, 2013; LeGrand, Guillaume, Latapy, & Magnien, 2009). This dissertation adds to the existing research and builds on the current literature in three important ways. First, this dissertation focuses on distribution via the WWW, through publicly accessible websites, rather than peer-2-peer networks such as Internet Relay Chat, Gnutella, and eMule. Second, it examines CE distribution at the entity (i.e., website) level rather than the individual level. Third, it addresses the transition of CE from a solo crime to a communal crime, by incorporating social network analyses. In addition to the literature on CE, this dissertation begins to develop a theoretical foundation for applying the *criminal career paradigm* (Blumstein, Cohen, Roth, & Visher, 1986) to cyberspace and validates the use of automated data collection techniques for cybercrime research.

In addition to the prevalence of the Internet within society, its functional structure may require an adjustment to the way we explain crime, given the inherent advantages of cyberspace for committing crime. The Internet is decentralized, adhering to no international boundaries or laws. This creates impediments for the regulation and monitoring of activities online. The lack of international law for combating cybercrime has been evidenced recently through the ongoing attempts to shutdown piracy giant The Pirate Bay (Gibbs, 2014b). Although law enforcement agencies have been able to cease the operations of the website for short periods of time, the website quickly resurfaces, hosted from a new location where the laws necessary for copyright law enforcement are lacking (Mazumdar, 2015).

The Internet provides plenty of advantages for those conducting criminal activities in cyberspace. It is a highly efficient system that facilitates quicker access to information, material, and social support. This efficiency also provides immediate access to a large, global, number of potential victims, significantly reducing execution time and risk of apprehension. The benefits of a global victim-pool are enhanced by economic differences between countries, resulting in the most vulnerable having the poorest security and being the easiest to exploit. Most importantly, the Internet affords offenders the ability to operate semi-anonymously. Research has shown that the lack of concern with being identified has resulted in many offenders sharing sensitive information, detection avoidance techniques, and tools/skills, in semi-public to public forums (Armstrong & Forde, 2003). The low risk of apprehension is not confined to third-world countries (IWF, 2014). Due to privacy concerns in North America and Europe, most Internet Service Providers and website hosting companies do not monitor the activities of users or websites they host (Wortley, 2012). Even popular blog-hosting services (e.g., Blogger©) are unable to effectively monitor the use of their products without public fears of privacy intrusion (Gibbs, 2014a). In situations where a website (e.g., blog) is found to be conducting illegal activities, and all the proper legal components have been followed, the website's removal is little more than an inconvenience to offenders, as they can quickly restore a backup, under a new address/domain, with a new username/email. While the Internet may provide multiple advantages for offenders, these same advantages provide opportunities for researchers to investigate criminal activity in manners unavailable offline. As a result, the Internet is a great research tool and an important avenue for understanding important theoretical constructs.

Offline criminal career research often suffers from difficulties in accessing complete and accurate offender data (Piquero, Farrington, & Blumstein, 2003). To properly assess the criminal career, research must include those who are detected, detected and released, and never detected (Nagin, 2008). Unfortunately, most research into the criminal process relies on post-crime arrest data. This data inherently ignores offenders who have avoided detection but remain active or offend for a period of time, without detection, and then desist. As a result, criminologists understand why offenders fail, but rarely do we understand the attributes that lead to success. The public, semi-anonymous, nature of the Internet means that illicit activities often remain in the public

domain; at least initially (Tremblay, 2006). The perceived low risk of apprehension provides a unique opportunity for criminologists to observe and conceptualize the criminal process as it occurs, rather than retroactively, and determine what facilitates success. This is especially true for public websites involved in illegal activities. The ability to monitor activity, in real-time, and collect *all* of a website's data provides an opportunity to address some of the limitations of previous criminal career research. However, questions still remain as to whether the data collected via the Internet is valid, representative, and complete. Within this dissertation, I will address these concerns and provide evidence that Internet-mediated research methods are valid, representative, and complete, and can be beneficial for understanding the criminal career, and criminal processes, for both online and offline criminal phenomena.

This manuscript-based dissertation draws from a longitudinal research design to allow for the measurement of websites involved in the distribution of CE material, at multiple points in time, via snowball sampling methods. There are four objectives that will be the focus. First, I will develop an automated webcrawler tool that integrates a law enforcement database of known CE images and keywords to identify websites involved in the proliferation of CE material. Second, I will develop a unique conceptual framework for analyzing the evolution of CE websites, drawing on the criminal career paradigm and social network analysis. Third, I will develop an analytical framework to study the evolution of CE websites and their corresponding networks, over time. This will include a) monitoring website failure; b) examining the social structure of communities; c) measuring changes in material present; and d) investigating the determinants of website popularity. Fourth, I will contrast the characteristics of websites within CE-seeded networks with two comparison genres of networks: sports and sexuality. This will include how they differ in website structure, presence of material (images, videos, and keywords), and rates of failure.

To address the objectives of this dissertation, I will present three studies. The first study will involve assessing a) the use of automated data collection tools and b) the criteria selected to identify CE-related data. The second study will explore website survival and how criminal persistence is facilitated or inhibited by offender (i.e., website) characteristics, corresponding to the dimensions of the criminal career paradigm. The

third study will examine social exchange between websites, using the community identification method faction analysis. This will provide information on the criminal processes that lead up to co-offender selection and how communities form and distribute material.

1.1. Criminal Career Paradigm

Coined by Blumstein, Cohen, and Nagin (1978), the criminal career paradigm encompasses when a person beings offending (onset), how they continue to offend (persistence), the change in offending patterns (escalation), and why a person ceases offending (desistance). Central to the active criminal career are the four dimensions *offending frequency, duration, crime-type mix,* and *co-offending.*

Offending frequency is defined as the change in frequency over a period of time and is often described in relation to chronic offenders (Blumstein & Cohen, 1979). Duration is the interval between when an offender initiates their career and when their career terminates. While the factors contributing to the initiation of offending have garnered extensive attention—with evidence that early onset offending increases later offending (Elliott, 1994; Loeber & Farrington, 1998)—research into the factors linked to criminal career termination have been impeded by difficulties in ascertaining the *true* desistance of the criminal career (LeBlanc & Loeber, 1998). Crime-type mix relates to the degree of specialization in offenders, the patterns of seriousness in offending, and the process of seriousness escalation over time. Finally, while some crimes (e.g., burglary and robbery) are more likely to involve co-offending, and juveniles engage in co-offending behaviour more than adult offenders, it has been argued (Reiss, 1988) that at some point within the criminal career, each offender engages in co-offending practices. The fourth dimension, co-offending, relates to the processes involved in co-offending practices and how they impact criminal career trajectory.

The criminal career paradigm has had an important influence on both theoretical and public practices. Many life-course and developmental theories incorporate the concepts of the criminal career, including Sampson and Laub's (1993; 1997) *age-graded* theory, Patterson and Yoerger's (1999) *social-interactional* developmental model,

Moffitt's (1993) *developmental taxonomy,* and Loeber and colleagues (1993) *developmental pathways* model. The criminal career paradigm has also informed general policy strategies (e.g., intervention programs and three strikes), sentencing practices and appropriate durations, and the identification of chronic career criminals. The growing priority amongst policy-makers to address the prevalence of cybercrime, highlights the importance of transitioning the criminal career paradigm to cyberspace, to facilitate the application of other theories to cybercrime and to aid in formulating policy strategies to account for the change to a virtual environment.

Piquero, Farrington, and Blumstein (2003) point to four general methodological issues that impede criminal career research: data, research design, analytic techniques, and analytic issues. First, the reliance on self-reports and official records impact reliability, representative sampling, and completeness, while difficulties in accounting for 'street' time (i.e., opportunities to offend) underestimates offending frequency. Second, the reliance on cross-sectional research designs to study a longitudinal phenomenon. Third, identifying the best analytic techniques for assessing change over time and incorporating random effects and clustering. Fourth, parsing state-dependence and persistent-heterogeneity effects and operationalizing, defining, and measuring career desistance. In this dissertation I address these issues by a) assessing the tool used to collect the data; b) using a longitudinal research design; c) addressing issues of dependent observations, unobserved heterogeneity, and ways to assess stability and change over time; and d) carefully defining, operationalizing, and measuring the transition, and manifestation, of criminal career dimensions to cyberspace and for online entities (e.g., websites).

Despite the abundance of research being conducted on the criminal career, DeLisi and Piquero (2011) outlined 16 existing knowledge gaps. Within this dissertation, I will specifically contribute to addressing three. First, co-offending practices have been shown to impact criminal career dimensions offending frequency (McCord & Conway, 2002), duration (Andresen & Felson, 2010; Carrington, 2009; D'Alessio & Stolzenberg, 2008; Piquero, Brame, & Lynam, 2004), and crime-type mix (Andresen & Felson, 2012; McGloin & Piquero, 2009). However, DeLisi and Piquero noted that the lack of social network data, on co-offending processes, results in the correlation between peer

influence and delinquency (e.g., Bouchard & Spindler, 2010) being unclear. To address this, I use social network analyses to examine the formation of communities around CE websites, focusing on the characteristics that bind hyperlinked partnerships and whether partnership characteristics are homogeneous or heterogeneous. Second, prior research has concluded that the majority of offenders are generalists, resulting in DeLisi and Piquero questioning the utility of typologies. However, amongst child molesters, there is modest support for specialization (Harris, Smallbone, Dennison, & Knight, 2009; Lussier, LeBlanc, & Proulx, 2005). For online child molestation, specialization is unclear (Babchishin, Hanson, & Hermann, 2011; Elliott, Beech, & Mandeville-Norden, 2013; Mitchell, Wolak, Finkelhor, & Jones, 2011; Wolak, Finkelhor, & Mitchell, 2005). To address the issue of specialization, I will examine the crime-type mix of CE websites and, less directly, determine if typologies can be garnered from the patterns found. Third, criminal achievement is an important factor in the duration of the criminal career (Lussier, Bouchard, & Beauregard, 2011; Morselli & Tremblay, 2004). However, central to achievement is an effective assessment of risk and reward. Online, the notion of risk is different, given increased anonymity, resulting in offenders operating with a reduced fear of detection (Armstrong & Forde, 2003; Decary-Hetu & Dupont, 2013; Decary-Hetu, Morselli, & Leman-Langlois, 2012; Maimon, Alper, Sobesto, & Cukier, 2014; Wolak, Finkelhor, & Mitchell, 2005). To address DeLisi and Piquero's call for research into situational context and decision-making, I examine persistence and how website characteristics increase (or decrease) the risk of premature failure. In addition to addressing three knowledge gaps outlined by DeLisi and Piquero, I will address two additional gaps. First, I will transition the criminal career paradigm to cyberspace and cybercrime. Second, I will develop the concept of the 'entity' criminal career by examining the evolution of websites directly and indirectly involved in the distribution of CE material.

The criminal career paradigm has been characterized as "the longitudinal sequence of crimes committed by an individual offender" (Blumstein, Cohen, Roth, & Visher, 1986, p.12). However, it would be naïve to assume that the actions of an offender were not influenced by a criminal enterprise, if they were associated (Tremblay, Laisne, Cordeau, Shewshuck, & MacLean, 1989). While criminal organizations, such as gangs, are continually moving towards decentralization and the formation of smaller,

close-knit, sub-groups (Bouchard & Morselli, 2014; Bouchard, & Spindler, 2010; Morselli, 2009), these sub-groups often still fulfil certain niches or roles within the larger organization. As such, their actions, and career trajectory, are influenced by the criminal enterprise. In this model, the goals and career of the criminal enterprise are actually the focal point and, more importantly, are not dependent on any specific member within the organization. Therefore, the criminal enterprise can be seen as having its own career, independent of any of the members. Understanding that criminal activities are conducted at both the micro, individual, level and the macro, entity, level, research has begun to examine the co-offending practices of criminal enterprises (Malm, Bichler, & Nash, 2011; Tenti & Morselli, 2014; Tremblay, et al., 1989; Tremblay, Bouchard, & Petit, 2009) and terrorist groups (Amirault & Bouchard, 2015).

The collective (entity) criminal career (Tremblay, et al., 1989) has yet to be developed as a concept directly. Rather, it has been developed empirically through research examining co-offending practices between groups or entities. Globalization has aided in the formation of a worldwide marketplace where criminal organizations form business partnerships with other criminal organizations. These entity-based partnerships center on resource sharing, through the exchange of goods and services (Albini, 1971, Haller, 1990) and intelligence (Goodson & Olson, 1995). Although entity-based partnerships have been shown to occur at the local (Potter, 1994) and national level (Canadian Centre for Justice Statistics, 1999), the majority of research has focused on the organizational features and structures of international partnerships. Within the drug-trade, Williams (2001) found that Colombian drug traffickers formed alliances with Sicilian, and other, criminal entities, to enter the European drug market, while Turkish and Moroccan entities controlled the Dutch heroin market by hiring Dutch criminal groups to transport their product into the Netherlands (Bruinsma & Bernasco, 2004; Bovenkerk, Siegel, & Zaitch, 2003). Finckenauer and Waring (1998) found that while not directly connected to a large number of organizations, Russian entities were indirectly connected to many others, allowing for opportunity-driven associations to exploit specific skill sets of other organizations. O'Neill (2010) found that these temporary partnerships aided Russian entities trafficking women in Western Europe. Malm, Bichler, and Nash (2011) found that entity co-offending was most common between ethnic groups; although not uniform across groups. They also found structural differences between co-

offending networks, with Asian entities having higher brokerage scores, but lower density scores than other ethnic entities. Tenti and Morselli (2014) found that entity co-offending networks were loosely based, with partnerships materializing in the pursuit of similar criminal goals. They also found that networks were 'chain-like', with low-density scores, but consisted of smaller, more densely connected, subsets of entities. Although entities within these subsets were not directly connected to entities from other subsets, through their associations with others, they were indirectly connected to most.

Although it is not labeled as such, nowhere may the concept of the *entity* criminal career be so apparent than in cyberspace. The past two years have been highlighted by the apprehension, and conviction, of Ross William Ulbricht, founder of the black market website *Silk Road*, and a $110 million settlement between the torrent-based, piracy, website *isoHunt*'s founder Gary Fung and the Motion Picture Association of America. In both cases, replacement websites (Silk Road 2.0 and isoHunt.to) were launched shortly after the original website was removed; operating with the same goals and objectives as their predecessors (Dolliver, 2015). Established in September 2003, the BitTorrent website *The Pirate Bay* has remained active despite numerous attempts by law enforcement agencies to remove the website *and* the incarceration of its three founding members Gottfrid Svartholm, Fredrik Neij, and Peter Sunde. Like *Silk Road* and *isoHunt*, *The Pirate Bay* has outlived its original creator(s), continuing to independently grow and evolve. As a result, each of these websites can be provided as examples of how an entity can formulate its own criminal career that, while influenced by those that comprise the entity, functions outside the criminal career of any specific person(s).

Compared to the structure and organization of offline entities, I expect that online entities (websites) will mirror some characteristics while the nature of the Internet will result in some differences between offline and online criminal entities. Like offline entity-based partnerships, websites should form hyperlinks based on resource sharing. This includes the exchange of illegal goods and services, and intelligence. Online entities will also form sub-groups based on shared characteristics such as the type of material being distributed or the medium being used. However, the nature of the Internet will result in more stable partnerships—as the maintenance cost of a hyperlink is minimal—and higher overall and sub-group density scores. Therefore, like offline entities, websites will

be indirectly connected to most websites, within the larger network, however they will be directly connected to others more than between offline entities.

1.2. Social Network Analysis

Social science research has traditionally explained the behaviours of an individual through their personal characteristics; whether those be primarily psychological or sociological. However, as the debate between peer influence and selection (Haynie, 2001; 2002) shows, there is acknowledgement that the subsequent behaviour of an individual may be influenced as much by their surrounding environmental structure as it is their personal characteristics. Social network analysis involves examining the linkages and interdependence of social interactions between different units (e.g., people) in a bound network to explain behaviour (Wasserman & Faust, 1994; Wellman, 1983). That is, if two groups of equally skilled people are compared, their overall performance will be dependent on the relationships between group members (Borgatti, Mehra, Brass, & Labianca, 2009).

The notion that structure mattered was first reinforced, in social science research, by Travers and Milgram (1969), who tested the hypothesis of small worlds that stated every American was connected within 'six degrees of separation'. After, important advancements to our understanding of social networks were made by Granovetter (1973), who surmised that those with common knowledge clustered but the people who bridged between clusters were the 'strongest'; Freeman (1979), who analyzed the importance, and devised three (centrality) measures, of network position; Burt (1992), who argued the importance of structural holes in social structures and that competition was dependent on relations rather than personal attributes; and Barabasi and Albert (1999), studied the World Wide Web and found that large networks follow a scale-free power-law distribution. Together, these studies bridged the gap between a mathematical equation, graph theory, and human interaction and behaviour.

Although derived from graph theory, SNA is not seen as a theory itself. Rather, SNA is a theoretical and methodological paradigm that can be used to analyze the relationships between nodes (e.g., people, organizations, and websites) and

connections (Borgatti & Halgin, 2011; Papachristos, 2011). These relationships can be positive or negative and can influence the behaviour of the individual nodes and/or the surrounding network. Borgatti, Mehra, Brass, and Labianca (2009) summarizes relationships (i.e., ties) into four types: 1) similarities (e.g., location, membership, or attribute), 2) social relationships (e.g., kinship, affective, or cognitive), 3) personal interactions, and 4) flows (e.g., information, beliefs, resources, or personnel). As a theoretical paradigm, SNA is characterized by three theoretical concepts (Borgatti, Mehra, Brass, & Labianca, 2009): 1) the shape and cohesion of the network structure, measured through network density, the clustering coefficient, and fragmentation; 2) the position of nodes within the network, measured through centrality, degree, closeness, and betweenness; and 3) the dyadic cohesion and equivalence, referring to the social closeness (dyadic cohesion) and the similarity in network roles (equivalence) of any pair of nodes. In this dissertation, I will examine two types of relationships: similarities (through location and attributes) and flows (through information and resources). I will also examine two theoretical concepts of SNA: the influence of structure and position of websites, through density and clustering, and the role of dyadic cohesion, through homophily.

Though previously underused in criminology, despite being considered a crucial paradigm in understanding criminal behaviour and advancing the field (Morselli, 2009), the use of SNA has begun to increase. This is because SNA allows criminologists to observe and measure the importance of social structure in criminal behaviour (Papachristos, 2011). Most recently, SNA has been most influential in the study of terrorism (e.g., Basu, 2014; Bouchard and Nash, 2014; Perliger & Pedahzur, 2011; Ressler, 2006), gangs (Bouchard & Konarski, 2014; Papachristos, 2009; Pyrooz, Sweeten, & Piquero, 2013; Tita & Radil, 2011) and drug markets (Kenney, 2007; Malm & Bichler, 2011; Morselli, 2010). The importance of SNA has also been acknowledged in the study of online CE, most notably by Krone (2004), stating that the linkages between distributors may be as important to understanding the phenomenon as the material that is being distributed. But beyond a few exceptions, the SNA studies taking advantage of the networked aspect of online CE are still rare.

The nature of online interactions and digital information make the Internet an ideal location for applying SNA (Hogan, 2008). Online interactions are inherently social network-based. Between people, this is conducted through direct communications between a sender and a receiver. Between websites, social networks are created through hyperlinked webpages. The nature of digital information also makes it conducive to SNA as network data are not dependent on in-person data collection, perception, or memory. Rather, there is concrete evidence of connectivity frequency and importance that is easy to measure. This means that the presence of a relationship (tie) and the strength of that relationship can be quantified through straightforward methods.

Hyperlink Network Analysis (HNA) is useful for studying the underlying structure of computer-mediated communication (Jackson, 1997; Park & Thelwall, 2003). Compared to social and Internet networks, that Park (2003) describes as focusing more on individuals, hyperlink networks can focus on people or entities and are "an extension of traditional communication networks in that it focuses on the structure of a social system based on the shared hyperlinks among websites" (p.51). Martinez-Torres and colleagues (2011) outline two different types of website hyperlink networks. The first is page networks, where the links between webpages of a given website are analyzed. The second is domain networks, where the links between websites are analyzed. In this dissertation, I focus on domain (i.e., website) networks and use many of the SNA features described by Martinez-Torres and colleagues, including data reduction techniques such as factor analysis. Prominent in cyberspace, and the key format in which websites express communication, Thelwall (2006) details five concerns with attributing meaning to hyperlinks between nodes (i.e., websites). First, the presence of hyperlinks between two websites does not identify the type of relationship as links can be used for a variety of purposes. Second, the sample size required for practical findings may not be worth the effort necessary for data collection. Third, in most hyperlink research the connections are not the primary focus but rather a means for drawing conclusions about the underlying processes of the network. This means that statistical analyses that assume independence of observations are not valid. Fourth, the dynamic nature of the Internet means that conclusions drawn from the structure and formation of hyperlinks may change over time or not be a valid representation of the current landscape. Fifth, the use of generic descriptors for hyperlinks that fail to differentiate the

value between hyperlinks. In addressing some of Thelwall's noted concerns, within this dissertation I a) have a large sample size across multiple (30) networks; b) account for the violation of independence of observations; and c) collect data over an extended period of time (60-weeks) and place my findings in the context of previous research to determine if the dynamic nature of the Internet has impacted the findings. To avoid assuming the relationship between two websites, given a hyperlink, or incorrectly interpreting the nature of a relationship, I focus only on the presence of a connection rather than the strength or importance of that connection.

Within criminology, HNA is most prominently used in studying the structure and purpose of online communities. This research has focused on hate groups (Burris, Smith, & Strahm, 2000; Chau & Xu, 2007), extremists (Zhou, Reid, Qin, Chen, & Lai, 2005), and terrorism (Fu, Abbasi, & Chen, 2010; Yang & Ng, 2007). This research has concluded that hyperlinked criminal communities are purposive and driven by homophilious ties; often operating in isolation from other criminal and non-criminal communities. More specifically, Burris, Smith, and Strahm found that network density among 80 white supremacists websites was 11% overall, but considerably higher within white supremacy sub-ideology communities. In this dissertation I will examine the patterns found in the hyperlinked networks of child sexual exploitation websites.

1.3. Role of Websites in Cybercrime

Websites are a foundational component of the Internet and, thus, have been appropriated in three ways for criminal activity. First, non-criminal websites have been exploited by offenders. For example, cyberbullying victimization increases with the use of social networking websites (Bossler, Holt, & May, 2012; Kwan & Skoric, 2013) and conversation applications, such as instant messaging and blogs (Moore, Guntupalli, & Lee, 2010)—sometimes with deadly outcomes (Hinduja & Patchin, 2010). Child molesters have used non-criminal websites, specifically social networks websites (Mitchell, Finkelhor, Jones, & Wolak, 2010), to engage in grooming tactics (Williams, Elliott, & Beech, 2013). Finally, fraudsters have used online purchasing websites to obtain identities (Holt & Turner, 2012; Pratt, Holtfreter, & Reisig, 2010).

Second, criminal websites have been created as virtual meeting places. These websites have been used by hackers to mentor skills (Decary-Hetu & Dupont, 2012) or exchange tools (Holt, 2009; Schell & Dodger, 2002), and by terrorists to distribute attack techniques (Weimann, 2005). They have also been used by hackers (Holt, Burruss, & Bossler, 2010), terrorists (Freiburger & Crane, 2008; TAG, 2013), software pirates (Burruss, Bossler, & Holt, 2012), and sex offenders (Holt, Blevins, & Burkert, 2010; Jenkins, 2003; Quayle & Taylor, 2002; Tremblay, 2006) to reaffirm behaviour, promote activities, recruit, and provide social support.

Third, criminal websites have been created as independent enterprises for criminal exchange. Online marketplaces have been setup for offenders selling credit card and bank account numbers (Chu, Holt, & Ahn, 2010; Holt, 2013; Holt & Lampke, 2010; Motoyama, McCoy, Levchenko, Savage, & Voelker, 2011). Websites have also been used as a means for centrally locating pirated music, videos, and software (Bachmann, 2007; Higgins & Marcum, 2011; Ingram & Hinduja, 2008; Moores & Esichaikul, 2011; Nhan, 2013; Steinmetz & Tunnell, 2013). Finally, drug traffickers and weapons dealers have created online marketplaces to advertise their products (Christin, 2013; Dolliver, 2015; van Hout & Bingham, 2013). These marketplaces have even implemented complicated quality control and customer feedback systems. Criminal website marketplaces have also been used extensively amongst sex offenders to advertise services (Holt & Blevins, 2007; Milrod & Weitzer, 2012; Sharp & Earle, 2003), sex tourism (Evans, Forsyth, & Wooddell, 2000; Leheny, 1995), and trafficking (Kotrla, 2010; Zhang, 2009).

Despite the reported prevalence of websites distributing CE material (Carr, 2004; IWF, 2014), the majority of research has focused on peer-2-peer networks (e.g., Fournier, et al., 2014; Rutgaizer, Shavitt, Vertman, & Zilberman, 2012; Steel, 2009). Research conducted on CE distribution on public websites has found that CE websites are not difficult to find, that websites with verified CE images differ from websites without verified content (Westlake, Bouchard, & Frank, 2011), and that functionality and structure differ within CE website types (blogs versus non-blogs; Frank, Westlake, & Bouchard, 2010). Westlake, Bouchard, and Frank (2012) found that multiple criteria were required for detecting CE websites and that the type of websites detected were

influenced by the criteria chosen, while Joffres, Bouchard, Frank, and Westlake (2011) identified the presence of hubs within CE-based networks and strategies for disruption, targeting hubs to reduce network density and clustering while using fragmentation tactics to reduce cohesion and average path length (minimum distance to access all websites). Public websites are important for understanding the process of online CE distribution. This is because they a) function as a starting point for offenders, given their easy accessibility and minimal risk of detection; b) function at the macro level of distribution and encompass individual offenders, providing a representation of the less-accessible micro level processes; c) provide information about the evolution of CE distribution, such as the preferred medium of distribution; d) provide evidence of the actions of individual offenders over time, such as the steps taken to avoid detection or new tools being distributed; and e) identify growing trends in exploitation, such as self-exploitation (Leary, 2007, 2009).

Examining CE distribution amongst public websites also has implications for advancing our knowledge about other types of cybercrime. As outlined above, public website have been exploited by all types of cybercriminals. Therefore, the examination of CE websites, and their surrounding network, has implications for other types of cybercrime. For example, hackers use public websites to exchange information and teach new hackers (Decary-Hetu & Dupont, 2012). This information is useful to researchers and law enforcement agencies because it explains the criminal process (e.g., mentoring) and how offenders evolve in their offender practices. It also gives advance information to law enforcement of new growth areas, providing intelligence on where subsequent enforcement should be focused. As the use of public websites is evident across types of cybercrime, the knowledge gathered from research into any cybercrime can then be applied to other types of cybercrime. For example, Westlake, Bouchard, and Frank (2011) developed an adaptable formula weighting content and connectivity to identify key websites within CE distribution networks. Given the adaptability of this formula, it can be used by researchers and law enforcement for different types of cybercrime.

Hartman, Burgess, and Lanning (1984) characterized CE collectors into four behavioural types: closet (secretive), isolated (personal collection), cottage (share with

like-minded for validation), and commercial (financial gain). Research into online CE has shown that network members often occupy specific roles and purposes (Alexy, Burgess, & Baker, 2005; O'Connell, 2001; Sullivan & Beech, 2004). Individually, online sexual offenders have been categorized into types. Briggs, Simon, and Simonsen (2011) identified two subtypes of groomers (fantasy driven and contact driven), while Lanning (2001) categorized online sexual abusers into seven types and Krone (2004) nine types– on a continuum of severity. Amongst Krone's nine types were the browser, private fantasizer, trawler, non-secure, secure, groomer, physical abuser, producer, and distributor. Although these studies were conducted at the micro level, many of the typologies suggested could be transposed to the macro, website, level. For example, some websites will be secretive (e.g., closet or secure collector), or open (commercial, non-secure, or distributor). Some websites will be dedicated to the collection of a specific abuser or abusers (isolated or producer), others will be directories pointing offenders to websites with specific material (trawler), while some will be 'accidental' offenders (browser). Although I do not directly address typologies within this dissertation, like individual offenders, websites play certain roles in the larger CE distribution network. Therefore, typology research should be conducted on websites involved in cybercrime to determine their main purpose and whether that purpose modifies functionality, structure, and/or social control combat. The findings within this dissertation aid in the initial construction of CE website typologies through the website characteristics analyzed and their influence on survival (Chapter 3) and network structure (Chapter 4).

1.4. Research Contributions

The growth in the use of the Internet has resulted in many benefits within our global society (Howard & Jones, 2004; Wellman & Haythornthwaite, 2002). However, as the Internet has increased in prominence it has been increasingly used to facilitate criminal activity. One cybercrime that has changed extensively, as a result of the Internet, is the sexual exploitation of youth. This change created three types of gaps in knowledge for criminologists. First, the absence of a theoretical framework that incorporates the cyber-environment, including the transition from a solitary crime to a global crime heavily reliant on social exchange. Second, the need for innovative

interdisciplinary research designs and the creation of tools to account for the change from an offline to online environment. Third, the need for designing efficient ways to improve the existing strategies and methods for combating online child exploitation.

From this dissertation, I will make three important contributions to the advancement of knowledge. First, I will continue the process of revising existing theoretical frameworks to account for the transition to cyberspace (e.g., Bossler & Burruss, 2011; Higgins & Marcum, 2011; Patchin & Hinduja, 2011). More specifically, I will apply the dimensions of the criminal career paradigm to websites and the distribution of CE material. Second, I will incorporate social network analysis and assess the use of automated data collection tools for Internet-mediate research. This will aid researchers in all fields, by improving the techniques and criteria used to collect data on the Internet, and law enforcement and private organizations in revising existing techniques for combating online CE. Third, the longitudinal aspect of this dissertation will assist in understanding how networks evolve and adapt over time, to ensure survival and promote dissemination. From this, law enforcement agencies can determine how networks adapt to social control efforts and develop a strategy for targeting offenders and establishing priorities. Finally, the analyzing of an automated webcrawler data collection tool will have direct implications for social control agencies. First, it provides a tool that can be used and adapted by law enforcement and private organizations to scan websites looking for illegal material of any type. Second, for those investigating online CE, this automation process reduces direct contact with material thereby reducing the psychological costs associated with the work. Third, a better understanding of the novel virtual environment and the theoretical principles that underlie the social exchange process of offenders will provide a foundation for understanding the phenomenon and how best to address it.

1.4.1. Study #1: Assessing the Validity of Automated Webcrawlers as Data Collection Tools to Investigate Online Child Sexual Exploitation

The abundance of data available on the Internet coupled with efficiency and cost-effectiveness of collection has facilitated the increased reliance on automated data collection techniques within Internet-mediated research. Despite increased use, their

validity, for investigating subjects where avoiding detection is important (e.g., forms of cybercrime), is unclear. Many automated data collection tools, such as webcrawlers, rely on researcher-defined criteria to guide the collection process and identify what information to include and what to discard. Cybercrime researchers who use webcrawlers have focused their criteria on keywords related to the topic of inquiry or specific attributes, such as types of file extensions. However, in both cases, the rates of false positives and false negatives are unknown. For example, the cross-indexing of keywords amongst subgroups brings in to question the validity of using such measures to identify one specific subgroup (e.g., child molesters) and exclude another subgroup (e.g., adult pornography). As cyber-related crimes continue to grow and criminologists transition more to automated data collection, it is important that the validity of these tools is tested and recommendations on how to create the best criteria are identified.

The first study of this dissertation is focused on assessing the use of automated data collection tools and uses the topic of CE material distribution on websites as its measure. To validate the inclusion criteria, I selected two contrasting genres for comparison. The first comparison genre, sports, was selected because of its dissimilarity to CE while the second comparison genre, sexuality, was selected because of its similarity. That is, CE criteria should appear minimally in dissimilar genres and modestly in similar genres. This study uses a custom-designed webcrawler to automatically collect data on three genres of websites. I then descriptively compare the derived networks to address two main objectives. First, whether automated data collection tools are effective for investigating, sometimes hidden, criminal activities in cyberspace. Second, whether commonly used criteria can distinguish between topic-related websites and unrelated websites.

The data used for this study was collected for the purpose of this dissertation and consists of 30 networks, each with 300+ websites and approximately 500,000 webpages. Each network began from a selected seed-website that was CE-related (10 seeds), sexuality-related (10), or sports-related (10). The networks were then formulated following hyperlinks to connected websites. Although the same data were collected on comparison networks, the inclusion criteria were only required within the CE networks. The 30 networks were recrawled 10 times, every six weeks, and their evolution was

plotted. Assessment of the webcrawler includes comparing the presence of the inclusion criteria across the three genres of networks and social network analysis measures of network cohesion. The manuscript presented here is currently under review at *Sexual Abuse: A Journal for Research and Treatment*.

1.4.2. Study #2: Criminal Careers in Cyberspace: Examining Website Failure within Child Exploitation Networks

Like desistance for the individual criminal career, the 'true' end of the criminal career for a website is difficult to measure. Similar to offline offenders, a website can have periods of inactivity or desist for reasons unrelated to crime (e.g., lack of interest, service interruptions, available time, or finances). Criminally, a website can be (temporarily) dormant because of detection or efforts to avoid detection. Moreover, after detection, a website may continue its criminal career under a different, but similar name (e.g., Silk Road 2.0). The second study of this dissertation is unique in that it explores the concept of persistence as it pertains to an entity rather than an individual. For the purpose of this study, persistence is conceptualized as the duration prior to first recordable failure, at a specific uniform resource locator (URL).

While questions exist around what constitutes persistence for websites, what may be more important is which characteristics result in prolonged survival and which actions result in premature failure? As many control agencies have databases of CE images, one survival strategy for websites would be to avoid distributing known images. However, it may be difficult for a website to know which images control agencies are aware of and which they are not. In addition to specific material, there are many characteristics that a website can control and may impact survival. Within this study I will examine offending habits (frequency, seriousness, and specialization), co-offending patterns (connectivity) and general attributes (size of website, amount of content, etc.).

The second study of this dissertation, accepted for publication in *Justice Quarterly,* examines the factors associated with persistence amongst websites, with two main objectives. The first objective is to identify the baseline survival rate for websites within CE networks and compare these Kaplan-Meier survival curves to those from sexuality and sports networks. To meet this objective, all 30 networks, across all 10

waves of data collection, will be used. The second objective of this study is to determine which, if any, website characteristics contribute to persistence or early failure, amongst websites within CE networks. This will result in multiple cox regression analyses being conducted, examining general website characteristics (e.g., website size) and each dimension of the criminal career paradigm. Offending frequency will be measured through known CE images and CE code words. Crime-type mix (i.e., specialization) will be measured through three composite measures: sex of the victim (boy/girl), severity (explicit/non-explicit), and media (videos/images/stories). Finally, co-offending will be measured through outgoing (connectivity) and incoming (popularity) hyperlinks. Although the Cox regression analyses will be conducted on all 30 networks, the focus will be on the CE networks.

1.4.3. Study #3: Liking and Hyperlinking: Community Detection in Online Child Exploitation Networks

One of the four key dimensions of the criminal career is co-offending. Those whom an offender elects to co-offend are chosen from a subject-pool of potential accomplices. However, very little is known about the subject-pool from which an offender selects and why some accomplices are selected while others are not. Offline, co-offending partnerships are typically homogeneous, across personal attributes (Sarnecki, 2001; Warr, 2002; Weerman, 2003), and dynamic, rarely lasting more than one offense (McGloin, Sullivan, Piquero, & Bacon, 2008). However, it is unclear whether these characteristics translate to cyberspace, given the inherent anonymity of the Internet.

The third study of this dissertation focuses on the communities formed within larger networks of CE-related websites and which characteristics connect them. Therefore, the purpose of this study is not to directly address co-offending patterns and processes. Rather, it provides a framework for exploring black market communities in cyberspace and a starting point for understanding the general structure of accomplice networks and how the connections that are formed, and emphasized, influence the subject-pool available for future partnership decisions. If the mechanisms that facilitate criminal practices online and offline overlap, as suggested (Grabosky, 2001; Moule,

Pyrooz, & Decker, 2014; Pyrooz, Decker, & Moule, 2013), then this study has implications for cybercrime and offline research.

This third study focuses solely on the 10 CE-seed networks, excluding the 20 comparison networks from analysis, and explores three main objectives. The first is to identify and describe the structure of website communities. The second is to determine if community structure is driven by homophily. The third is to determine whether website communities are stable over time. To address the first objective, I use the community detection method of faction analysis to identify sub-groups within each network and then compare the characteristics of the websites that comprise each sub-group. To address the second objective, I use a formula adapted by van Mastrigt and Carrington (2013) to measure co-offending homophily, to determine whether websites cluster/connect to other websites that have similar outward characteristics. The characteristics selected for comparison are: sex of victim focus (boy/girl), severity of language (explicit/non-explicit), and the preferred media of distribution (videos/images/stories). To address the third objective, I select the *best* community structure and determine how many websites switch communities, across the duration of the study period (60 weeks), and what precipitated those transitions. The manuscript presented here is currently under review at *Social Problems*.

1.5. Summary

The focus of this dissertation is the evolution of websites within child sexual exploitation networks with an emphasis on designing a custom-designed webcrawler for data collection and creating a modified criminal career framework that can be applied to all cybercrime research. This dissertation provides a unique approach to investigating the criminal career as the career typically pertains to one individual. However, in the online world, the crime is the distribution of illegal material while the career pertains to the life of the website 'overseeing' the crime. Therefore, this dissertation ventures into new territory within criminology addressing the absence of a theoretical framework and empirical research into the link between online social exchange and CE material distribution. In turn, this new way of thinking about online CE will result in a better understanding of how distribution websites differ from non-CE websites. The first study

addresses the validity of relying on automated data collection tools for identifying CE websites and compares their characteristics to sports and sexuality websites, to improve inclusion criteria. The second study focuses on criminal persistence and which website attributes contribute to increased survival. The third study examines the construction of communities within networks, which website traits unite a community, and how websites interact within theirs and surrounding communities.

In addition to taking a mainstream criminological paradigm and applying it to cybercrime, this dissertation has practical implications for law enforcement efforts. First, the continued refinement of the webcrawler used in this dissertation can help aid subsequent investigations, through its ability to accurately analyze webpages at a much higher rate than if the webpages were to be examined manually. This increases the efficiency of investigations and the decreases the psychological strain placed on investigators by decreasing the amount of content they need to observer. Second, study 2 informs law enforcement about how child exploitation websites adapt to law enforcement efforts and where efforts are lacking or need to be improved. Third, study 3 aids law enforcement in understanding the organizational structure of child exploitation websites and how they cluster together.

Chapter 2. Assessing the Validity of Automated Webcrawlers as Data Collection Tools to Investigate Online Child Sexual Exploitation[1]

2.1. Introduction

The increasing prevalence of cybercrime over the last two decades has resulted in a re-examination of existing criminological paradigms and theories, to incorporate cyber aspects or to transition entirely to cyberspace (Bossler & Burruss, 2011; Higgins & Marcum, 2011; Patchin & Hinduja, 2011). One of the challenges to this transition is that the change in environment often requires the application of different analysis methods and/or data collection techniques (e.g., Burris, Smith, & Strahm, 2000; Grabosky, Smith, & Dempsey, 2001; Holt, Blevins, & Burkert, 2010; Karpf, 2012; Layton, Watters, & Dazeley, 2011). Cyberspace provides a unique environment for data collection, but one that is vast, complex and thus, yet to be fully understood. To rectify this, social scientists have formed interdisciplinary partnerships with computer scientists to design innovative methods and tools to collect online data. Because of the abundance of data available, interdisciplinary partnerships have focused on methods for simplifying the collection process of relevant data. In most cases, this simplification involves partly or entirely automating data gathering through the building of custom-designed webcrawlers (e.g., Ball, 2013; Bouchard, Joffres, & Frank, 2014; Chen, 2012; Kontosthathis, Edwards, & Leatherman, 2009). While each webcrawler differs slightly, depending on the intended purpose, in general, webcrawlers scan text on webpages, or in forums or databases,

[1] This manuscript is currently under review at *Sexual Abuse: A Journal of Research and Treatment*.

and compile the data for further analysis (Kanich et al., 2011)[2]. To guide and target the data collection, webcrawlers often follow user-defined criteria placed on the crawling process. Despite the growing abundance of cybercrime researchers using automated data collection tools, and developing them for police investigations (e.g., Dykstra & Sherman, 2013; Saari & Jantan, 2013), no studies have been undertaken, within criminology, with the purpose of validating their ability to discriminate between relevant and irrelevant data. Instead, it is assumed that, through the user-specified rules and their use in other fields, the data collected is accurate and on-topic. However, the criminal nature of the material being collected, and the tendency for people to hide their activities, suggests that a close assessment of the data produced by webcrawlers is needed. This is especially important in the context of online child exploitation where illegal content is often both accessible to the public, and hidden within otherwise legitimate content. Webcrawlers can separate the wheat from the chaff, but with the benefits of automation come the potential for these tools to drift away from its intended target.

The Internet has helped create new types of crimes such as malicious software production and hacking (McGuire & Dowling, 2013). However, it has also aided in modifying existing offline crimes such as fraud and sexual exploitation. As a result of the Internet, the distribution of child sexual exploitation (CE) material has boomed in

[2] Although automated data collection tools can be used to index a singular website, they are just as often used to index larger networks. One example, though with a slightly different purpose, are those used by search engines to index websites throughout the world. While the network limits are sometimes partly (Burris, Smith, & Srahm, 2000; Chau & Xu, 2008) or fully pre-determined (Dykstra & Sherman, 2013; Saari & Jantan, 2013), autonomous webcrawlers have also been used to study unknown criminal networks. For example, Layton, Watters, and Dazeley (2011) used automated data collection techniques to analyze phishing campaign networks. Allodi, Shim, and Massacci (2013) explored the online black market trading of tools that could be used to exploit computer system vulnerabilities. Kanich and colleagues (2011) scanned and monitored spam-advertising websites. Where automated data collection techniques have been most prominent are in the study of online terrorist networks (Ball, 2013; Chen, 2012; Fu, Abbasi, & Chen, 2010; Zhou, et al., 2005).

prevalence to a level never before seen offline (Beech, Elliott, Birgden, & Findlater, 2008). Given the graphic nature of the content and its abundance, researchers have been strong proponents of using webcrawlers for data collection. While the data collected are often assumed, and described, as 'child exploitation' in nature, because of the presence of specific keywords or known images, the ability for a webcrawler to discern between child exploitation material and other, legal, material is unclear.

In the current study, we design a webcrawler tool to collect data on websites associated with the distribution of CE material. Using the Child Exploitation Network Extractor, we construct hyperlinked networks surrounding 10 CE-seed websites from our inclusion criteria: presence of CE-related keywords selected from previous research in the field, and/or known CE images from a database provided by law enforcement. We then compare the CE-seeded networks to those constructed starting from a similar, but legal, genre (non-CE sexuality) and a dissimilar genre (sports). We compare how frequently our CE inclusion criteria appear in non-CE genres' networks, especially sexuality, and what differentiates CE-seeded networks from non-CE-seeded networks. Through this comparison, we can provide support, or recommendations, to the currently used standard for CE identification criteria. We can then determine if the currently used criteria can allow for autonomous searching, and detection, of CE-related websites on the public World Wide Web. We begin with an overview of the benefits and concerns with conducting research in cyberspace, paying specific attention to their impact for online sexual offending research. We follow with a systematic review of the advances in methods used to investigate online sexual offending with an emphasis on CE material distribution and automated data collection tools.

2.2. Literature

The use of, and reliance on, Internet-mediated research (IMR) methods for analyzing various phenomena continues to grow in both primary and secondary research (Hewson, Yule, Laurent, & Vogel, 2003). Across disciplines, the Internet has been utilized to carry out surveys and questionnaires, experiments, interviews, observations, and document analyses (Rasmussen, 2008). Within criminology, IMR has primarily centered on victimization surveys, conducted amongst high school and college

students (e.g., Bossler & Holt, 2010; Choi, 2008) and content analyses of criminal networks (e.g., Décary-Hétu & Dupont, 2012). While some disciplines, such as psychology (Gosling & Mason, 2015), have been quick to embrace IMR methods, scholars in other disciplines have noted the hesitancy of their colleagues to fully embrace them (Farrell & Peterson, 2010).

The rise in IMR has been attributed to several important advantages. Whether the data being collected is related to cyberspace practices, scholars with limited budgets and time restraints find IMR to be a suitable substitute to in-person methods (Hewson & Laurent, 2008). More importantly, cyber-based research provides access to a vast, globally diverse, subject-pool, accessible 24-hours-a-day. For (deviant) sexual behaviour research, this means connecting with specialized populations and subgroups that may otherwise be difficult to access (Denney & Tewksbury, 2013; Durkin, Forsyth, & Quinn, 2006; Spink, Ozmutlu, & Lorence, 2004). IMR also provides safer and more secure connection to sex offender populations, benefiting researchers and participants. For researchers, the semi-anonymity of the Internet has allowed researchers to covertly interact with online offenders (Quinn & Forsyth, 2005; Williams, Elliott, & Beech, 2013). For participants, it has increased participation from those who are unable or unwilling to do so offline, reducing bias while increasing objectivity (Hewson, Laurent, & Vogel, 1996). IMR has also led to marginalized populations feeling empowered to share their experiences (Hughes, 2012; Mann & Stewart, 2012).

Hesitation to fully embrace the Internet as a domain for conducting research has focused on the lack of control over the participant environment, while taking part in the study, and the difficulties in managing study access and appearance (Reips & Birnbaum, 2011; Reips, Buchanan, Krantz, & McGrawn, 2011). Even proponents of IMR have raised concerns regarding the quality, representativeness, reliability, and validity of Internet data (Schonlau, van Soest, Kapteyn, & Couper, 2009; Shropshire, Hawdon, & Witte, 2009). More specifically, because of the clear advantages of IMR, there are fears that the new and adapted techniques and methodologies have yet to undergo rigorous validation. To address some of these concerns, research examining the limitations of Internet data collection have grown exponentially. These studies have concluded that the quality and representativeness of Internet-mediated data is comparable to offline

survey data (e.g., Chang & Krosnick, 2009; Dillman, 2007). While there is evidence that IMR data is valid, in general, there are specific situations where the evidence is less clear. Examinations of online and offline sexual offenders has shown that the two differ on key demographics (Babchishin, Hanson, & van Zuylen, 2015; Elliott, Beech, Mandeville-Norden, & Hayes, 2009; Seto, Hanson, & Babchishin, 2011). Given the possible differences between online and offline sexual offenders, it is unclear whether the assumptions about offline sexual offending practices can be directly applied to online practices and vice versa. Therefore, the collection of online sexual offending data needs to be validated independently, and differ from other type of material found online, including legal, sex-related material. A key part of that validation process is identifying the best criteria to differentiate between websites that may potentially contain online child exploitation and others.

2.2.1. Innovative Online Sexual Exploitation Research Methods

Unique to the cyber-environment is the opportunity to safely observe, control, and possibly even induce, the criminal event; thus providing important information regarding the objectives and criminal processes unable to be effectively studied offline. This opportunity has facilitated the creation of fictionally vulnerable computer systems, known as honeypots, to observe the motives and techniques used by hackers (Almutairi, Parish, & Phan, 2012; Marin, Naranjo, & Casado, 2015; Provos & Holz, 2007; Spitzner, 2003). This ability to observe and collect data unobtrusively has also led to the study of strategies, tactics, motives, and rules of cyber-communities involved in the distribution of live webcam recordings of sex (Roberts & Hunt, 2012).

Offline, non-incarcerated sexual offenders, or those in the early stages of their criminal career, are often difficult to access in part to real, or perceived, risks to safety. Online, perceptions of anonymity, by offenders, have allowed researchers to survey sexual interests in children (Wurtele, Simons, & Moreno, 2014), online viewing habits (Seto et al., 2015), and the personality and demographic differences between consumers and non-consumers (Ray, Kimonis, & Seto, 2014; Seigfried, Lovely, & Rogers, 2008). Anonymity has also provided the opportunity to obtain background characteristics, sexual preferences, attitudes, and motives for men seeking sexual

services online (Milrod & Monto, 2012) and compare them to offline customers (Monto & Milrod, 2014), to investigate sex tourism networks (Chow-White, 2006; Evans, Forsyth, & Wooddell, 2000), and to understand the interplay between online negotiations and offline sexual risk reduction (Rice & Ross, 2014). However, the abundance of data available for collection on the Internet, and the desire of researchers to have as much information as possible, has led to the development of automation of data collection techniques.

While sex offenders take considerable risks conducting their illicit affairs in the public domain of the World Wide Web (WWW), we know from offline deterrence research that the threat of a long prison sentence does not deter crime. On the Internet, the perceived, and often real, anonymity provided lead many offenders to conduct criminal activities in the public domain (Armstrong & Forde, 2003; Holt, Blevins, & Burkert, 2010; Maimon, et al., 2014). For offenders operating in the public domain, one of the easiest ways to avoid detection would be to continually modify one's moniker. However, research reveals that offenders maintain the same pseudonym throughout their 'online criminal career' as it becomes associated with notoriety and respect, which outweigh the corresponding costs (Decary-Hetu & Dupont, 2012; Decary-Hetu, Morselli, & Leman-Langlois, 2012). Specific to child sexual exploitation, Wolak, Finkelhor, and Mitchell (2005) found that very few CE-related offenders used any security measures to hide their activities. The argument could be made that those who avoided detection were the ones who used security measures and that the knowledge of detection-avoidance tactics were significantly less in 2005. However, the current availability of CE-related material on the WWW suggests that the findings of Wolak, Finkelhor, and Mitchell still hold today.

Examining the distribution of CE images, Carr (2004) identified the WWW (e.g., public websites) as the second most prominent method of acquisition. More recently, O'Halloran and Quayle (2010) conducted a qualitative study of boy love support forums and found that despite public forums being 'old technology', they were still being used prominently. Similarly, Tremblay's (2006) study of boy love forums highlighted the use of public spaces for the discussion of illicit activities, such as the distribution of CE material. Within peer-to-peer networks, researchers have highlighted the excessive distribution of CE-related material on public and semi-public networks (Fournier et al., 2014; Rutgaizer

et al., 2012; Steel, 2009). Finally, an annual analysis by the Internet Watch Foundation (2014) identified “31,266 URLs contain[ing] child sexual abuse imagery” and that 77% of these URLs were located at .com, net, ru, org, and info domains. Combining these findings, not only is the public WWW a suitable domain for conducting research on illicit sexual activities, it appears to be reflective of an important population among child sexual exploitation offenders, and subsequent distribution techniques.

2.3. Current Study

The distribution of CE material is conducted using a variety of online and offline methods (Callanan, Gercke, DeMarco, & Dries-Ziekenheiner, 2009; Faber, Mostert, Faber, & Vrolijk, 2010; Fortin & Corriveau, 2015; van Wijk, Nieuwenhuis, & Smeltink, 2009). For those researching CE distribution, the graphic nature of material, combined with the abundance of data available, has made automated data collection tools appealing; especially for longitudinal studies (Latapy, Magnien, & Fournier, 2013; Wolak, Liberatore, & Levine, 2014). Automated data collection tools (e.g., webcrawlers) have incorporated CE-related keywords (Frank, Westlake, & Bouchard, 2010), to guide the process, and have examined peer-to-peer networks such as Gnutella (Steel, 2009), eDonkey (Fournier, et al., 2014), and BitTorrent (Rutgaizer, Shavitt, Vertman, & Zilberman, 2012), and publicly accessible websites (Westlake, Bouchard, & Frank, 2011). More recently, researchers have begun to incorporate CE image databases into existing search criteria, after determining that keywords alone did not provide a full picture of the distribution network (Westlake, Bouchard, & Frank, 2012). Incorporating social network analyses, researchers have used webcrawlers to identify key players (Westlake, Bouchard, & Frank, 2011) and cliques (Iqbal, Fung, & Debbabi, 2012) within CE networks and the most optimal strategies for network fragmentation (Joffres, Bouchard, Frank, & Westlake, 2011). Despite the increasing use of CE-related keywords and, more recently, CE image databases, no research has determined whether these criteria adequately distinguish between CE and non-CE-related data. To investigate the effectiveness of commonly used CE identifying criteria we a) determine the prevalence of inclusion criteria in CE, non-CE sexuality, and sports seeded networks; b) identify other website and network characteristics that distinguish CE-seeded networks from

comparisons, across multiple waves, including more than a year after the start of data collection. By identifying the differences between CE-seeded networks and comparisons we can refine and improve on the existing criteria used within the field.

2.4. Methods

2.4.1. Webcrawler (Child Exploitation Network Extractor)

The webcrawler, referred to as the Child Exploitation Network Extractor (CENE), designed for this study follows a similar structural and functionality design as automated data collection tools used by search engines and researchers to index websites (Burris, Smith, & Strahm, 2000; Chau, Shiu, Chan, & Chen, 2007; Chau & Xu, 2008; Frank, Westlake, & Bouchard, 2010). CENE was designed to follow a method similar to that of a person browsing the Internet looking for illegal material. CENE begins at a user-specified website, analyzes the hypertext markup language (HTML)[3] of a webpage, and collects information about the website's structure and pertinent characteristics. It then recursively follows hyperlinks[4], found on a webpage, to other websites and continues the analysis process. Similar to when a person was browsing, the webcrawler scans the linked webpage to determine if the website is relevant to the topic being searched. If the website does not meet the pre-defined criteria, it is discarded; as it would be by a user who views the webpage, sees that it is not what they are looking for, and closes the website.

As the Internet is infinitely large, first, data collection size limits were placed on CENE with regards to the number of websites (approximately 300) and webpages[5] (500,000) included. Second, a set of 'safe' websites were identified and programmed into the webcrawler as being off-topic. Included in this list were popular businesses such

[3] HTML is the standard coding language used to create webpages.

[4] Text found on a website that, typically when clicked, is activated and transports the user to a separate webpage or website.

[5] Websites are comprised of a collection of webpages, each connected to one another under the main website name (e.g., www.website.com). Therefore, the 500,000 webpages were collected from the 300 websites.

as Microsoft©, Google©, Facebook©, and potential false positives (e.g., Disney©). Third, and most importantly, criteria associated with CE material were implemented to guide the webcrawler's search and decision-making process about the inclusion of each website it scanned. For a website to be included, the hyperlinked webpage had to contain at least 7 of our 82 keywords and/or 1 known CE image, as identified by the integrated database provided by the Royal Canadian Mounted Police (RCMP). Again, this simulated the search process of a user in requiring specific material to be present for them to remain on the website. Given the limitations placed on CENE, the network of websites collected should not be viewed as exhaustive but rather a representation of what a typical user might do during their search. This also means that the data collection on each website may also not be exhaustive; depending on size of the website. However, the size is large enough to draw conclusions regarding the validity of automated data collection methods and for mapping illegal networks online.

Once the defined limitations of the data collection were met, CENE compiled the data and aggregated it to the server level. In other words, the data on www.website.com/webpage1 and www.website.com/webpage2 were summed and listed under www.website.com. A list of all websites and webpages scanned by CENE were stored and reused during subsequent crawls. This ensured that the same webpages, if they were still online, were analyzed at each time point.

Webcrawler Criteria

To identify CE-related websites, CENE integrated a database of 2.25 million hash values[6] collected during RCMP investigations, for the purpose of prosecution. Last updated on June 1st, 2012 for our purposes (CENE was launched July 2012), the database was divided into three categories (see Table 2-1). These three categories were based on Canadian legal definitions of CE material. As important differences exist internationally regarding legal definitions (see Gillespie, 2012), we quickly summarize

[6] The database is a collection of hash values identifying CE images. A hash value is a 32-hexidecimal code which functions similar to a digital fingerprint. Each computer file is given a hash value based on its binary composition. When a file is edited, even minimally, a new hash value is created. Tretyakov, Laur, Smant, Vilo, and Prins (2013) state that the chances of two distinct files having the same hash value is 'negligibly small' ($1/2^{2048}$).

Canadian law. Under section 163.1 (1) of the Canadian Criminal Code (1985), *child pornography* includes any 'photographic, film, video, or other visual representation…written material…or audio recording' of a person under the age of eighteen, engaged in an explicit sexual act, or advocating sexual activity. Those depicted can include imaginary people.

Table 2-1: Description of the categories of keywords and hash values used by the web-crawler

	Keywords (Number)	*Hash Values (Number)*
Category 1	Child Exploiter-Code (27)	Child Exploitation (618,632)
Category 2	Thematic (23)	Child Nudity (652,223)
Category 3	Sex-Oriented (32)	Collateral (981,231)

The first database category (*Child Exploitation*) contained 618,632 images all of that were classified, under the Canadian Criminal Code, as being CE. The second (*Child Nudity*) contained 652,223 images that would probably be considered CE by a judge. However images in this category were not blatant and, thus, the risk averse nature of law enforcement resulted in these images being placed into a separate category. The third (*Collateral*) contained 981,231 images that were important enough to be collected by offenders but would not be defined as CE under the Canadian Criminal Code. For example, the initial images in a photo-shoot whereby a child was still clothed and not being directly sexually exploited would be in this category. For an image to be included in a website's count, it had to be at least 150 pixels (approximately 2 inches, or 4 centimeters) by 150 pixels.

CENE used a set of 82 keywords, selected from previous research conducted on the topic of online CE (LeGrand, Guillaume, Latapy, & Magnien, 2009; Steel, 2009; Vehovar, Ziberna, Kovacic, & Dousak, 2009). The 82 keywords were initially classified into three broad categories (see Table 2-1). The first were *code* keywords (27) commonly used by offenders to alert one another to material (e.g., pthc[7]). The second were *thematic* keywords (23) not directly linked to CE but typically present (e.g., boy, girl, child). The third were *sex-oriented* keywords (32) that referenced sexual organs or

[7] Pthc is an acronym for the term preteen hardcore and is one of the most prevalent code words.

acts (e.g., pussy, cock, oral). As the webcrawler only searched for the presence of keywords, the context of the keywords' use was not able to be determined. This is a limitation of keywords as a criterion; however, the general patterns found using keywords commonly linked to CE material are important. To aid in addressing this limitation, we selected roots of words (e.g., bath instead of bathing) and included multiple spellings (e.g., paedo and pedo). However, the use of short keywords, examined outside of their context, means that the webcrawler could, for example, identify a website as containing the word 'anal', when it actually contained 'analyze' or 'analogy'. We address this limitation in more detail below.

2.4.2. Data

CENE collected data on 30 networks surrounding *seed* websites. Ten of the networks began from a CE-seed while the remaining 20 comparison networks began from a sexuality-seed (10) or sports-seed (10) website. Using a repeated measures design, data were collected in 10 waves, at an interval of 42 days. Network composition followed a snowball sampling method via hyperlinks between websites (Burris, Smith, & Strahm, 2000; Westlake, Bouchard, & Frank, 2011). As the nature of the seed can bias the sample derived from snowball sampling (Heckathorn, 2007; Salganik & Heckathorn, 2004), half of our seed websites began with a *blog* while the other half began with a *site*. This aided in maximizing network diversity, allowing us to determine whether the starting point influenced the created network. A *blog* was defined as a website with user-generated posts, in a traditional web-log setup. A *site* was defined as a website with interlocking webpages that did not meet the criteria of a blog. This included discussion forums and photo galleries.

Seed Website Selection

Each of the initial 30 CENE crawls began with a seed website selected by the researchers. For the 10 CE networks, seed websites were selected from two sources. The first source was a list, provided by the RCMP, of websites known to be involved in the distribution of CE material. This list accounted for four (two blogs and two sites) of

our 10 seeds. These four were chosen because they did not require registration to view content on the website[8]. The second source was a list of websites identified, and inspected in previous research, to be involved in the distribution of CE material. This list accounted for the remaining six seed websites. Each CE-seeded network initially included an average of 305.10 (s.d. = 2.33) websites and was recrawled every 42.14 (s.d. =4.45) days.

For the 10 non-CE sexuality networks, seed websites were chosen using Google© search engine. Several search terms were used, to include a broad spectrum of websites related to sexuality. Four (two blogs and two sites) seeds were sex-education websites selected using the terms *sexuality* and *education*. The remaining six seeds were adult pornography websites. Three (one blog and two sites) were selected using the term *BDSM*[9] while the remaining three were selected using the term *sex*. For each search, the most popular websites (i.e., the websites that were first in the search results) that met our criteria were chosen. Although the data collected on the non-CE sexuality and sports websites were the same as for the CE websites, no inclusion/exclusion criteria were specified. Each non-CE sexuality network began with an average of 306.30 (s.d. =6.00) websites and was recrawled every 41.62 (s.d. =7.71) days.

For the 10 sports networks, blog seeds were selected using a sports blog (popularity) ranking website[10] while the site seeds were selected using a sports marketing (i.e., popularity) website[11]. Websites that tailored to specific teams were excluded while those covering an array of sports were preferred. Each sports network began with an average of 301.40 (s.d. =1.65) websites and was recrawled every 41.69 (s.d. =4.44) days.

[8] We excluded websites that required registration for three reasons. 1) Websites use a variety of methods for registering users. Therefore, the additional coding required to address each method was beyond the capabilities of CENE; 2) Even if multiple registration methods are included in CENE coding, websites use different tools, such as CAPTCHA images and sounds and unique questions to minimize bot registration; 3) There were potential legal and ethical issues with accessing private websites.
[9] 'BDSM' stands for a) bondage and discipline; b) sadomasochism; and c) dominance and submission (Wiseman, 1996).
[10] http://labs.ebuzzing.com/top-blogs/sports
[11] http://www.marketingcharts.com/

Composite Measures

Network and website characteristics were compared across the 30 networks, between seed-type (blog/site) and genre-type (CE, non-CE sexuality, and sports). Two additional composite measures were created using subsets of keywords.

Sex Focus: Websites were classified as being either boy or girl oriented based on the relative frequency of specific keywords. *Boy* keywords were: boy, son, twink, penis, and cock while the *Girl* keywords were: girl, daughter, nimphets/nymphets, Lolita/lola/lolli/lolly, vagina, and pussy.

Content Focus: Websites were classified as being either explicit or non-explicit focused based on the relatively frequency of specific keywords. *Explicit* comprised 21 keywords related to severe sexual abuse (e.g., cries, torture, and rape) while *Non-Explicit* comprised 15 keywords related to personal characteristics (e.g., innocent, lover, smooth).

Network Measures

The Internet is an ideal medium for extracting and analyzing social networks, given the inherent nature of online interactions and digital information (Hogan, 2008). Online interactions are directed in that there is usually a sender and a receiver, while the encoding process of digital information makes identifying network connections straightforward. For websites, this is evidenced by the directing from one website to another through encoded text, known as hyperlinks. Thus, website networks can be conceptualized as evidence-based hyperlinked networks (De Maeyer, 2013; Park, 2003; Rodriguez, Leskovec, & Scholkopf, 2013). Using measures of network cohesion—density, clustering coefficient, and reciprocity—provides important information regarding how networks containing websites involved in the distribution of CE material compare to non-CE networks. Cohesion measures were compared across network genre and seed-type at Wave 1 and Wave 10.

Density: Measures the proportion of direct connections present between websites in relation to all possible network connections (Garton, Haythornthwaite, &

Wellman, 1997). Density is used to determine how cohesive is a network and how effectively websites communicate with one another.

Clustering Coefficient: Examines the likelihood that two connected websites are connected to a same third website (Median, Matta, & Byers, 2000). Compared to density, the clustering coefficient informs whether network ties are evenly distributed or whether websites are clustered in sub-groups.

Reciprocity: Identifies the proportion of nodes that directly reference one another (Wasserman & Faust, 1994). That is, if website A references website B, does website B also reference website A? Like density, reciprocity measures global camaraderie or whether websites operate in isolation.

2.5. Results

2.5.1. Validating Selection Criteria

While CE-related keywords have been used prominently to identify material, the only true valid measure is hash values, because it confirms the presence of a known illegal image. Yet, there are trade-offs to both. The problem with hash values is that they are dependent on the images having been unmodified and detected in prior police investigations. A keyword strategy brings more false positives but avoids the narrow funnel of a pure hash value strategy. The total frequency of known CE images criterion and per webpage frequency of keywords criterion, along with subsets of keywords (per webpage) are presented in Table 2-2. An initial within-genre comparison showed no significant differences, across any set of keywords, between seed-type; therefore Table 2-2 displays the combined blog-seed and site-seed networks. Measures for websites within CE-seeded networks differed from non-CE sexuality or sports websites are noted.

The use of hash value databases and keywords appear to be an effective criterion for delineating between CE websites and both similar (non-CE sexuality) and dissimilar (sports) websites. First, regardless of classification category, all but 2 of the 22,729 hash values in our database identified at Wave 1 were located in CE-seeded

Table 2-2: Presence of webcrawler criteria across three network genres at Wave 1 and Wave 10

	CE Networks	Non-CE Sexuality Networks	Sports Networks
		Total Images	
		Child Exploitation	
Wave 1	12,966	0	0
Wave 10	239	7	0
		Child Nudity	
Wave 1	481	2	0
Wave 10	6	0	0
		Collateral	
Wave 1	9,180	0	0
Wave 10	2,782	1,240	5
		Keywords Per Webpage	
Code	0.11	0.13	0.16
Thematic	195[ab]	64	68
Sex-Oriented	124[ab]	77	39
Boy	146[ab]	25	39
Girl	16[b]	16	3
Explicit	46[ab]	17	8
Non-Explicit	191[ab]	31	21

a: statistically different compared to non-CE sexuality networks ($p<0.01$).

b: statistically different compared to sports networks ($p<0.01$).

networks. This includes all 12,966 *Child Exploitation* images identified. By Wave 10 there were 81% fewer identified hash values, suggesting those that were present a year previous (Wave 1) had been removed. Of the 4,274 CE images identified in Wave 10, 71% were located in CE-seeded networks. Although 1,240 *Collateral* images were identified within non-CE sexuality networks, 94% were located on three websites.

Overall, the use of keywords was a valid criterion to discriminate between the websites connected directly and indirectly to a CE seed, compared to others (Table 2-2), but its use comes with a caveat. *Code* keywords, the expected most reliable subset, was found equally across network genres. Moreover, their presence was carried by several

outliers. For example, in one sports network, a website had 140,047 references to *pthc*[12], while in a non-CE sexuality network, one website referenced *paedo*[13] 35,630 times.

When these outliers were removed, each network's average fell from 475 to 10 and 131 to 14 respectively. The impact of outliers was not limited to comparison networks. Within one CE-seeded network, a website referenced *paedo/pedo* 351,445 times while in another a different website referenced it 260,307 times. For each website, these accounted for 99% of all code keyword references. Likewise, their removal resulted in similar drops in network averages. Most important to our findings is: of the 27 code keywords, almost 60% (16) were present less than 0.1 times *per website* in CE-seeded networks, with 6 never being present. This stresses the importance of selecting current criteria, rather than criteria that has been useful in previous research, and is discussed in more detail below.

Beyond code keywords, *thematic, sex-oriented, explicit,* and *non-explicit* subsets of keywords were significantly ($p<0.01$) less frequent within the sports networks and non-CE sexuality networks (Table 2-2). While the lower frequencies in sports networks are expected, the lower frequencies within non-CE sexuality networks is important to highlight. The nature of adult pornography suggests that sex-oriented (e.g., sex and naked) and explicit (e.g., anal and fuck) keywords would feature prominently within non-CE sexuality networks. The higher frequency in non-CE sexuality networks, compared to sports networks, supports this hypothesis. However, the even higher frequency in CE-seeded networks suggests that while the mere presence of explicit and sex-oriented keywords does not distinguish CE-related websites from non-CE sexuality websites, the quantity of the keywords can. As a result, the keywords function as an important quantity criterion rather than a presence criterion. Combined, our hash values and keywords

[12] A manual examination of the websites with abnormally high rates of code keywords, in this example 'pthc', found that their abundance was the result of the html code. On fantasysnews.cbssports.com, 'pthc' was a part of a coding reference to 'depthchart'. Although only identified in isolated cases, it is a limitation that we address in more detail below.

[13] A manual examination of several non-CE sexuality websites with high frequencies of code keywords revealed an important pattern. The use of 'paedo' was actually a security measure. Like the criterion we placed on the webcrawler to exclude 'safe' websites, some websites had a script built into their source code that scanned a user's post and would filter inappropriate keywords. In the example highlighted here, the website had a list of 'bad words' that included 'paedo'.

findings suggest each distinguishes one genre (e.g., child exploitation) from both similar (non-CE sexuality) and dissimilar (sports) genres. However, the frequency and current relevance (i.e., popularity) of keywords have mediating effects.

2.5.2. Comparing CE Networks and Non-CE Networks

Hash value databases and code keywords have been the primary indicators used by researchers to distinguish CE related websites from non-CE related websites. However, a key question is whether websites within networks beginning with a CE-seed differ in any other significant ways. Preliminary comparisons of within-network-genre show that the seed-type did not modify the website characteristics within a network. Table 2-3 summarizes the general website characteristics of each network genre: average number of websites, webpages per website, and images and videos per webpage. The median value for each characteristic is also displayed while characteristics where CE-seeded networks differed ($p<0.01$) from non-CE sexuality and/or sports networks are noted.

General website characteristics, summarized in Table 2-3, show that CE-seeded networks are easily distinguishable from sport networks, having more webpages per website, and images and videos per webpage. At the same time, they are less distinguishable from non-CE sexuality networks, having more images per webpage and outgoing and incoming hyperlinks. The limit placed on the size of each network (approximately 300 websites and 500,000 webpages) means that we expect the average number of websites and webpages to be comparable across genres. However, websites within CE-seeded networks, and non-CE sexuality networks, were significantly ($p<0.02$) larger. The larger size of sex-based websites (CE, pornography, or education) may be the result of volume of material or, in the case of CE, a function of the detection avoidance strategy of hiding content on hard-to-reach webpages.

Table 2-3: General website characteristics across three network genres at Wave 1

	CE Networks	Non-CE Sexuality Networks	Sports Networks
Avg. Nb. of Websites	306	306	302
Avg. Nb. of Webpages	1,583[c]	1,837	1,084
(Median)	(77)	(71)	(20)
Nb. of Images per	20.7[ab]	5.9	4.9
Webpage (Median)	(7.0)	(1.9)	(2.0)
Nb. of Videos per	0.16[b]	0.21	0.04
Webpage (Median)	(0)	(0)	(0)
Outgoing Hyperlinks	23[ab]	15	31
(Median)	(10)	(8)	(12)
Incoming Hyperlinks	23[ab]	15	31
(Median)	(20)	(8)	(13)

a: statistically different compared to non-CE sexuality networks ($p<0.01$).

b: statistically different compared to sports networks ($p<0.01$).

c: statistically different compared to sports networks ($p<0.02$)

Advances in video recording technology and Internet bandwidth suggest that there may be an increase in the amount of CE videos being distributed going forward. Despite this hypothesis, Table 2-3 shows that distribution remains heavily image-based, suggesting that CE images can be an effective criterion to find child exploitation websites. Networks beginning with a CE-seed website averaged 21 images per webpage while non-CE sexuality and sports networks averaged 6 and 5 respectively. Although CE-seeded networks and non-CE sexuality networks displayed similar rates of videos per webpage, when non-CE sexuality networks were divided into pornography and sex education (SE), the results differed. Within SE-seeded networks webpages averaged 0.07 videos (similar to sports-seeded networks), while within pornography-seeded networks webpages averaged 0.44 videos. The abundance of images in CE-seeded networks may reveal: a) the continuing tendency towards image-based distribution, as it is easier; b) a slower movement towards CE video distribution; or c) that videos increase risk of detection, through possible recognition of video features. Regardless, the higher rate of videos per webpage in pornography-seeded networks suggests that, over time, CE distribution may include more video distribution.

Individual characteristics are important for describing websites, however, websites and their corresponding network do not operate independently. From the structure of a network, information regarding individual website behaviour and patterns can be determined. Table 2-4 summarizes the change in network cohesion across the three network genres, at Wave 1 and Wave 10. Sports networks are divided between blog-seed and site-seed as the two significantly differed on hyperlinking and each network cohesion measure. Significant differences between network-genres (and seed) are noted.

Hyperlinks to and from a website can be viewed as a proxy measure of network insulation (outgoing) and popularity (incoming). For illegal websites, hyperlinking potentially increases the risk of detection but also the opportunity to attract consumers. CE-seeded networks hyperlinked more than non-CE sexuality-seeded and sports site-seeded networks and less than sports blog-seeded networks (Table 2-4). While the median number of outgoing and incoming hyperlinks did not vary within non-CE sexuality and sports networks, they did for CE-seeded networks (10 and 20). This

Table 2-4: Network cohesion measures at Wave 1 and Wave 10

	CE Networks	Non-CE Sexuality Networks	Sports Blogs	Sports Sites
	Hyperlinking at Wave 1 (Median)			
Outgoing	23[abc] (10)	15 (8)	45 (17)	15 (10)
Incoming	23[abc] (20)	15 (8)	45 (17)	15 (10)
Density				
Wave 1	39%[abc]	5%	15%	5%
Wave 10	7%[abc]	5%	14%	5%
Clustering Coefficient				
Wave 1	0.43[b]	0.42	0.59	0.45
Wave 10	0.43[c]	0.40	0.48	0.38
Reciprocity				
Wave 1	23%[b]	24%	55%	24%
Wave 10	22%	23%	36%	19%

a: statistically different compared to non-CE sexuality networks ($p<0.01$).

b: statistically different compared to sports blogs ($p<0.01$).

c: statistically different compared to sports sites ($p<0.01$).

suggests that within CE-seeded networks there are websites that act as directories, connecting users to all websites, and those that act as suppliers, connecting to only selective websites. Combined with similarities in reciprocity, to comparison networks, it appears directory-based websites hyperlink to all suppliers, while suppliers do not hyperlink to one another but *do* hyperlink back to some directory-based websites. This conclusion is supported by network density and clustering coefficient. CE-seeded networks had a higher proportion of all possible network connections present (density) that held until Wave 10. However, they were equally as likely as comparison networks to have two websites that were connected to one another be connected to a common third (clustering). In other words, more websites acted as hubs (brokers) within CE-seeded networks, compared to within non-CE sexuality and sports networks. This finding has implications for how we conceptualize distribution and competition within online illegal networks and removal strategies.

2.6. Discussion

Automated data collection for Internet-mediated research is an efficient, cost-effective, technique and can be useful for collecting data on topics of a sensitive or graphic nature (e.g., child sexual exploitation). Although useful, these tools come with the caveat that by not looking at the data directly there are questions regarding the validity of the data collected. In the current study we began to address questions surrounding validity by designing a webcrawler to collect data on networks surrounding CE-seeds and compare the presence of CE criteria to similar (non-CE sexuality) and dissimilar (sports) networks. From the topic-specific criteria, selected to guide the webcrawler, we a) provide recommendations for improving the reliability of selection criteria for CE investigation and researchers; and b) outline how CE-seeded website networks differ from comparisons and what these differences tell us about how they function.

Research into the distribution of CE material in cyberspace has primarily used image hash value databases and keywords to identify CE websites and content. Our results show that images from all three of our database's categories were prevalent in CE-seeded networks and minimally in non-CE-seeded networks (Table 2-2). *Code*

keywords –those suggested as identifiers by past researchers- were equally prevalent across all three genres of networks; however, other 'non-code' keywords were more frequent in CE networks. Based on these results, we have recommendations for improving selection criteria. First, specialized databases (e.g., hash values) are a valid criteria for identification. However, their reliability, and hence usefulness, is contingent on the completeness of the database. For CE research, the abundance of images being distributed coupled with the simplicity of changing a hash value means that hash value databases, alone, are not a reliable criterion (Westlake, Bouchard, & Frank, 2012). Second, the Internet is a fast-paced, evolving, environment. Subsequent research cannot rely solely on the findings of previous research for choosing selection criteria. Of the 27 *code* keywords from past research (LeGrand, Guillaume, Latapy, & Magnien, 2009; Steel, 2009; Vehovar, Ziberna, Kovacic, & Dousak, 2009), more than half (16) were minimally present on *any* website. Even code keywords that persist are not particularly useful, if used alone, as we found them on other genres of websites. On other genres' websites, code keywords were identified as parts of unrelated keywords and on lists of 'banned' keywords. As the webcrawler examined the html source code for a webpage, it registered the code keyword's presence despite its presence being an attempt by the website to exclude the keyword. Therefore, we believe that keywords are a useful selection criterion provided that they a) reflect the current, not historic, landscape of the topic of research; b) cover a wide range of related keywords (e.g., thematic); and c) are used more as an inclusion criterion rather than an exclusion criterion, as some keywords have different meanings in different contexts, fall out of favor or are adopted by other sub-groups over time.

For researchers, and companies attempting to detect and remove CE content from the Internet, our findings have important implications. Any search for CE-related content needs to include a multi-criterion approach that is updated regularly. In addition, the variety of content available suggests that research, and identification, needs to focus on specific types. This may be type of exchange platform (e.g., public/private website or peer-2-peer networks), victim (e.g., age or sex), or media. The graphic nature of the material being examined has also been a hurdle for continued research into online CE-related distribution. From our research, those studying, or combating (allied professionals), online child sexual exploitation have evidence that criteria can be used to

autonomously identify CE-related sources (websites) using automated data collection techniques. This extends beyond distribution, as researchers examining discussion forums (e.g., Fortin & Corriveau, 2015; Tremblay, 2006) or other user networks can modify the criteria to specific topics. For example, add specific keywords to target 'boy love' forums, trafficking, or live webcam broadcasts.

Our second objective was to identify how CE-seeded networks differed from similar and dissimilar genre networks. Hash values and the frequency of keywords easily distinguished sports networks (dissimilar) from CE-seeded networks but were less effective at distinguishing networks with overlapping focus (sex). While the higher quantity of (all) images in CE-seeded networks, compared to non-CE sexuality networks, point to CE distribution still being heavily dominated by images rather than transitioning to videos –an important findings regarding the evolution of online web-based CE distribution- there was one finding that we wanted to pay particular attention. Networks beginning with a CE website differed from comparison networks in hyperlinking practices (Table 2-3) and subsequent network cohesion (Table 2-4). Specifically, CE-seeded networks were more densely connected, but were structured with hubs that controlled connectivity and thus information. This practice, unique to CE-seeded networks, may be a survival tactic. If consumers are aware of 'hubs' that list websites, then distribution-based websites can use these hubs to redirect previous consumers to the new website address, should their original website be removed by control agencies. Hubs also provide a location where distribution-based websites can inform the larger community of their available content and, by extension, increase traffic to their website. Equally as interesting, the proportion of hyperlinks reciprocated and the degree of clustering between subsets of websites within their larger network was similar across all three network genres. This could mean that outside of hubs CE-seeded networks operate similarly to other, legal, online networks and that CE websites may compete with one another at equivalent rates to non-CE websites. This potential idea of competition may appear to counter previous research suggesting a communal aspect to CE (e.g., Beech, Elliott, Birgden, & Findlater, 2008; Estes, 2001; Tremblay, 2006) and cybercrime in general (Basamanowicz & Bouchard, 2011; Dupont, 2013; Holt, 2007). However, we argue that the two are not incongruent. While the Internet has transitioned CE from a solo crime to more community forum-based, for virtual interactions and distribution, this

does not mean that competition is completely non-existent. Instead, we argue that there may be more competition between illegal websites than currently suspected and that it is similar to rates found amongst legal websites.

Expanding beyond individual distribution and understanding how the larger network functions as a whole has implications for control efforts by law enforcement and private agencies (Krone, 2004). More specifically, research into the network structure facilitates identification of the most effective methods for disruption and for how new content is circulated. For those researching online CE distribution, the linkages formed between distributors may be as important as the individual distributors themselves, given the communal aspect of this cybercrime. Coupled with the automatic detection techniques proposed, this research provides an important framework for future research into CE distribution at the micro and macro levels.

2.6.1. Limitations

The validity of an automated webcrawler tool is contingent on the criteria used to guide the crawling process; including the starting websites. Among the 10 CE websites used as seeds, eight were identified as boy-focused. It is likely that these seed websites contributed to approximately 80% of websites within CE-related networks being classified as boy-focused. In the two networks beginning with a girl-focused seed, less than 20% of websites were classified as boy-focused, and *Child Exploitation* images were statistically less frequent.

Given that our study used a Canadian definition of CE material, our criteria does not necessarily translate to research where the definition of CE material is more or less stringent. In addition, the use of individual keywords as a criterion, rather than groups of keywords, can result in false positives. A manual verification of CE-seeded and non-CE-seeded websites with high frequencies of code keywords found that they were used in a different context. This was most evident within the html source code of a webpage, where references to automated scripts or to non-related files (e.g., videos) resulted in a sports website, for example, appearing to include code keywords. Nevertheless, we believe that the patterns found within this study have global applicability. While countries

have differing CE laws, the general context is similar and the issue of cybercrime is not unique to any location. Although individual CE keywords will be used in non-related environments, sets of keywords (including sex-oriented and thematic) will still be more prevalent among CE websites.

Increases to the capability of video recording devices and download speeds point to the potential rise in the number of CE-related videos being distributed in cyberspace and/or live webcam performances by sexually exploited children. As of yet, no database of known CE videos exists. Given this, our webcrawler did not include any video-based criteria. While we collected descriptive information on the number of videos being distributed, on each website CENE visited, we were unable to verify if any were CE-related. The increase in video distribution, across all genres, on the Internet (e.g., YouTube©) point to the need for a) a video-based database, and b) a video-based criteria for CE research and investigations. Specific to the current research, the additional storage space required to display videos, compared to images, may point to video-based distribution being more prevalent on private websites or networks, where users pay for access. Growth in the number of live webcam performances adds further complexity to the collection of data on CE-related websites as the webcam stream may not be directly connected to the website, would regularly include 'new' content and yet be interpreted by the webcrawler as a single video (stream), and would not be catalogued in any video database. Therefore, this study was focused on the distribution of CE-related images and limited with regards to the implications for how CE video-based networks function.

The current study examined publicly accessible websites, excluding websites that required a password or registration to access hidden areas and those found on the *Deep Web*[14]. Given the illegal nature of CE content, a method for avoiding detection is to hide illegal material or requiring registration to access. Another method, that is growing quickly, is to use the dynamic and anonymous Deep Web. Although our study was targeted towards public webcrawler searches, such as on websites or peer-to-peer

[14] The Deep Web refers to the bottom layer of the World Wide Web that is not indexed by search engines and is comprised of dynamic webpages (Bergman, 2001).

networks, our findings reinforce the need for current criteria relevant to the type of network being examined.

The targeting of our study to public websites points to a typical limitation of many automated webcrawler tools. Given the simplicity of their build, many are unable to access private/hidden data. The added complexity of coding a webcrawler to register an account or 'verify' that it is human may be beyond the capabilities of many social scientists. This inherent limitation impacts the validity of any data collected from their use. However, this does not limit the use of webcrawlers as a strategy for gathering data on protected networks, or the Deep Web. Although the scope would not be as large, a webcrawler can be launched from within a password protected area allowing for complete data collection from one source (e.g., discussion forum website). Continued interdisciplinary partnerships are necessary to integrate methods for accessing more secure/private data. This is especially true for criminologists interested in researching underground criminal activities in cyberspace.

2.7. Conclusion

Growth in automated data collection, for cybercrime research, has occurred without the necessary, vigorous, validation of these techniques. Comparing the composition of a series of networks beginning with known child sexual exploitation (CE) websites, to those derived from non-CE sexuality and sports websites, we a) provided recommendations for improving the validity of automated webcrawler tools for CE research; and b) identified criteria and characteristics that can be used to distinguish CE-seeded networks from non-CE-seeded networks. Using criteria selected from previous research on the topic of online CE distribution, we found that image databases was a valid criterion but their lack of completeness limit reliability, while keywords was a reliable criterion but the constant evolution of the Internet limit their validity. Comparing CE networks to non-CE sexuality and sports networks, we found that websites within CE-seeded networks were larger (than sports websites) and more image-based with different hyperlinking properties, while the CE networks were more dense but equal in clustering and reciprocity suggesting the presence of hubs.

The use of automated data collection tools, such as webcrawlers, provide a great advantage to researchers as they are more efficient and require minimal intervention. However, for automated data collection techniques to be useful in any discipline, prior research into the field in question must be completed to ensure that the best and most recent selection/inclusion criteria are chosen. Failure to select relevant criteria, or relying on one criterion rather than several, can lead to high rates of false positives and an inaccurate representation of the current landscape. While their validity and reliability is dependent on the criteria used and their objective, webcrawlers are a great asset for third-party companies (e.g., Blogger©), to ensure that their terms of service are being adhered to by clients, and software engineers, enforcing copyright laws or identifying security vulnerabilities being traded on the black market. In these situations, automated data collection tools can be a first line of defense, by conducting an initial scan of the source, to flag potential issues, that can then be manually verified by humans. The customizability of automated data collection tools mean companies would be able to select the best criteria to fit their needs, or what they are trying to identify, thereby proving to be a reliable criterion. For researchers and organizations combating the distribution of CE material in cyberspace, the graphic nature of some of the material viewed can have substantial impacts on the retention and psychological health of employees (Bourke & Craun, 2014; Krause, 2009). Continued improvements to the validity and reliability of the criteria selected may aid with prolonging the careers of those investigating the crime and increase the number willing to research the topic. Together, better detection techniques and strategies can be developed while the criminal processes involved in distribution can be better understood and explained.

Chapter 3. Criminal Careers in Cyberspace: Examining Website Failure within Child Exploitation Networks[15]

3.1. Introduction

Central to the understanding of criminal activity is how the career of offenders extends longitudinally. Ever since the publication of *Criminal Careers and Career Criminals* (Blumstein, Cohen, Roth, & Visher, 1986) much has been learned about how criminal careers start, change, and end (Laub & Sampson, 2003; Piquero, Farrington, & Blumstein, 2003). However, this research has predominantly focused on individual offenders, excluding entities like illegal websites, and has yet to be fully applied to cybercrime. This is despite the growing concern of cybercrime growth amongst policy makers and an increasing number of criminologists who are dissatisfied with the current state of empirical knowledge in this area (Holt & Bossler, 2014). The criminal career paradigm marked a turning point in our understanding of crime and criminals, and we argue that approaching cybercrime through this lens is a conceptually and empirically productive way to move forward in this new area.

There is no denying the importance of Internet-related crimes to the mix of crimes reported to police over the past 20 years. Growth in global World Wide Web usage has been met with a paralleled growth in offenders either incorporating a digital component to their offline offenses or transitioning entirely to cyberspace (Holt & Bossler, 2014). Despite this growth, researchers, for the most part, have yet to determine whether existing criminological theories and paradigms require new provisions to account for cyberspace or whether they can be directly transposed to

[15] This manuscript has been accepted for publication in *Justice Quarterly* (http://dx.doi.org/10.1080/07418825.2015.1046393).

cybercrimes (for exceptions, see Bossler & Burruss, 2011; Higgins & Marcum, 2011; Patchin & Hinduja, 2011). Applying traditional criminological concepts and measures to cybercrimes is not without challenges. For example, typical offline co-offending predictive measures, such as proximity, race, and age, are not readily available and may not even matter to cyber offenders. The process is further complicated by the relative anonymity of online crime (Holt, Blevins, & Burkert, 2010), the novelty of the methods some use (e.g. Grabosky, Smith, & Dempsey, 2001), and the lack of offenders to study in official crime databases or among inmate populations – resulting in the use of college samples (e.g., Bossler & Holt, 2009, 2010; Choi, 2008). As a result, it remains unclear how characteristics such as offending frequency, crime-type mix, duration, and co-offending manifest online and whether these career dimensions are consistent across different cybercrimes and offending partnerships.

In this study, we propose to analyze websites with illegal content as the main object of inquiry. Specifically, we draw from a repeated measures design to examine the predictors associated with the persistence of websites potentially containing child exploitation (CE) material, over a sixty-week period. In the spirit of snowball sampling studies, our study takes advantage of the networked nature of the Internet by creating our sample through a systematic analysis of the hyperlinks included on “seed” websites. We start with 10 seed websites, with known CE material, and use a custom-designed web-crawling tool to map the surrounding networks, up to a limit of approximately 300 websites per network. The web-crawling tool allows us to automatically collect a number of website characteristics that can be operationalized under the various dimensions of the criminal career. Our goal is to examine which website characteristics are associated with the failure. We begin by reviewing the online persistence literature, and the process of transferring the criminal career dimensions to cyberspace, before turning to a more detailed description of the data and methods.

3.2. Literature

3.2.1. Online Persistence

In approaching the persistence of illegal websites (and using websites as the main unit of analysis), we start by examining the literature on website survival more generally. While a website may persist without being successful (e.g., number of visitors, new consumers, etc.), e-business research notes the importance of appealing to consumer demand for maintaining survival (Robbins & Stylianou, 2003). The ability to appeal to consumers is often predicated on accessibility, speed, navigation, and content quality (Miranda & Banegil, 2004). Although to potentially varying degrees of importance, illegal websites face these same challenges. However, like many illegal market suppliers (e.g. Bouchard, 2007; Bouchard & Ouellet, 2011), websites with illegal content, such as CE material, must balance their appeal to consumers with their ability to avoid detection. At some point, a website (operator) may deem the time and energy required to maintain this balance exceeds the perceived benefits and elect to cease operations.

While an offender may desist from crime for personal reasons, an important component of maintaining survival is an ability to avoid detection. For websites involved in illegal activities, persistence requires the avoidance of detection by various control agencies (e.g., activist & law enforcement) while ensuring that there is enough support to maintain the website. In order to avoid detection, website owners have begun to implement survival tactics such as closing ranks, using passwords for newsgroup access, and taking advantage of remailers, anonymous servers, and Internet Protocol spoofing. Despite these tactics becoming more commonplace, Wolak, Finkelhor, and Mitchell (2005) found that only 20% of offenders charged with possession and/or distribution used any type of security measure to hide their content. Of those, only eight percent used anything beyond simple password protection. In fact, online offenders feel so secure about their anonymity that many share information, detection avoidance techniques, and barter tools/skills, in semi-public to public forums (Armstrong & Forde, 2003). The minimal personal security measures used by offenders may be a result of the “natural” security provided by the Internet that includes anonymity, globalization, and the

ambiguous legal status of many online behaviors (Maimon, Alper, Sobesto, & Cukier, 2014).

3.2.2. Transitioning Criminal Career Dimensions to Cybercrimes

Outside of death, it is difficult to ascertain the cessation of an individual's criminal career as offenders often experience career interruptions (Piquero, Farrington, & Blumstein, 2003). Likewise, the criminal career duration of a website is difficult to quantify. Although it may be possible to observe websites long enough to see them go offline, their start dates are unlikely to be available. Therefore, conceptualizations of desistance are often described as the process whereby an offender decelerates their offending frequency, de-escalates their offending seriousness, and becomes more specialized in their offending crime-type (LeBlanc & Loeber, 1998). In cyberspace, measuring criminal career duration is affected by the ease with which an offender can modify their identity (i.e., change pseudonyms). While maintaining anonymity online, through the changing of a pseudonym, provides a clear advantage for detection avoidance, offenders often maintain the same pseudonym throughout their career as it can become associated with notoriety and respect (Decary-Hetu & Dupont, 2013; Decary-Hetu, Morselli, & Leman-Langlois, 2012). Because of this, we can conceptualize the criminal career duration of a specific pseudonym, and similarly, the duration of a specific website address hosting illegal content[16]. That is, the website address is similar to a pseudonym for an individual offender in that there is associated notoriety and respect with certain website names. With this in mind, we conceptualize website criminal career duration as the amount of time a website remains active at a specific address. The websites analyzed for this study were followed for sixty-weeks, with more than 15% going offline during the observation period.

[16] While continually changing a website's address would function similarly to changing one's pseudonym and thus appear to be a strategic detection avoidance strategy, it comes at the cost of losing website traffic and associated gains. Therefore, the website address should function similarly as a user's pseudonym and be an accurate representation of said website's criminal career duration. Nevertheless, we acknowledge that websites may also experience career interruptions and thus the concept of criminal career duration for a website is similar to that of offline career duration and not contingent on absolute desistance.

Duration does not operate independent of the other three criminal career dimensions –offending frequency, crime-type mix, and co-offending– outlined by Blumstein, Cohen, Roth, and Visher (1986). In this study, we use measures of offending frequency (volume of illegal content), crime-type mix (or crime seriousness), and co-offending (connectivity) as the main predictors of the career duration (survival) for illegal websites. Yet, these measures first need to be adapted to the specific online context in which we apply them. Offending frequency, or lambda (λ), refers to the estimated change in offending frequency - most often amongst chronic offenders - over time (Blumstein & Cohen, 1979; Cohen, 1986). While participation can be inferred simply from a website being online and distributing illegal content, the concept of frequency is a bit more ambiguous when applied to the online context. For websites involved in distributing illegal material (e.g., movies, music or CE media), offending frequency can be measured in multiple ways. If a website begins distributing 10 movies illegally, within a one-hour time frame, do we consider that 10 offenses or one? If a website continues to distribute those 10 movies the following day is it considered a new offense or a continuation of the previous offense? If a website is still distributing those 10 movies two years later, is it still part of the original offense or is it a new offense? One of the difficulties illustrated by this example is that cyber offending is often a cumulative process rather than a series of independent events. That is, a website does not need to stop distributing one item to begin distributing another. Instead, the new item can be added to the existing distribution chain. As a result, offending frequency in cyberspace, and especially for websites, is more appropriately measured via the *volume* of content. This can be viewed as total volume currently available, or the change in volume over a specific period of time.

Crime-type mix often refers to an offender's tendency towards specialization and/or escalation in seriousness. Although we acknowledge the importance of escalation in criminal career persistence, our focus will be on specialization. DeLisi and Piquero (2011) summarized the debate on specialization by stating that almost all offenders are generalists. While the majority of sexual offenders also fit into the category of generalists (Lussier, 2005: Miethe, Olson, & Mitchell, 2006) there is support for modest levels of specialization, especially amongst child molesters (e.g. Harris, Smallbone, Dennison, & Knight, 2009; Lussier, LeBlanc, & Proulx, 2005). For websites disseminating CE

material, the tendency towards general offending or specialization is unclear. Mitchell, Wolak, Finkelhor, and Jones (2011) and Wolak et al.'s (2005) studies of offenders arrested for online child sexual exploitation are the most revealing on this issue. While 30% of those arrested for possessing CE material had prior arrests for nonsexual offenses, 18% had prior arrests for sexual offenses and 13% for sexual offenses against minors (Mitchell et al., 2011). Offenders were also found to possess content in multiple formats (39% possessed videos) and specialized at different rates across age (25%) and sex (76%) of victim and severity of content (Wolak et al., 2005). However, research has also shown that online CE offenders have varying profiles, specifically when comparing various types of consumers and producers (see Babchishin, Hanson, & Hermann, 2011; Elliot, Beech, & Mandeville-Norden, 2013). As websites may reflect consumer demand, we explore their tendency towards specialization through three composite measures of content focus: sex of the victim (boy/girl), severity of content (explicit/non-explicit), and media distribution (images/videos/stories).

Co-offending usually refers to the partnership between those directly involved in a crime (Reiss, 1986). However, some prior studies have argued for the inclusion of the larger pool of associates from which co-offending decisions are made (Warr, 2002), or simply the inclusion of offenders who are indirectly involved in a specific crime as facilitators (Morselli & Roy, 2008; Tremblay, 1993). The larger pool of associates, in which offenders are embedded, are important in providing criminal capital to offenders: offenders often acquire methods or items for a crime, access to criminal organizations, the identification of potential targets, or detection avoidance techniques (McGloin & Piquero, 2010). The network also provides social exchange benefits that extend beyond the criminal event (Gallupe & Bouchard, in press; McGloin & Nguyen, 2013; Weerman, 2003). Likewise, online criminal networks provide criminal and non-criminal benefits (Holt, 2007; Rosenmann & Safir, 2006; Tremblay, 2006) that can translate into co-offending opportunities (Holt, Strumsky, Smirnova, & Kilger, 2012). For website owners, forming connections with other websites may provide similar criminal and non-criminal benefits. Connections between websites are formed through hyperlinks, whereby one website owner places the address of another website on their own. Hyperlinks may be reciprocated (both websites provide a hyperlink to the other), but this is not always the case. As such, it is useful to distinguish between outgoing hyperlinks (the number of

times websites reach out to others) and incoming hyperlinks (the number of times other websites receive a hyperlink), with incoming considered to be a measure of popularity (e.g. Gallupe and Bouchard, in press; Haynie, 2001). Through these two types of hyperlinks we are able to identify the network surrounding a website and, thus, variations in online criminal embeddedness.

3.3. Current Study

Once a solitary crime, with sparse networks, the sexual exploitation of children has seen a boom in prevalence as a result of the ease of access to high quality recording devices, the Internet, and virtual communities (Beech, Elliot, Birgden, & Findlater, 2008; Tremblay, 2006). Quayle and Taylor (2011) found that the social networks formed in cyberspace are a critical factor in the exploitation of youth. Within those social networks, websites act as collection and distribution points for material, while facilitating support and acceptance of offenders (Akdeniz, 2013; O'Halloran & Quayle, 2010; Prichard, Watters, & Spiranovic, 2011). Because of the importance of online social networks, the websites that create the foundation of these networks need to be studied and the determinants of their survival and failure to be better understood.

The current study provides two contributions to the field. First, we propose a research design to monitor an under-explored unit of analysis in criminology - websites hosting illegal content. Second, we conceptualize the illegal website as an entity with its own trajectory or career, and propose criminal career dimension measures adapted to the context of CE networks. To better understand illegal website persistence and failure we aim to answer two questions. First, do survival rates for websites involved in CE differ from two comparison group networks (i.e., sports and sexuality)? Second, do CE websites that fail differ from those that persist on cyber-specific criminal career dimensions?

3.4. Methods

The data for this study come from a longitudinal project with the objective to analyze the evolution of websites involved in the dissemination of child exploitation (CE) material. The dataset consists of 10 CE networks and 20 comparison networks –10 sports and 10 (legal) sexuality. Using a repeated measures design, data were collected 10 times, every 42 days, using a custom-designed web-crawler that followed a snowball sampling method via hyperlinks between websites (Burris, Smith, & Strahm, 2000; Westlake, Bouchard, & Frank, 2011).

The Child Exploitation Network Extractor (CENE) is a revised version of a web-crawler described in prior research on this topic (Frank, Westlake, & Bouchard, 2010; Westlake et al., 2011). In short, CENE operates similar to the web-crawlers used by Google© to index websites. Following a set of researcher-specified criteria, CENE starts from a small subject pool (*seed* websites) and follows the connections (hyperlinks) to other websites and determines if the connected website also meets the study criteria. If it does, CENE collects data on the size of each website, and the number of images, videos, and keywords. This process is repeated until the desired sample size is met.

3.4.1. Seed Websites

Each initial web-crawl began on a seed website selected by the researchers. For the CE networks, seed websites were selected from two sources. The first source was a list of websites provided by the Royal Canadian Mounted Police (RCMP) that were known to be distributing CE material. This accounted for four of our 10 seeds. These websites were selected from the list because they fit our criteria of seed-type and did not require registration to enter[17]. The second source was a list of websites identified, and inspected in previous research, to be involved in the dissemination of CE material. This

[17] We excluded websites that required registration for three reasons. 1) Websites use a variety of methods for registering users. Therefore, the additional coding required to address each method was beyond the capabilities of CENE; 2) Even if multiple registration methods are included in CENE coding, websites use different tools, such as CAPTCHA images and sounds and unique questions to minimize non-human (bot) registration; 3) There were potential legal and ethical issues with accessing private websites.

included, but was not limited to, image or video distribution and being an access point for CE websites categorized and listed by type of content. For the inclusion of every subsequent website the hyperlinking webpage had to contain at least 7 of our 82 keywords *or* at least one known child exploitation image. Prior research in this area showed that a combination of the two criteria was the most effective at minimizing both false positives and false negatives (Westlake, Bouchard, & Frank, 2012). As with any snowball sampling study, the nature of the seed can bias the sample derived and the results (Heckathorn, 2007; Salganik & Heckathorn, 2004). To account for this, we drew from 10 different seeds in order to maximize network diversity and allow us to answer our research questions more accurately than we could if we had a single case study network. Half of all networks began with a "blog" while the other half began with a "site". A blog was defined as a website with user-generated posts in a traditional web-log setup. A site was defined as a website with interlocking webpages that did not meet the criteria of a blog. This included discussion forums and photo galleries. CE networks began with an average of 305.10 (s.d. =2.33) websites and were re-crawled every 42.14 (s.d. =4.45) days.

As the focus of our study was illegal websites *and* the communities that surround them, we did not discriminate between legal or illegal connected websites. Given that some of these websites contained thousands of webpages and hidden directories it was beyond the scope of this study to confirm that the primary focus of each website was to disseminate CE material. However, we attempted to control for website focus through the use of indicators tapping into illegal content dissemination such as CE code words and a police database of CE images. Although we cannot confirm or deny the illegality of each website we believe that survival across the entire community needs to be analyzed as their connectivity provides points of access to confirmed, illegal, content (e.g. our 10 seeds).

In order to determine whether the survival rates of our networks generated from CE websites differed from networks generated from legal websites, we created two comparison groups. One centered on sports websites and the other centered on legal sexuality-related websites. For the sports networks, blog-seed websites were selected

using a sports blog ranking website[18] that ranked blogs based on popularity. The site-seed websites were selected using a sports marketing website[19] that ranked the most popular sports websites on the Internet. Websites tailored towards specific teams were excluded while websites covering an array of sports were preferred. For the sexuality networks, seed websites were comprised of four sex education websites (two blogs and two sites) and six adult pornography websites (three blogs and three sites). Each sexuality seed was selected using Google©. The four sex education websites were the most popular websites (i.e., the websites that were first in the search results) using the search terms *sexuality* and *education*. The six adult pornography websites were selected using a variety of search terms in an attempt to acquire a broad spectrum of pornographic websites. The first search term used was *BDSM*[20] for which three seed websites were selected (one blog and two sites). The second search term used was *sex* for which an additional three seed websites were selected (two blogs and one site). Although the same data were collected on the comparison networks as with the CE networks, no website inclusion criteria specific to keywords or images were required. Sports networks began with an average of 301.40 (s.d. =1.65) websites and were re-crawled at an interval of 41.69 (s.d. =4.44) days while sexuality networks began with an average of 306.30 (s.d. =6.00) websites and were re-crawled at an interval of 41.62 (s.d. =7.71) days.

3.4.2. Web-Crawler Criteria (Keywords & Known C.E. Images)

The 82 keywords used by CENE to identify CE websites were selected from previous research and were found to be prevalent on CE websites (Latapy, Magnien, & Fournier, 2013; LeGrand, Guillaume, Latapy, & Magnien, 2009; Steel, 2009). The 82 keywords were grouped into three categories. The first were code words (27) commonly used by offenders to alert one another to material (e.g., pthc[21]). The second were thematic words (23) not directly linked to CE but typically present (e.g., boy, girl, child).

[18] http://labs.ebuzzing.com/top-blogs/sports
[19] http://www.marketingcharts.com/
[20] 'BDSM' stands for a) bondage and discipline; b) sadomasochism; and c) dominance and submission (Wiseman, 1996).
[21] Pthc is an acronym for the term preteen hardcore and is one of the most prevalent code words.

The third were sex-oriented words (32) referencing sexual organs or acts (e.g., pussy, cock, oral).

The presence of CE images was verified using a database provided by the RCMP. Last updated on June 1st, 2012, this database contained 2.25 million hash values[22] and was classified into three categories, in accordance to Canadian law. As a result, using any of the existing scales (eg., COPINE or SAP) was not feasible. According to section 163.1 (1) of the Canadian Criminal Code (CCC; 1985), *child pornography* includes any '…visual representation…written material…or audio recording' of a person under the age of eighteen, engaged in an explicit sexual act or advocates sexual activity'. Unlike the United States and Australia, the definition of child exploitation material is uniform across the country and, like the United Kingdom, includes real or imaginary/computer-generated visual representations (see Gillespie, 2011). The first database category (*Child Exploitation*) contained 618,632 images and were classified, under the CCC, as being CE. The second (*Child Nudity*) contained 652,223 images that would probably be considered CE by a judge. However images in this category were not blatant and, thus, the risk averse nature of law enforcement resulted in these images being placed into a separate category. The third (Collateral) contained 981,231 images that were important enough to be collected by offenders but would not be defined as CE, according to the CCC. For example, the initial images in a photo-shoot whereby a child was still clothed and not being directly sexually exploited.

3.4.3. Measures

Duration

This study referred to duration as the time between the beginning of data collection and the first, recordable, interruption. We defined an interruption as being a

[22] The database is a collection of hash values identifying CE images. A hash value is a 24-hexidecimal code which acts like a digital fingerprint for any file. If a file is edited, even minimally, a new hash value is created. Hardy and Kreston (2004) state that the probability of two files having the same hash value is 10^{38}.

data collection point where the website was not reachable (i.e., offline)[23]. While we assumed that some of the interruptions would be due to detection and removal by control agencies, it was impossible for us to determine the true nature of the interruption.

Each of the websites were active prior to the data collection. This means that the true start dates of our websites were unknown[24]. The start date of the observation period was the same for all websites included within a network. In addition, our data were right-censored because some of our websites remained active at the conclusion of our sixty-week observation period. Although we were unable to accurately determine the full criminal career duration, our primary focus was to examine the characteristics of websites that failed during the observation period and compare them to those that persisted. Duration varied between being active a single wave of data collection to being active for the full observation period. The relative "age" of a website was controlled for via various indicators of size, as discussed below.

Volume of illegal content

We measured volume as specifically pertaining to the count of all (27) code words, *child exploitation* images, and/or borderline images (*child nudity or collateral)* found on a website.

Crime-mix type

Focusing on specialization amongst websites, we created three composite variables that measured a website's content preferences. To control for the size of a website (i.e., the number of webpages comprising the website), each of our specialization variables were calculated at the "per webpage" level.

[23] While it is possible that an identified interruption could be short-lived and not actually reflective of a website going offline permanently, a six-month follow-up revealed that only 5.2% of our failed websites came back online after our identified interruption.
[24] We attempted to minimize the potential bias in survival estimates drawing on Cain et al.'s (2011) procedures for dealing with left truncation and censoring. Using these procedures resulted in either a) too much data loss, or b) an inability to apply semi-parametric methods (e.g., cox regression).

Sex of Victim (Boy/Girl): A website was determined to be boy or girl focused based on the relative frequency of keyword specific to boys (boy, son, twink, penis, and cock) in comparison to keywords specific to girls (girl, daughter, nymphet/nymphet, lolita/lolly/lola/lolli, vagina, and pussy). With the exception of two, CE networks were predominantly boy focused[25].

Severity (Explicit/Non-Explicit): The severity focus of a website was determined by the higher relative frequency of explicit or non-explicit keywords. The composite variable *explicit* consisted of 21 keywords related to severe sexual abuse (e.g., cries, torture, and rape). The composite variable *non-explicit* consisted of 15 keywords related to sexual characteristics (e.g., innocent, lover, smooth). Within the child exploitation networks 40.0% of websites were classified as being explicit.

Media (Video/Image/Story): For both *videos* and *images*, we used the average number of instances found of each medium on a webpage. For *stories,* we used the average number of our 82 keywords found per webpage. Stories may include vivid descriptions or comments attached to an image (or video) as the detail and excessive use of keywords would point to the written depiction being more central or indicative than the visual content. Each website was given a standardized score (0.00 to 1.00) for each type of media, relative to the other websites within the network. For example, the website with the highest average number of videos per webpage received a *videos* score of 1.00. A website's standardized score on each media type was then compared and a website was determined to be video, image, or story focused based on which measure they scored highest. Image focused websites were the most prevalent accounting for 83.7% of websites while 9.2% were video focused and 7.1% were story focused.

Connectivity

To measure the connectivity of a website within its own network we used the number of incoming and outgoing hyperlinks to/from other websites in the network. We simply recorded if two websites were connected, not the number of hyperlinks going

[25] These findings coincide with our seed websites as eight were identified as boy focused. However, within our dataset boy focused websites accounted for only 62.85% of the websites.

from one website to another. For example, if Website A hyperlinked to Website B four times it would count as one outgoing hyperlink for Website A and one incoming hyperlink for Website B. Child exploitation websites averaged 19.37 outgoing hyperlinks and 20.49 incoming hyperlinks.

3.4.4. Analytic Methods

To determine whether baseline survival rates for websites in CE networks differed from sports and sexuality websites, Kaplan-Meier estimates (KME) were calculated. KME was selected as it is adept at accounting for right-censored-data (Cleves, Gould, Gutierrez, & Marchenko, 2010). In our study, we were interested in the effects of website characteristics on subsequent failure. Consequently, we needed to merge our 10 CE networks into one database. The merging of datasets has the advantage of partially accounting for the selection biases inherent in selecting specific seeds; the networks of which may not be representative of other networks constructed from other seeds. Selecting multiple seeds and merging the datasets gives the analysis more power and less dependence on unconventional websites and their networks. However, such merging can be problematic if the datasets have heterogeneous properties. The most cited examples are randomized drug trials at different facilities, different treatment methods or selection criteria, and different data collection methods (Lijoi & Nipoti, 2014; Yasrebi, Sperisen, Praz, & Bucher, 2009). Given that the 10 networks were collected simultaneously, using the same tool and selection criteria we felt confident in merging the networks into one large dataset.

To ensure that each website was included only once, websites that appeared in multiple networks were removed. In addition, websites found to consist of only one webpage, with no content, were also removed. Their removal ensured that websites that had been replaced with a notification of removal, or similar notifications, were not included in the analyses. This resulted in our original sample of 3,051 being reduced to 1,580 unique websites. Using the sample of 1,580 websites, five proportional hazard (Cox) regression models were calculated. Our first model examined the effects of general website characteristics—webpages per website, images, videos, and keywords per webpage—while models two through four examined the effect of each criminal

career dimension: offending frequency (volume of illegal content), crime-type mix, and co-offending (connectivity). The fifth model was an amalgamation of the four previous models.

For each model we interacted each continuous variable with itself to test for a quadratic effect (Cleves et al., 2010). If the interaction term was significant it was included in the final model and the turning point was calculated. Assumptions of proportional hazard were assessed using Schoenfeld residuals (Grambsch & Therneau, 1994). Ties in failure were handled using the Efron approximation. This method was chosen over the Breslow approximation and the Kalbfleisch-Prentice approximation as the former tends to underestimate the continuous-time calculation while the latter overestimates (Hertz-Picciotto & Rockhill, 1997). As the failure of one website may have been influenced by the failure of another, the assumption of independent observations was violated. Therefore, robust standard errors were used to deal with potential clustering (Hoechle, 2007). Finally, goodness of model fit was determined two ways. First, Cox-Snell residuals for each model were compared to the estimated Nelson-Aalen cumulative hazard (Cox & Snell, 1968). Second, a Harrell's C concordance statistic was calculated (Harrell, Lee, & Mark, 1996).

3.5. Results

Do websites derived from legitimate seeds (our comparison groups) survive longer than websites derived from seeds hosting child exploitation (CE) material? Figure 3-1 displays the Kaplan-Meier estimates averaged for each genre and seed-type. Results show that websites within networks starting from a CE seed did not fail at a greater rate than those starting from a legal sexuality or sports seed. More specifically, Figure 3-1 shows that the survival curves of the sexuality seeds sample could not be distinguished from the ones obtained for the CE seeds. Further analyses suggest that illegal websites may even survive longer than others (Figure 3-2), although this may not hold true when examining only the CE sample at the multivariate level. Examining the survival rates of websites with any known CE images (n =80) from websites without any known CE images (n =1500), though directly or indirectly connected with a CE seed, Figure 1b shows that survival rates are significantly higher for websites with confirmed CE images from past police investigations (X^2 =8.86; p <0.01).

Figure 3-1: Comparing survival rates across genres by seed-type

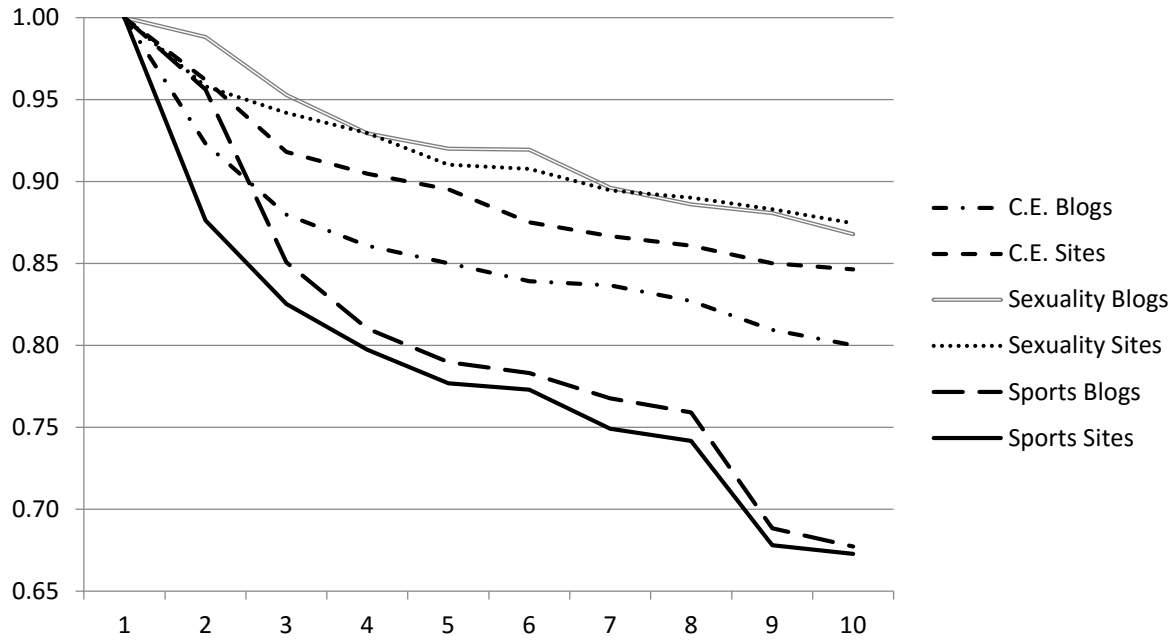
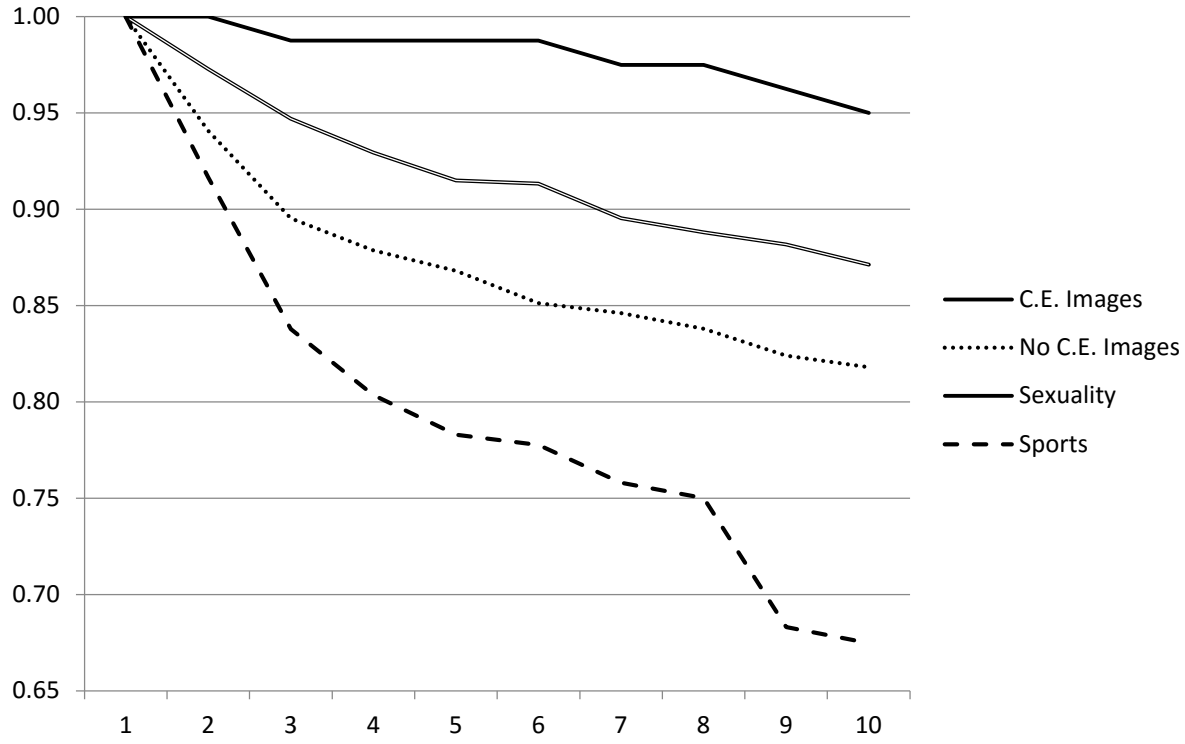


Figure 3-2: Comparing survival rates of CE network websites with and without known images to comparison networks



Selecting only the 1580 unique websites derived from the 10 CE seeds, Table 3-1 compares characteristics of websites that survived (n =1303) to those that failed (n =277)[26]. The most interesting finding was that website "hubs", as defined by general website characteristics, volume of content, and connectivity, were more likely to survive than websites that were smaller, had less CE material, and were less connected with the rest of the community. These results were not unlike what we would expect of the survival rates of legitimate businesses where newer, smaller businesses tend to have lower survival rates. The size of a website, in number of webpages, or images, also acted as a proxy for the age of the website. Older websites have had a longer period of time to publish content, establish links with the online community, and find a *raison d'être* that provides incentives for website owners to maintain the website as active.

[26] We also ran this analysis within the comparison networks, for non-CE measures. Sports websites that failed were smaller (fewer webpages) and had fewer images per webpage, outgoing, and incoming hyperlinks. Overall, no differences were found between sex websites that failed or survived.

Table 3-1: Chi-square analyses comparing website attributes of surviving to failing CE websites and early failing to late failing CE websites

	All websites		Failed websites	
	All Survived (se) n=1303	*All Failed (se) n=277*	*Early Failures (se) n=89*	*Late Failures (se) n=188*
Website Characteristics				
Webpages (Per Website)	1171.08** (6401.92)	114.90 (470.07)	15.35* (54.92)	162.03 (563.23)
Keywords (Per Webpage)	2021.58 (50956.86)	1051.20 (13177.10)	198.00 (185.30)	1454.64 (15978.52)
Videos (Per Webpage)	0.67 (3.76)	0.44 (2.62)	0.02 (0.12)	0.64 (3.17)
Images (Per Webpage)	39.86** (55.33)	16.94 (33.63)	11.88 (19.30)	19.33 (38.37)
Volume of Illegal Content				
Code Keywords	197.10 (3155.72)	33.45 (336.04)	14.98 (125.70)	42.20 (398.33)
C.E. Images	0.24 (4.17)	0.79 (13.01)	0.00 (0.00)	1.16 (15.78)
Borderline Images	0.06** (0.23)	0.01 (0.12)	0.00 (0.00)	0.02 (0.14)
Crime-Type Mix				
Boy-Focus	0.63 (0.48)	0.66 (0.47)	0.71 (0.46)	0.64 (0.48)
Explicit-Focus	0.43** (0.50)	0.54 (0.50)	0.72** (0.45)	0.46 (0.50)
Media Focus: Image	0.90** (0.30)	0.75 (0.43)	0.71 (0.45)	0.77 (0.42)
Media Focus: Video	0.06 (0.23)	0.06 (0.23)	0.01* (0.11)	0.08 (0.27)
Media Focus: Story	0.04** (0.21)	0.19 (0.39)	0.28** (0.45)	0.15 (0.36)
Connectivity				
Outgoing Hyperlinks	16.50** (27.95)	8.80 (14.10)	7.45 (7.91)	9.44 (16.19)
Incoming Hyperlinks	12.71** (10.61)	10.31 (9.41)	9.53 (9.70)	10.68 (9.71)

*p<0.05 **p<0.01

A common occurrence of survival designs, where entities are already active at the start of the study, is that a disproportionate number of failures will occur early in the observation period due to the capturing of a more diverse sample. Early website failures included websites that were bound to fail rapidly (e.g. true early failures) as well as those failing from natural attrition (e.g. websites that were active for a period prior to the start of the study, but happened to fail during the early waves). Of the 277 failures that occurred during the study period, 89 (32.13%) failed between Wave 1 and Wave 2[27]. While this early attrition was expected, we still needed to examine whether early failures had different characteristics than subsequent failures. Table 3-2 examines the characteristics of these two groups and shows that very little can help differentiate them. We found that four of 14 characteristics were shown to significantly differentiate these two groups. Early failures were smaller, less likely to be focused on videos, but more likely to be explicit focused, or story focused. These differences suggest that early failures were also younger websites, something that would not be unexpected for any new projects or businesses. That so many early failures were focused on presenting explicit content or code words is perhaps indicative of the vulnerability of these websites to detection. Despite the similarities, the disproportionate rate of failure between Wave 1 and Wave 2 resulted in our Cox regression models violating the assumption of proportional hazard. To address this issue we controlled for the effect of failure during Wave 2 in the multivariate models presented below.

3.5.1. Predicting Time to Failure

Table 3-2 presents the results of the Cox regression models predicting time to failure for the 1,580 websites in our sample (hazard ratios are shown). Model 1 examines general website characteristics: number of webpages on a given website and average number of images, videos, and keywords per webpage[28]. Our results show that

[27] Within the sports and sex networks, we also compared early and late failures. Amongst sports websites, early failures had three times more videos per webpage. Amongst sexuality websites, early failures had *more* webpages per website but fewer images per website.
[28] All continuous variables were tested for quadratic properties. Each model began with all linear and quadratic effects. Quadratics found not to be significant were removed, one at a time, until the final model was determined. This final model was reported in Table 2 and any quadratic effects were noted.

Table 3-2: Proportional hazard regression models of time to first website failure

		Model 1[a] n=1580	Model 2[a] n=1580	Model 3[a] n=1565	Model 4[a] n=1580	Model 5[a] n=1565
	Webpages (Per Website)	1.000[c]*** (0.000)	-	-	-	1.000[c]** (0.000)
	Keywords (Per Webpage)	1.000 (0.000)	-	-	-	
	Videos (Per Webpage)	1.009 (0.019)	-	-	-	
	Images[b] (Per Webpage)	0.982*** (0.003)	-	-	-	
	Volume of Illegal Content					
	Code Keywords	-	1.000 (0.000)	-	-	1.000[d]* (0.000)
	C.E. Images	-	1.013** (0.005)	-	-	1.014*** (0.004)
	Borderline Images	-	0.358** (0.180)	-	-	0.637 (0.369)
	Crime-Type Mix					
	Girl-Focus	-	-	0.925 (0.100)	-	0.961 (0.105)
	Non-Explicit Focus	-	-	0.944 (0.102)	-	1.007 (0.108)
Media Focus	Image	-	-	REF	-	REF
	Video	-	-	1.407 (0.352)	-	1.402 (0.357)
	Story	-	-	1.833*** (0.251)	-	1.687*** (0.221)
	Connectivity					
	Outgoing Hyperlinks	-	-	-	0.980*** (0.008)	0.986** (0.007)
	Incoming Hyperlinks	-	-	-	0.987** (0.006)	0.988** (0.006)
	Outgoing x Incoming	-	-	-	1.000 (0.000)	1.000 (0.000)
	Harrell's C	0.804	0.723	0.725	0.723	0.769

*p<0.10 **p<0.05 ***p<0.01 REF=reference category

a: Early failure (between Wave 1 and Wave 2) was controlled for to better assess the effect of our predictors.

b: images per webpage was the only continuous variable found to function as a quadratic. Therefore, Model 1 included a quadratic effect for images per webpage. The hazard ratio was 1.000036 p>0.001.

c: in Models 1 and 5, the non-rounded hazard ratio of Webpages per website were 0.99957 and 0.99964 .

d: in Model 5, the non-rounded hazard ratio of code words was 1.00008

large and image focused websites had higher survival rates (0.04% and 1.88% for each additional webpage or image). However, the effect of images on survival was not linear as each addition image per webpage above 256 resulted in a 0.004% ($p<0.01$) decrease in survival rates. We also ran a Cox regression (not shown) within each comparison genre, analysing the effects of non-CE measures on survival. Controlling for early failure[29], results were similar within sports and sexuality networks. Survival was increased for each additional webpage and/or image (per webpage) present.

The next three models each introduced one of the three (remaining) criminal career dimensions. Model 2 removed the general characteristics and introduced our indicators for variations in the volume of illegal content. Results suggest that volume of illegal content matters. Model 2 (Table 3-3) showed that for each additional known *CE image*, odds of survival decreased by 1.27%. Interestingly, having grey area images (i.e. *child nudity* or *collateral*) were related to persistence. Crime-type mix indicators introduced in Model 3 were not shown to be as important in predicting time to failure. Neither the *sex of the victim* nor the relative *severity* were shown to be associated with failure. The type of media, however, emerged as a significant factor. Websites focused on *stories* had increased odds of failure compared to websites focused on *images*.

The importance of the location of a website in the online community is evidenced in Model 4 (Table 3-3), which examined the effect of *outgoing* and *incoming* hyperlinks on time to failure. The additional visibility created by a website reaching out to other websites (*outgoing*) and becoming popular (*incoming*) appears to outweigh the potential increased risk of detection from this same visibility[30]. For each additional outgoing hyperlink, odds of survival increased by 2.04%. For incoming hyperlink, odds of survival

[29] To satisfy the requirement of proportional hazard, 'early failure' within sports networks was prior to Wave 3.

[30] Amongst comparison networks, sexuality website survival was increased by 4.03% and 2.24% for each outgoing and incoming hyperlink. For sports, each outgoing hyperlink increased survival by 0.59%

increased by 1.32%. However, their effects were independent as their interaction was not significant.

While the individual influence of general characteristics and criminal career dimensions on website persistence is important, the presence of multiple website characteristics may provide a better explanation for survival. Model 5 (Table 3-3) simultaneously examined the effects of webpages, volume of illegal content, crime-type mix, and connectivity. There were little changes to the main results: small, story focused websites as well as those with CE images and fewer connections were shown to fail more quickly. Two changes from previous models were worth noting. First, although not significant in Model 2, the volume of code words became significant in the full model, with each additional code word resulting in a 0.008% decrease in survival. Second, the result from Model 2 that *child nudity* and *collateral* images increased survival no longer held, when controlling for the other relevant factors in Model 5. In the end, the volume of illegal content (*CE images)* appeared to matter most in predicting website failure in our sample.

3.6. Discussion

The fact that an increasing number of crimes are committed in cyberspace, or use it as a tool for committing crimes, is starting to get the attention of mainstream criminology. Beyond the sheer size of the phenomenon (in costs to society, and number victims and offenders), one reason for the growing attention to cybercrime is because of the impact it has on offline, traditional crime. Closer to our concerns, the growing use of the Internet has created an online support community for certain types of offenders, like child molesters, where none existed before (Tremblay, 2006). An important task of scholars will be to test whether the existing conceptual frameworks and theories can be used to study crimes and criminals in cyberspace. Another is to develop new conceptual and methodological tools to account for the unique elements of cybercrimes and apply them to situational crime prevention strategies (Clarke, 1997; Newman & Clarke, 2003).

The current study attempted to provide a contribution towards both. Conceptually, we began transitioning the criminal career paradigm to cyberspace, more

specifically in the context of websites hosting illegal, child exploitation (CE) material. We focused on the dimensions of offending frequency, seriousness/crime-type mix, and co-offending and developed indicators that would tap into these dimensions, for this type of crime. Methodologically, we developed a repeated measures design around a web-crawling tool that was uniquely adapted to the fact that we were studying a relatively new unit of analysis (i.e. the illegal website), and was embedded in an online network of other websites, connected through hyperlinks; unique tools for unique data.

The main objective of this study was to examine the factors associated with failure of websites hosting CE material and the community of other websites surrounding these illegal websites. A secondary objective was to compare the rates of survival of our sample to comparison groups consisting of websites associated to 10 sports seed websites, and 10 sexuality seed websites. We discuss the two main findings below.

First, we found variations in the survival rates, based on the nature of the seed from which the networks were constructed. It comes as no surprise that websites surrounding, or involved in, illegal activities have different baseline survival rates than legal websites. However, what is surprising is that it was survival amongst the confirmed illegal websites that was found to be longer (Figure 3-2). While the high demand for CE material coupled with better detection avoidance tactics for these websites may partly explain this finding, the fact that we conducted this study retrieving only “open websites”, available to the public, implies inherent limits in the ability or efforts invested in avoiding detection for these websites. As many websites, legal or not, are quickly abandoned because of the time commitment, resources required, and/or an inability to attract users (Hsu & Lin, 2008; Urboniene, 2014), it may be that the demand for illegal content motivates an operator to maintain their website longer. Put another way, compared to sexuality and sports, where supply outpaces demand and, thus, new websites face higher competition from established websites in the genres, CE websites may not be as concerned with competition given the inherent fluctuation of supply. As a result, the importance of accessibility, speed, navigation, and content quality may be weighted differently for illegal websites compared to legal websites. This conclusion is supported by the findings of our multivariate analyses. Namely that larger and more popular

websites persisted (Table 3-3) and that early failures were predominantly comprised of small websites with little to no content (Table 3-2).

Second, a series of findings emerged from the Cox regression analyses focused on the websites that were part of the social structure surrounding our 10 CE seeds. Here, two of the three website-specific criminal career dimensions emerged as important predictors: volume of illegal content and connectivity of the website. Both the volume of illegal content and connectivity tap into the notion of online "visibility" that could potentially play a role in detection. Yet, the former dimension is associated with failure while the latter is associated with survival. There are at least two potential angles of interpretation for these findings. One, popular websites do not necessarily have a high volume of known illegal content. In fact, popularity may stem from the novelty of the material found on a website as opposed to older material, albeit material confirmed to be CE. Two, a rational choice perspective suggests that popularity may breed motivation for owners to persist with a website. The more visitors, the more successful, the more incentives to persist and publish new material, or simply keep it online. This interpretation is supported by research into other cybercrimes that emphasizes the importance of connectivity and relationships in persistence (Decary-Hetu & Dupont, 2012; Holt, 2007). Recall that website failures are not solely due to the work of control agencies. Therefore, while survival and failure of some website entities may be economically driven, a website's ability to integrate into the larger community and become a congregation point that provides acceptance (Bossler & Burruss, 2011) and social and criminal resources for members may play an equally important role in survival and failure (see Tremblay, 2002).

Next, the increased risks of failure associated with higher volumes of CE material suggest a potential role for law enforcement activities in detecting some of the websites we followed that went offline. After all, including material that the police have already classified as illegal naturally increases the vulnerability of a website to detection. However, we believe that many website owners take this risk unknowingly as it is unlikely that they are aware of their website containing (some) images classified by the police as illegal. Beyond law enforcement agencies, our findings also suggest a role for hosting companies in removing some of the illegal websites they encounter. For

example, the lower baseline survival for websites within blog seed networks (Figure 1a) could be suggestive of hosting companies playing a role in removing offending websites. Many personal blogs are hosted by larger companies such as Blogger© or LiveJournal©. When a user creates a blog, they agree to a terms of service (TOS). Most TOS include clauses about disseminating illegal content, including copyrighted material. If a blog is found to be in violation of the TOS, the hosting company can immediately remove the blog. Conversely, websites that are hosted independently are more difficult to remove as control agencies must identify the country the website is hosted, the laws of the hosting country, the TOS of the hosting company, and the cooperation of the host. These additional hurdles may result in fewer efforts, or abilities, being taken to shutdown sites in comparison to blogs. Despite the hard work potentially being conducted by blog hosting companies to remove illegal websites, the ease of setting up a blog, coupled with the anonymity afforded by the Internet, point to the removal of websites disseminating CE material as being a cyclical process. That is, there is very little preventing a user from creating a new blog, on the same or different service, every time their previous blog is removed.

From a policy perspective, our findings highlight the need for additional tools combating CE material. For example, the similarity in failure rates for video focused websites emphasizes the lack of a CE video database, similar to existing image databases. Currently there are no reliable techniques for detecting CE video content. While the dissemination of images remains the primary form of CE material, increases in bandwidth speeds, coupled with the growth in personal video equipment (e.g., webcams), point to video content becoming more prevalent. One possible option is digital video fingerprinting. This technique allows for the matching of similar videos through unique features; however, it has limitations as it is open to user-collusion and requires the original video, to ensure that all edited videos share the unique features (for further discussion see Kumar & Kaliyaperumal, 2012). By creating additional tools and databases, the efficiency and accuracy of removal can be increased and the responsibility of removing such material can be shared. As a result of shared responsibility, society can more effectively address the five situational crime prevention strategies outlined by Wortley and Smallbone (2012) for combating CE distribution: increase effort, increase risk, reduce rewards, remove excuses, and reduce provocation.

3.6.1. Limitations

Our study was subject to four primary limitations. First, much like offline networks, which include both offenders and non-offenders (e.g. Haynie, 2002), it is likely that contained within our CE networks were websites that did not disseminate CEM (i.e., false positives). However, their connection to websites hosting illegal material made them part of the community of websites from which users can access this content, via hyperlinks or keyword searches. We believe that including the networked community of websites in such studies preserves the reality of the seed websites existing within a larger online social structure that is not perfectly homogenous. Our design, however, limits the interpretation of our findings to websites within the community of 10 illegal seed websites, as opposed to simply “illegal websites”.

Second, because we only measured websites every six weeks, we do not know the precise date of failure. Rather, we have an interval during which the website failed. It is a common limitation of survival studies where the occurrence of a condition is only learned at a follow-up examination or assessment. The Efron approximation method we used performs well in these conditions and, overall, it is not expected that this limitation significantly alters the substance of our findings.

Third, although Wolak et al. (2005) found that online child exploiters use few security measures, and subsequent research into other cybercrimes has found that offenders ignore warnings (Maimon et al. 2014), those that do implement some type of security measure are probably more likely to survive. Nevertheless, our study findings should be viewed as a representation of website with minimal security measures, with CEM often openly advertised.

Fourth, our study was conducted on the *Surface Web*. The Surface Web is the portion of the World Wide Web (WWW) that is indexed by search engines and is the most readily accessible to the average user. The majority of material is found in the

Deep Web[31], the bottom layer of the WWW. The Deep Web is comprised of dynamic webpages and is not indexed. As a result, this is where many of those wishing to keep their activities hidden reside. While future research needs to examine illegal websites on the Deep Web, our focus on the Surface Web can be seen as the starting point for many first-time offenders and where new offenders acquire knowledge and skills. Thus, it remains paramount to understand what is available easily and publically.

3.6.2. Future Research

The findings of this study identify at least three important areas for future research. First, the importance of connectivity between websites was a key contributor to increased persistence. As such, subsequent research needs to examine more in-depth the role of connectivity in the criminal career of websites. More specifically, whether subsequent desistance can be traced along website connections and whether failures are clustered within smaller sub-communities. In addition, examinations of community effects may lead to a better understanding of how material is distributed throughout a network. Finally, how the connections of the website operator(s) may influence the embeddedness of the website through operators' management of multiple websites.

Second, researchers have noted that groups of offenders follow different life course trajectories and that these trajectories coincide with offender characteristics, life events, and types of crime. For offline sexual offending several trajectory models have been suggested (Lussier, Tzoumakis, Cale, & Amirault, 2010). Therefore future research needs to examine, generally, whether the criminal career of websites can be categorized into specific trajectories and, specifically, whether existing trajectory models can be translated to cybercrime. More generally, future work is needed to conceptualize and measure the trajectories of websites hosting illegal content other than child CEM.

Third, the geographical location of the material and its' novelty may contribute to longevity or premature failure. Subsequent research needs to identify the country

[31] Comprised within the Deep Web is the more recognized Dark Web. The Deep Web refers to the entire system while the Dark Web refers to the illegal activities that are conducted in the Deep Web.

hosting the material in order to control for the corresponding laws and the difficulties of international law enforcement operations. Additionally, the importance of volume of content to website survival may be mediated by the novelty of content. The inclusion of *new* material may increase a website's popularity, in turn increasing longevity, while old material may carry little weight in the online CE community and may be easier for control agencies to detect and remove. However, it is also possible that the inclusion of new material may lead to premature failure due to law enforcement priorities focused on rescuing children currently being abused. Including a measure that controls for the newness of material, such as the date of its inclusion into the hash value database, may provide valuable insight into website survival.

3.7. Conclusion

Drawing from a repeated measures design that followed, for sixty-weeks, over 1500 unique websites connected directly and indirectly to 10 illegal ones**,** this study finds that the volume of illegal content on a CE website is associated with increased risks of failure. At the same time, the study also found that such websites are no more, no less likely to fail than websites found in legitimate adult sexuality communities, or sports websites communities. The paper also provides a framework for future research transitioning the criminal career paradigm to cyberspace, and websites specifically. We provide evidence of website characteristics that can be linked to three dimensions of the criminal career paradigm. The growth in cybercrime over the past decade has made it paramount that existing criminological theories and frameworks be either adapted and revised for online contexts, or simply rejected as fruitful theories to adopt for cybercrime. Novel theoretical developments are also needed, and should catch up on the methodological advances made by recent research on online criminality.

Despite the challenges of research in online settings, cybercrime research also provides unique opportunities for innovations in research designs and contributions to the field as a whole. For example, websites and online discussion forums with illegal material emerged as a new object of criminological inquiry providing unique insights into illegal markets operating online (Holt, 2012; 2013), how online subcultures around deviant interests form, evolve, and disappear (Decary-Hetu & Dupont, 2012; Holt, 2007;

Jordan & Taylor, 1998), how the logic deterrence may apply online (Maimon et al., 2014), or how the existence of the Web changes the practices of criminal networks and groups such as street gangs (Moule, Pyrooz, & Decker, 2014; Pyrooz, Decker, & Moule, 2015). In some cases, like ours, longitudinal data can be collected in real-time, as opposed to retroactively. The design we adopted only scratched the surface of the array of meaningful innovations that can be adopted by researchers as we try to not fall too far behind in our understanding of cybercrimes and criminals.

Chapter 4. Liking and Hyperlinking: Community Detection in Online Child Exploitation Networks[32]

4.1. Introduction

Research on the sociology of the Internet has shown that online communities matter; perhaps just as much as offline communities and social interactions. Previously seen as small and bound by a neighbourhood or village, the minimization of communication barriers afforded by the Internet (Rheingold, 2006) have facilitated the transition of communities into large, global, social networks (Wellman, Boase, and Chen, 2002). This transition significantly impacted how people with similar interests connected. For those conducting illicit activities, the Internet opened up a new avenue for conducting business and targeting victims. For distributors and consumers of child pornography, the Internet transitioned a traditionally solitary crime into a globally communal crime, where material could be accessed instantaneously (Hillman, Hooper, and Choo, 2014). Previously isolated and facing constant challenges to find new material, the Internet provided consumers with the opportunity to connect across long distances with like-minded individuals, who provided moral validation, social support, and access to a constant stream of new material, often free of charge (Beech, et al., 2008; Estes, 2001; Quayle and Taylor, 2011; Taylor and Quayle, 2003; Tremblay, 2006). Central to the opportunities and support provided online are the networks that form between individuals *and* entities (e.g., websites), and the communities that develop from these networks. As a result, in this paper we take a unique approach to understanding communities by focusing on the larger, macro-level, networks created between website entities, via hyperlinks. Specifically, we examine the structure and long-term functionality of CE-related website communities.

[32] This manuscript is currently under review at *Social Science Research*

The online distribution of child sexual exploitation (CE) material, depicting the sexual abuse of children, is conducted using a variety of public and private platforms. Central to the online distribution of CE material are the producers and distributors. However, within the World Wide Web (WWW), the practice of distribution is heavily influenced by the websites that host the material. Although initially uploaded by individuals, consumers do not directly access the supplier. Rather, they rely on the intermediary website to acquire material, and use said website to gain access to other websites with similar (or different) material. Therefore, the process of CE distribution, within the WWW, cannot be fully understood without considering the role websites play, as they provide the initial means for accessing distributors and like-minded offenders. In fact, websites may play an even greater than suspected role in distribution, as they can dictate, to some degree, who and what individual consumers have access to, via their connections (hyperlinks) to other suppliers (i.e., websites). Therefore, the analysis of websites is central to understanding CE distribution within the WWW and cyberspace. More specifically, the analysis of the structure of the communities websites form, how these communities function, and how they evolve over time can have implications for the development of social policies, the methods used by researchers to study online child sexual exploitation, the strategies used by law enforcement agencies, and the role of public and private companies in detection and removal.

In this study, we examine the communities that form around public websites involved in the distribution of CE material. Although inclusion criteria are used, it is worth nothing that not all websites we examine are actively distributing CE images. Instead, we study the communities *surrounding* 10 known child sexual exploitation websites and thus characterize the networks as related to them. Connected via hyperlinks, we explore the social structure of website communities, how these communities are formed, and how they evolve and adapt over time. A supplementary objective of this study is to introduce the Child Exploitation Network Extractor (CENE), an automated data collection tool that can be useful in increasing the efficiency of data collection processes for online research on social problems, and studying graphic and/or traumatizing topics (e.g., child sexual abuse). In this study, CENE allowed us to process automatically the content and hyperlink information of 4,831,050 webpages.

4.2. Literature

4.2.1. Online Communities and Illegal Websites

Website communities are formed through hyperlinks connecting websites directly to one another. Hyperlink network analysis (Park & Thelwall, 2003) has shown that hyperlinking practices are purposive, following underlying communicative rationale and providing opportunities to create and foster alliances (Foot et al., 2003; Park, Kim, and Barnett, 2004; Park, Thelwall, and Kluver, 2005). Hyperlinked websites often share ideological similarities and are rarely connected directly to dissimilar websites (Ackland and Shorish, 2009; Adamic and Glance, 2005; Hargittai, Gallo, and Kane, 2008). As a result, hyperlinking can be viewed as a tool used to create smaller communities, within a larger social network, and provide website visitors access to like-minded others and targeted content. In social network terms, the action of hyperlinking is a recognition of the existence of another, important (or strategic) enough to warrant a public display whereby visitors of the first website can easily transfer to the second.

A central component of the World Wide Web, websites have been appropriated by offenders in three important ways. First, social networking websites have been exploited by sexual offenders for grooming victims (Whittle, Hamilton-Giachritsis, and Beech, 2014a; 2014b; Whittle, et al., 2013), while e-commerce websites are used to acquire financial information (Holt and Turner, 2012; Pratt, Holtfreter, and Reisig, 2010). Second, criminal websites have been created to provide offenders with *convergent settings* (Felson, 2003) where they can ply their trade (Decary-Hetu and Dupont, 2012) or acquire social support (Tremblay, 2006; Wortley and Smallbone, 2012). Third, websites are developed for the primary purpose of distributing illegal goods and services (Dolliver, 2015; Steinmetz and Tunnell, 2013; van Hout and Bingham, 2013) and promoting illicit activities (Milrod and Weitzer, 2012; Chow-White, 2006).

Criminal-based websites have a tendency to operate in isolation from mainstream and dissimilar illegal websites, preferring to form small communities with like-minded others (Burris, Smith, and Strahm, 2000; Chau and Xu, 2008; Zhou, et al., 2005). Research into online criminal communities has found that even for solitary

crimes, such as hacking, there exist strong ties whereby tools and information are actively shared across intricate network connections (Holt, 2009). Within these website communities offenders acquire criminal capital (prestige, notoriety, and status) through the sharing of tools and information with other offenders (Decary-Hétu, Morselli, and Leman-Langlois, 2012; Dupont, 2013; Holt, 2007). In acquiring criminal capital, offenders are more able to form friendships with others that may later translate into co-offending opportunities (Holt, Blevins, and Burkert, 2010; McCarty and Hagan, 1995). Subsequently, websites may play an important role in the co-offending selection process given they function to facilitate offender transactions by creating points of congregation. As a result, websites can been characterized as also acquiring criminal capital, through being seen as a place to acquire tools and information, and to connect with other offenders. By hyperlinking to homogeneous others, the websites create the boundaries of the community, facilitating and controlling some criminal opportunities afforded to offenders.

Among offline offender, those who subsequently co-offend often share homogeneous personal attributes, such as age, sex, ethnicity, residency, or criminal experience (Carrington, 2014; Schaefer, 2012; van Mastrigt and Farrington, 2011; Weerman, 2003). However, research has been divided as to whether this proclivity towards homogeneity is the result of preference (Reiss and Farrington, 1991) or structural opportunities (Carrington, 2002). Recently, van Mastrigt and Carrington (2014) found that homogeneous tendencies are not strictly the result of opportunity structures (e.g., more young men involved in crime) but rather that offenders also make conscious choices to select co-offenders similar to themselves.

It is unclear whether proclivities towards homogeneity translate to cyberspace. This is because the structure of the Internet means that co-offenders rarely interact in-person, instead relying on more secure digital communication methods (Decary-Hetu and Dupont, 2012). As a result, personal characteristics such as age, sex, ethnicity, residency, or experience are not readily available –and may not actually matter. Instead, online partnerships may be formed from broad interests in specific types of crimes, such as racism (e.g., white supremacists) and extremism, or, like between some offline gang members, facilitated by joint memberships to the same website.

Given that criminal-based websites hyperlink to similar websites (e.g., Burris, Smith, and Strahm, 2000) and offline criminal enterprises co-offend more along functional ties than ethnic ties (Malm, Bichler, and Nash, 2011), it follows that the first step in studying CE-related website communities is to identify the attributes that form the foundation for hyperlinking. As websites are innate, they do not have any personal characteristics that can be used to promote connectivity. However, CE-related websites can be characterized by their homogeneity in sex of victim (boy/girl), type of content (explicit/non-explicit), or medium of distribution (images/videos/text). Although, it is also possible that CE-related communities are solely based on broad interests and that connectivity between websites is more akin to a pure availability model.

There is support for the hypothesis that illicit website communities may function closer to a pure availability model. Within large piracy communities, websites, such as The Pirate Bay, appear indiscriminate with regards to the type of content they provide or the websites they associate. Even those who provide specific types of content, such as television shows (e.g., EZTV), do not appear to function outside the larger piracy network, or specifically among a community of television-based websites. However, within smaller Warez-release communities (Basamanowicz and Bouchard, 2011) and CE communities (Wortley and Smallbone, 2006), the connectivity selection process appears to be more discriminate. In the distribution of CE-related material, public websites may try to maximize embeddedness within the larger network by creating communities with any website, irrespective of content. A third possibility is that they congregate with dissimilar websites to avoid direct competition and to maximize general appeal, increasing overall consumer traffic.

4.3. Current Study

The distribution of CE-related material in cyberspace is conducted using a variety of methods operating strictly through online means or through a combination of offline and online tactics (van Wijk, Nieuwenhuis, and Smeltink, 2009). While a large percentage of CE-related material is distributed through Internet Chat Relay, the Deep Web, and other private networks, Carr (2004) identified the World Wide Web –which includes blogs, discussion forums, live webcam feeds, and photo galleries- as the

second most prominent source for obtaining CE images. Supporting Carr's assertion, recent research suggests that a substantial amount is still distributed through public, or semi-public, methods via peer-to-peer networks (Latapy, Magnien, and Fournier, 2013; Steel, 2009; Wolak, Liberatore, and Levine, 2014), BitTorrent networks (Rutgaizer, et al., 2012) and public websites (Westlake, Bouchard, and Frank, 2011). .

While parallels can be drawn between virtual and non-virtual communities, we propose that virtual communities are distinct enough to warrant separate conceptual and empirical treatment. One key reason for this distinction is that virtual communities do not require direct contact between members nor do they even directly require people for functionality. Instead, websites connect, through hyperlinks, with one another to form virtual 'entity-based' communities. Through these hyperlinks, public CE-related websites form communities, through hyperlinks, that facilitate the transmission of content, information, and offender connections, criminal opportunities, belonging, and validation.

In the current study, we investigate the communities that form within 10 networks of 300+ websites each, 4,831,050 webpages total, beginning with a known child sexual exploitation website. Using a custom designed web-crawler, we monitor these networks for 60 weeks, allowing us to study how the communities formed change over time. Drawing on community detection techniques, we a) describe the website characteristics that comprise and differentiate communities within the overall network; b) identify the stability of virtual communities over time; and c) whether hyperlinking tendencies are driven by homogeneous characteristics.

4.4. Methods

4.4.1. Data

The 10 networks and 4.8 million plus webpages analyzed for this study come from a longitudinal project with the objective to analyze the evolution of the websites involved in the dissemination of child sexual exploitation (CE) material. Using a repeated measures design, we analyzed the networks ten times at an interval of 42 days (six weeks). In order to manage each network's size and relevance, we implemented three

conditions. First, we created an exclusion list of popular websites (e.g., Facebook©) and false positives (e.g., Disney©) from previous CENE searches, known not to be directly associated with CE material. Second, we limited the size of the networks to ~300 websites and ~500,000 webpages. Third, an image had to be larger than 150 pixels by 150 pixels; or two inches (four centimeters) by two inches. Once the network size limitation were met, the data was aggregated up to the server level. In other words, the data on www.website.com/webpage1 and www.website.com/webpage2 were summed and listed under www.website.com. A list of all websites and webpages scanned were stored and reused at each data collection interval to ensure that the same webpages, if they were still online, were analyzed at each interval.

Data were collected using a custom-designed web-crawler that followed a snowball sampling method via hyperlinks between websites (Burris, Smith, and Strahm, 2000; Chau and Xu, 2007; Frank, Westlake, and Bouchard, 2010; Fu, Abbasi, and Chen, 2010; Zhou, et al., 2005). Referred to as the Child Exploitation Network Extractor (CENE), the webcrawler designed for this study operated similarly to automated data collection tools used by various search engines, to index websites. CENE followed a method similar to that of a person browsing the Internet, looking for CE material. Beginning with a *seed* website, known to be related to CE material, CENE scanned the hypertext markup language (HTML) for our pre-determined CE-identifying criteria. If the webpage was determined relevant, CENE continued to scan the website, collecting structural and website characteristics data. Similar to a person viewing the website, CENE then followed hyperlinks found on the website, to other websites, repeating the criteria search process. If the hyperlinked webpage was deemed irrelevant (i.e., did not meet the criteria), the website was discarded and CENE moved on to the next. A website was deemed to be CE-related if the webpage contained at least one known CE image, from a database of images provided by the Royal Canadian Mounted Police (RCMP), or at least seven keywords, from a list of 82, relevant to child sexual abuse.

Seed Websites

Each network began with a seed website selected from one of two sources. The first source, accounting for four seeds, was a list of websites, provided by the RCMP, known to distribute CE-related material. The second source, was a list of websites

identified in previous CENE searches to be engaged in CE-related dissemination. This included, but was not limited to, image or video distribution or acting as an access point to CE websites. To partially account for the nature of the seed biasing the characteristics of the sample derived, five of our ten seeds were *blogs* while the other five were *sites*. A blog was defined as a website with user-generated posts in a traditional web-log setup. A site was defined as a website with interlocking webpages, which did not meet the criteria of a blog, including discussion forums and photo galleries. Each network began with an average of 305.10 (s.d. =2.33) websites and was re-crawled every 42.14 (s.d. =4.45) days.

Website Inclusion Criteria

Keywords: The 82 keywords used by CENE were found to be the most prevalent in online CE dissemination networks (Hurley et al., 2013; Latapy, Magnien, and Fournier, 2013; LeGrand, et al., 2009; Steel, 2009; Vehovar, et al., 2009) and categorized into three groups. The first group were *code keywords* (27) commonly used by offenders to alert one another to content, such as pthc (pre-teen hardcore). The second group were *thematic keywords* (23) not directly linked to child sexual abuse but typically present on such websites (e.g., boy, girl, child). The third group were *sex-oriented keywords* (32) referencing sexual organs and acts (e.g., pussy, cock, oral).

Child Exploitation Images: CE images were identified using a hash value database provided by the RCMP. A hash value is a 32-hexidecimal code that functions similar to a digital fingerprint. Each computer file is given a hash value based on its binary composition. When a file is edited, even minimally, a new hash value is created. Tretyakov, and colleagues (2013) state that the chances of two distinct files having the same hash value is 'negligibly small' ($1/2^{2048}$). Last updated on June 1st, 2012 (CENE was launched in July 2012), the database contained 2.25 million hash values classified into three groups. According to section 163.1 (1) of the Canadian Criminal Code (CCC, 1985), child pornography includes "… any written material, visual representation, or audio recording" of a person "under the age of eighteen years and is engaged in or is depicted as engaged in explicit sexual activity". Unlike the United States and Australia, the definition of child pornography is uniform across the country and, like the United Kingdom, includes real or computer-generated visual representations (see Gillespie,

2012). The first group (*Child Exploitation)* contained 618,632 images that, according to the CCC, met the definition of child pornography. The second group (*Child Nudity)* contained 652,223 images that would probably be considered child pornography but have not yet been brought before a judge. The third group (*Collateral)* contained 981,232 images that do not meet the definition but were important enough to be collected by offenders. Images in this category may have included initial photographs taken by an offender, of a child, prior to the removal of clothing.

4.4.2. Website Composite Variables

During data collection, CENE identified the total number of webpages, images, videos, keywords, and incoming and outgoing hyperlinks for each website. From these we created composite variables to describe the characteristics of websites and the communities they form.

Type of Focus

Sex (Boy/Girl): Using the relative frequency of specific keywords, we classified each website as being *boy* or *girl* oriented. The keywords used for these classifications were: boy, son, twink, penis, and cock, or girl, daughter, nimphets/nymphets, Lolita/lola/lolli/lolly, vagina, and pussy.

Content (Explicit/Non-Explicit): Using the relative frequency of specific keywords, we classified each website as being focused on *explicit* or *non-explicit* material. The composite *explicit* measure consisted of 21 keywords related to severe sexual abuse (e.g., cries, torture, and rape). The composite *non-explicit* measure consisted of 15 keywords related to personal characteristics (e.g., innocent, lover, smooth).

Medium (Image/Video/Story): Using the relative frequency of three types of media found on websites, we classified each website as primarily distributing images, videos, or stories. First, for *image* and *video,* we used the average number of instances of each medium found on a webpage. For *story*, we used the average number of our 82 keywords found on a webpage. Second, each website was given a standardized score (0.00 to 1.00) for each medium, relative to the other websites within the network. For

example, the website with the most images per webpage was given a score of 1.00 while every other website was standardized against this website, on the same measure. For whichever medium a website received the highest score was its classification.

Connectivity

Incoming/Outgoing Hyperlinks: A website's connectivity was determined through the number of (unique) incoming and outgoing hyperlinks found. Whether Website A hyperlinked to Website B eight times or two times, the connections counted as one outgoing hyperlink for Website A and one incoming hyperlink for Website B. *Incoming* can be viewed as a measure of popularity while *Outgoing* a measure of a website's attempt to reach out to the community and integrate.

4.4.3. Community Detection

We considered a variety of community detection methods (e.g., Moody and White, 2003[33]) to identify the cohesive sub-groups that comprised a larger network of websites, and narrowed our selection to the faction analysis algorithm available in the UCINET software (Borgatti, Everett, and Freeman, 2002; also see de Amorim, Barthelemy, and Ribeiro, 1992; Glover and Laguna, 2013). Our preliminary analyses also included the Girvan-Newman method (Girvan and Newman, 2002; Newman and Girvan, 2004). Although both methods seek to find distinct sub-groups within a larger network, by maximizing the density[34] between group members and minimizing density between non-group members, they differ with how they identify the communities. The Girvan-Newman method identifies cohesive sub-communities by generalizing Freeman's (1979) concept of betweenness centrality to all edges in a network—what Girvan and Newman termed edge betweenness. The sequential removal of edges with high betweenness centrality effectively separates groups of actors that are more cohesive. As

[33]Although alternative method allow for the identification of nested subgroups allows for multiple group memberships (see Moody and White, 2003), we chose to focus on the "home base" community of these websites as a first step to understanding connections within this environment.

[34] Density is the proportion of direct connections found between nodes in relation to all possible connections between nodes (Garton, Haythornthwaite, and Wellman, 1997).

a result, Girvan-Newman is less adept at handling directed (non-reciprocal connections) networks such as those studied here. Our comparative analysis confirmed this limitation, as the Girvan-Newman did not produce satisfactory results (i.e., low modularity and similarity within communities).

The faction analysis algorithm we used begins with a random partition and tries to build the solution that maximizes the density in a number of groups selected in advanced by the researcher, moving individuals from one group to another until a solution deemed optimal is determined (de Amorim, Barthelemy, and Ribeiro, 1992; Glover and Laguna, 2013; Zhao, et al., 2011). Because faction analysis begins with a random partition, it is possible to obtain different results each time the method is conducted. Using a range between 2 and 20 communities, faction analyses were conducted multiple times, on each network, at each data collection point (wave), to ensure consistency. Based on goodness of fit models and visual inspection, optimal community configurations were selected and used for subsequent analyses. Goodness of fit was determined through the 'final proportion correct', which is the sum of the number of ties between websites in different factions divided by the total number of ties, and 'Q value', which provide the percentage of network ties that are within a community (Carolan, 2013). The average goodness of fit, across all ten waves of data collection, ranged from 0.81 and 0.85. Within each network, there was minimal between-wave variance (<0.01) in goodness of fit values.

4.4.4. Homophily

The tendency for two, similar, entities to associate with one another is called homophily (McPherson and Smith-Lovin, 1987). In delinquency research, the study of homophily often refers to similarities between co-offenders on attributes such as sex and age (Carrington, 2014; van Mastrigt and Carrington, 2014). However, the probability that two connected entities (e.g., people or websites) will be similar on an attribute is dictated by the demographics of the subject pool. For example, in a subject pool consisting of five boys and two girls, the baseline probability that a boy will randomly be connected with another boy is greater than the probability they will be connected with a girl. Therefore

the preference for homogeneity needs to account for the connections that exist and compare them to the expected connections, based on the subject pool structure.

To determine whether websites were more likely to connect with similar websites, we used a formula adapted by van Mastrigt and Carrington (2014), to measure expected and observed homophily. Expected homophily is the number of connections expected between similar websites, based on a random distribution of connections. As the probability of homogeneity *and* heterogeneity need to be determined, expected homophily is calculated using combinatorics. The formulas for expected homophily are as followed: $EH_{1n}=p^n$; $EH_{2n}=q^n$; $EH_{3n}=1-(p^n+q^n)$ and explained through the following example, where *p= proportion of explicit*, *q= proportion of non-explicit,* and *n=2* (number of nodes in partnership). The probability that explicit websites are connected is determined by the explicit/non-explicit composition of the network's websites and the size of the desired homogeneous group. In a network with 75% explicit websites, the probability of a dyad explicit partnership is 0.75^2 and 0.25^2 for a non-explicit dyad. Therefore, the probability of an explicit/non-explicit dyad is $1-(0.75^2+0.25^2)$. These probabilities are then compared to observed homophily[35] using a X^2. We measured homophily across the website type of focus: *sex of victim, content,* and *medium*.

4.5. Results

The 10 networks generated very similar social structures. Across the networks, Wave 1 consisted of two large, central, communities –accounting for 77.97% of websites– and three to five smaller, surrounding, communities. Average community density for site networks was 0.445, and 0.341 for blog networks. Network websites were overwhelming boy-oriented (79.2%), non-explicit (86.7%) in their sexual content, and image-focused (81.6%). As shown by Table 4-1, predominant network characteristics appeared influenced by the seed website. All 10 seed websites were non-explicit; the two networks beginning with a girl-oriented seed contained the highest percentage of girl-oriented websites (89.0% and 71.4% compared to 5.9% for the other

[35] Observed homophily is simply the *actual* number of ties between explicit and non-explicit websites

eight networks); and the two video-focused seeds had above average counts of video-focused websites (7.2% and 11.2% compared to 6.7% for the other eight networks). With the exception of blog network 4 and 5, each network's seed was located within one of the two central communities. Finally, all but blog seed 3 was active at the conclusion of the study.

Table 4-1: Characteristics of the seed, and of the network of hyperlinked websites around it

	# of CE Images	# of Child Nudity Images	# of Collateral Images	% Explicit Focused	% Boy Oriented	% Image Focused	% Video Focused
Site 1	342	24	329	8.9	96.4	76.7	17.7
Seed	4	0	0	Non-Explicit	Boy	Image	
Site 2	3112	47	2544	12.0	95.5	88.2	4.2
Seed	0	0	0	Non-Explicit	Boy	Image	
Site 3	516	161	414	6.6	97.4	85.5	5.3
Seed	27	0	27	Non-Explicit	Boy	Image	
Site 4	1292	1	1186	14.8	99.0	87.5	7.2
Seed	0	0	0	Non-Explicit	Boy	Video	
Site 5	0	17	0	12.3	11.0	72.6	8.4
Seed	0	0	0	Non-Explicit	Girl	Image	
Blog 1	5255	125	3146	23.9	74.5	85.3	2.3
Seed	0	0	0	Non-Explicit	Boy	Image	
Blog 2	1308	14	904	12.0	98.4	81.2	6.8
Seed	2	0	14	Non-Explicit	Boy	Image	
Blog 3	610	70	305	7.5	97.7	86.9	5.2
Seed	0	0	0	Non-Explicit	Boy	Image	
Blog 4	898	9	351	11.2	93.7	75.6	11.2
Seed	0	0	5	Non-Explicit	Boy	Video	
Blog 5	1	13	0	24.2	28.6	82.7	3.3
Seed	0	0	0	Non-Explicit	Girl	Image	

It is useful to focus on 'representative' networks (closest to the average across network characteristics) in order to illustrate some of the more detailed findings. Figure 4-1 displays representative site network 1 (Sweet Love[36]), and Figure 4-2, blog network 2 (Teddy Bear), with the seed-website circled. To aid in identification and subsequent recall of communities being described, their names are derived from the most prominent characteristic(s). Although the networks are large enough that the specific connections between websites cannot be emphasized, the figures still provide an important visual display on how dense are the hyperlinked networks formed and how the pockets of websites, cohesive enough to be identified as communities, form around the core. Figure 4-1, for example, shows different communities forming around the distribution of videos, despite the seed being image focused (only 0.08 videos per webpage). Figure 4-2, instead, shows a pattern around sub-communities specializing in the distribution of images, much like the seed. The first network (4-1), which started from a boy-oriented illegal website, even included a community of adult gay videos (right side of the figure). That community is slightly more isolated from the rest of the network, and as shall be seen below, is the only one without a trace of illegal or grey area material.

Table 4-2 summarizes the community characteristics of the two network representatives Sweet Love and Teddy Bear. As was the case across all ten networks, websites were predominantly boy-oriented; 96.4% (294/305) in Sweet Love and 98.4% (304/309) in Teddy Bear. The two core communities within each network differed from the surrounding communities in two key ways. First, the majority of illegal images were within the two core communities. Within Sweet Love, the community *Network Core* (n=142) contained 95.3% of CE images and 94.5% of Collateral images. Within Teddy Bear, the community *CE Image Core* (n=117) contained 72.9% of CE images and 88.8% of Collateral images. Second, the two core communities within each network contained the vast majority of websites publishing videos (Table 4-3).

[36] To preserve the anonymity of the websites, we use fictitious web domain names for the purpose of this study.

Figure 4-1: Network of site Sweet Love (bold white circle) at Wave 1, displaying its six communities

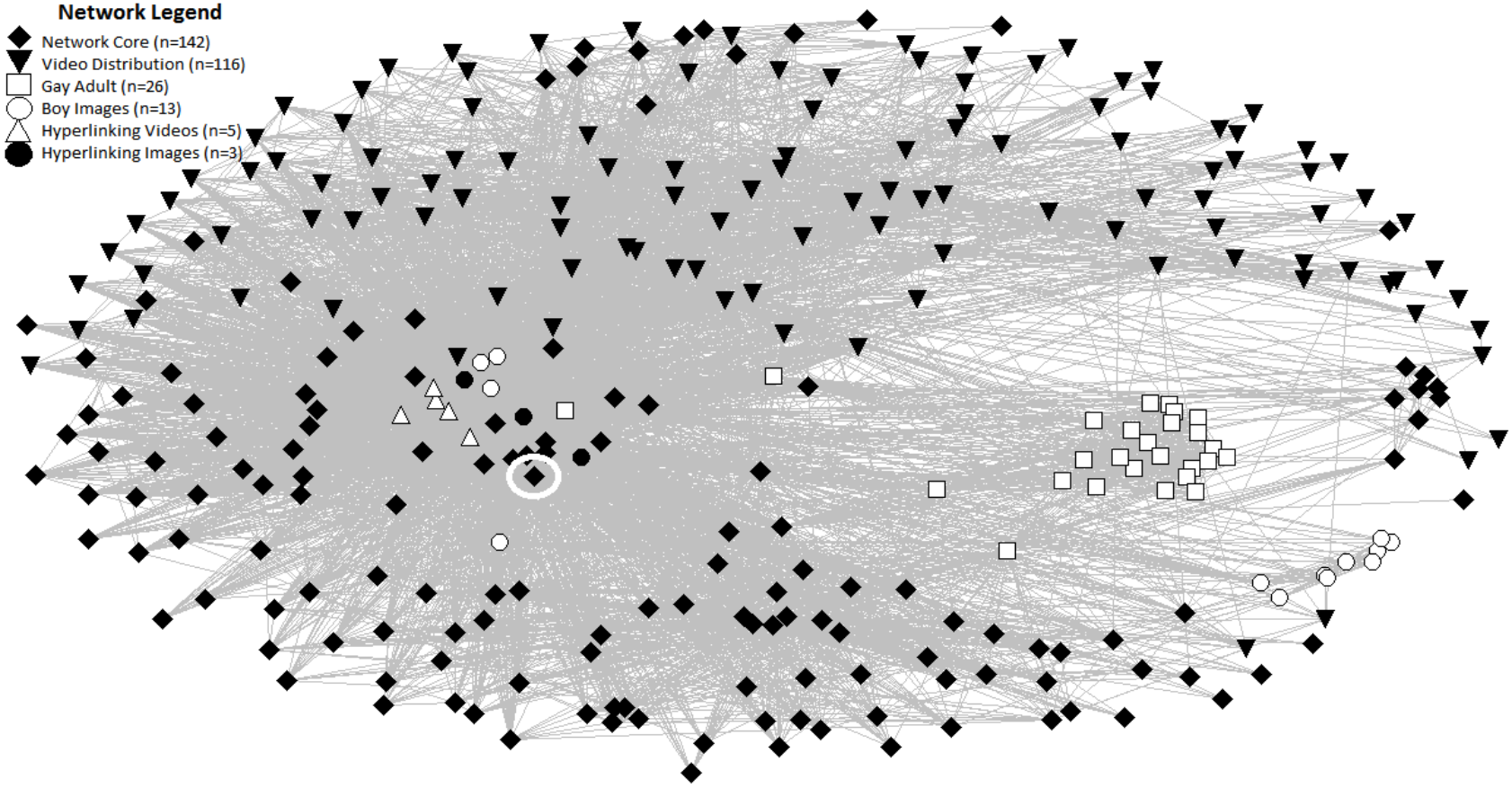
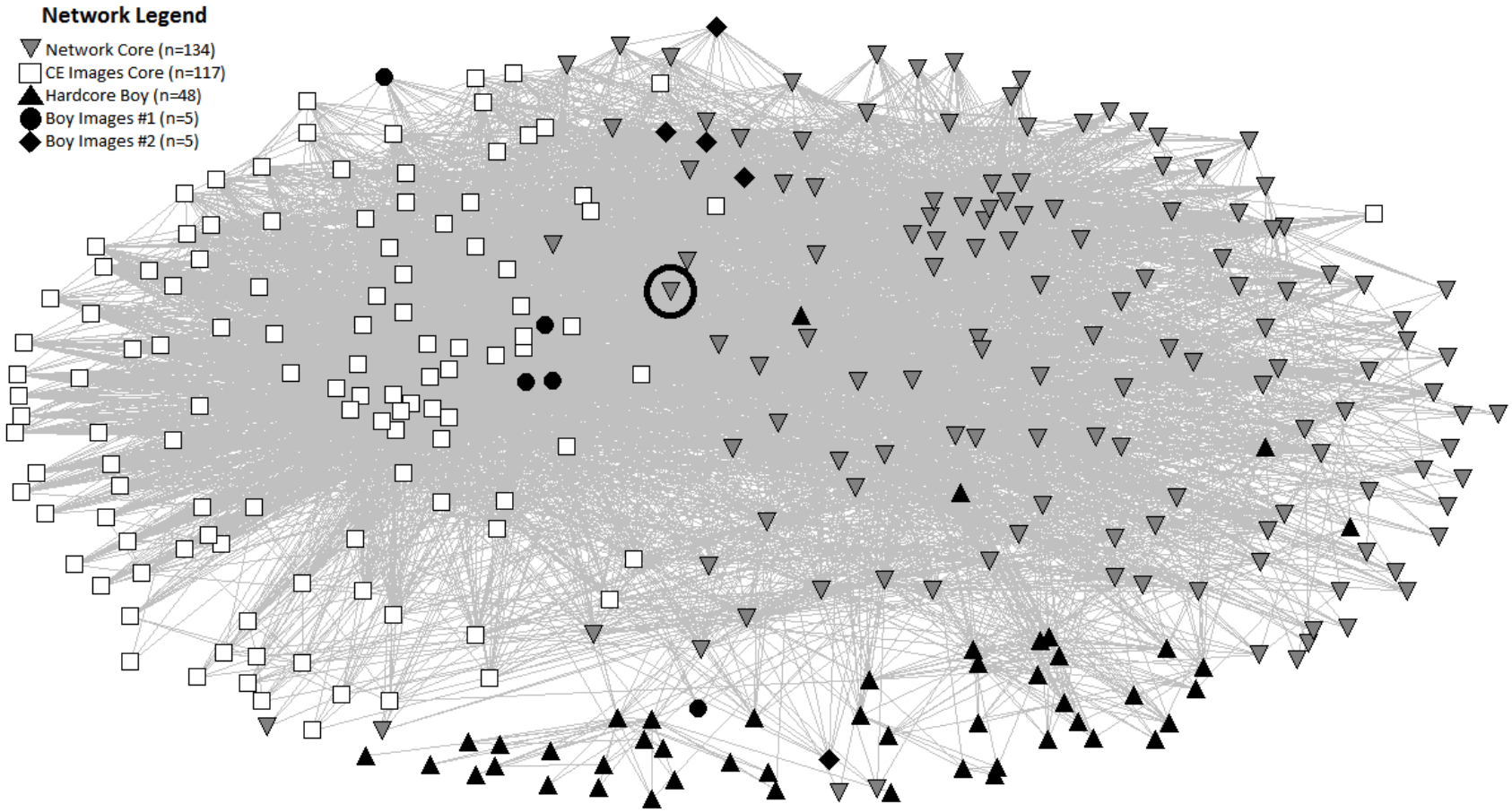


Figure 4-2: Network of blog Teddy Bear (black bold circle) at Wave 1, displaying its five communities



Within each network, websites that clustered together often shared some prominent characteristics. In the two networks where the seed-website was not located in one of the primary two communities, the seed did not appear to influence the community characteristics. For example, blog seed 4 was the *only* video-focused website within their community while blog seed 5's community was more than 40% boy-oriented; despite the seed being one of two girl-oriented seeds. Focusing on the two representative networks, within Sweet Love, communities *Hyperlinking Videos* (n=5) and *Hyperlinking Images* (n=3) contained websites that could be characterized as brokers; hyperlinking out to between 30 and 50% of all the websites within the network. *Gay Adult* (n=26) was comprised of adult gay content -based on website names and type of content- while *Video Distribution* (n=116) was core to the network but distributed video content. In Teddy Bear, similar patterns emerged. Communities *Boy Images #1* (n=5) and *Boy Images #2* (n=5) distributed boy images while *CE Images Core* (n=117) predominantly distributed known illegal images. Together, these communities, especially Sweet Love's *Video Distribution* (#6) and Teddy Bear's *CE Images Core* (#2), suggest that some websites will become close-knit with websites that they share important similarities.

Central to the notion of community cohesion is mutual acknowledgement. In network terms, mutual acknowledgement can be categorized as the percentage of connections (i.e., hyperlinks) that are reciprocated amongst community members. Across the ten networks, reciprocity was 23.0% (22.6% for site networks and 23.4% for blog networks). However, we observed higher cohesion when examining within-community reciprocity. As would be expected, smaller communities were substantially more cohesive than larger communities. For example, weighting each community equally, communities with more than 100 websites had a reciprocity of 27.4%, compared to 53.1% within communities with less than 100 websites. For communities with less than 30 websites, reciprocity averaged 47.7%. While smaller communities were more reciprocal, the presence of known CE images did not enhance or decrease reciprocity. This finding held at multiple cut-off points for both number of CE images and websites.

Table 4-2: Community size and CE image composition for representative networks Sweet Love (site) and Teddy Bear (blog)

	Community Name	Count (% of Network Websites)	Avg. Nb. of Pages per Website	CE Images (% of Websites)	Child Nudity Images (% of Websites)	Collateral Images (% of Websites)
Sweet Love	Network Core	142 (46.6)	917.8	326 (7.8)	24 (16.9)	311 (6.3)
	Video Distribution	116 (38.0)	1671.0	1 (0.9)	0 (0.0)	10 (0.9)
	Gay Adult	26 (8.5)	299.2	0 (0.0)	0 (0.0)	0 (0.0)
	Boy Images	13 (4.3)	243.5	13 (30.8)	0 (0.0)	6 (23.1)
	Hyperlinking Video Sites	5 (1.6)	1864.4	1 (20.0)	0 (0.0)	1 (20.0)
	Hyperlinking Image Sites	3 (1.0)	1866.7	1 (33.3)	0 (0.0)	1 (33.3)
Teddy Bear	Network Core	134 (43.4)	972.7	233 (2.2)	1 (0.8)	14 (0.8)
	CE Images Core	117 (37.9)	521.5	953 (10.3)	13 (11.1)	803 (6.8)
	Hardcore Boy	48 (15.5)	4298.7	43 (2.1)	0 (0.0)	42 (2.1)
	Boy Images #1	5 (1.6)	617.0	66 (40.0)	0 (0.0)	32 (40.0)
	Boy Images #2	5 (1.6)	2695.4	13 (40.0)	0 (0.0)	13 (40.0)

Table 4-3: Content and connectivity descriptives, by community, for representative networks Sweet Love (site) and Teddy Bear (blog)

	Community Name (Count)	% Explicit	% Boy	% Video	% Image	% Story	Avg. Outgoing	Avg. Incoming
Sweet Love	Network Core (142)	4.9	93.0	19.0	71.8	9.2	21.8	26.4
	Vid. Distribution (116)	17.2	99.1	21.6	75.0	3.5	18.4	23.5
	Gay Adult (26)	0.0	100.0	0.0	100.0	0.0	29.7	29.4
	Boy Images (13)	0.0	100.0	7.7	92.3	0.0	48.3	17.2
	Hyperlinking Video Sites (5)	0.0	100.0	20.0	80.0	0.0	109.6	23.4
	Hyperlinking Image Sites (3)	0.0	100.0	0.0	100.0	0.0	157.3	21.7
Teddy Bear	Network Core (134)	16.4	99.3	9.0	76.9	14.2	23.8	30.0
	CE Images Core (117)	4.3	97.4	6.8	88.0	5.1	32.0	32.0
	Hardcore Boy (48)	20.8	97.9	2.1	72.9	25.0	17.3	10.7
	Boy Images #1 (5)	0.0	100.0	0.0	100.0	0.0	65.8	16.2
	Boy Images #2 (5)	0.0	100.0	0.0	100.0	0.0	64.6	12.4

4.5.1. Change in Community Composition

The stability, or volatility, of virtual communities is important for understanding how quickly online networks evolve. If connections between websites are transient, in place only to serve a specific purpose and then removed, then distribution networks can be seen as operating closer to a pure availability framework. However, if the connections are long-lasting, this suggests that relationships are more likely to form between websites and their operators. That is, they are in a better position to develop a certain level of trust with specific websites and work together to achieve a common goal. To measure stability, we had to force select the number of factions at each time point to

allow for cross-wave comparison[37]. Upon selecting the *best* stable community structure, across the duration of the study, community movement was calculated at each wave. If a website moved from one community to another at any point during the study period, it was classified as being dynamic.

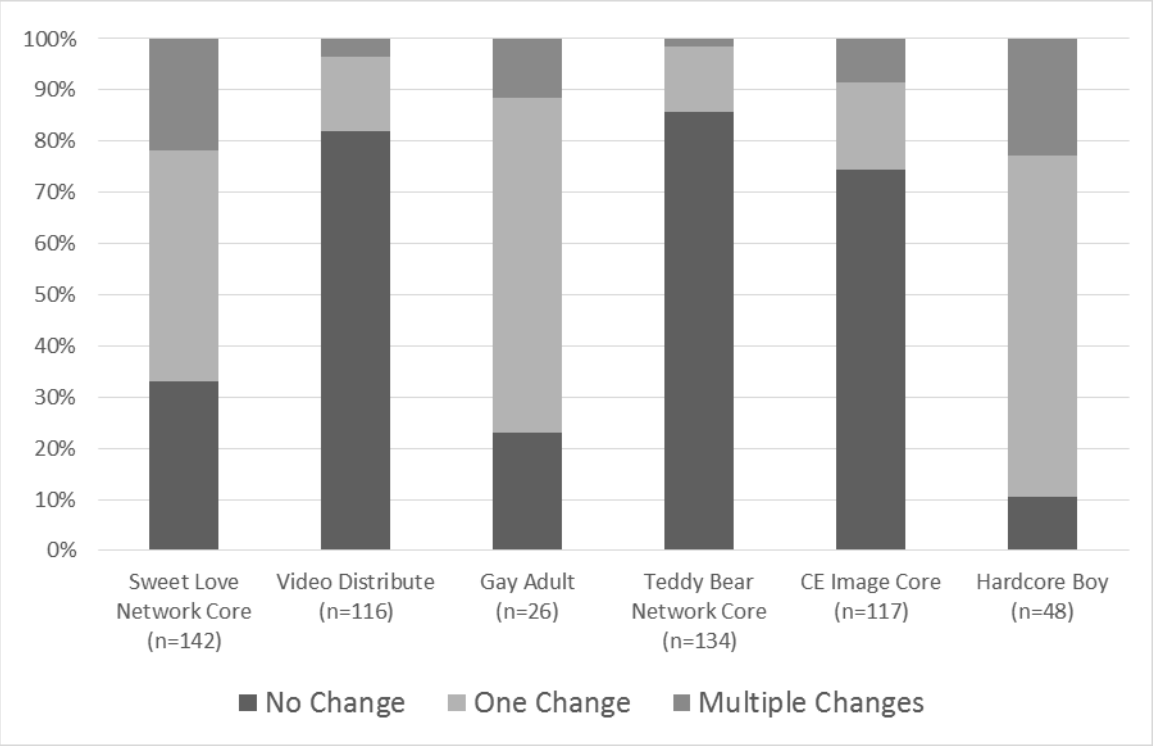
Results of the community stability analyses reveal equal consistency and instability in group membership. Nearly half (48.86%) of all websites did not transition to another community within their network, with less between-network variability amongst blog seed networks. Within blog networks stability ranged from 44.1% to 67.0% (s.d. =9.9%), while community stability ranged from 14.8% to 76.0% (s.d. =20.8%) in site networks. Amongst the ten seed websites, three remained in the same community throughout while five switched communities three or more times. The five most stable seeds remained in the two largest communities. Where community stability was the lowest –site network 5, blog networks 3 and 5– seed-websites experienced the highest rates of community movement. Figure 4-3 displays stability for the three largest communities –where stability could best be measured– in Sweet Love and Teddy Bear. We also identified the number of websites that switched between two communities and more than two.

Community transition most often involved websites from smaller communities moving to one of the two core communities; potential evidence of smaller websites being 'accepted' into the larger community. In some situations, community transition was cluster-based. For example, in Sweet Love, 14 websites moved from *Gay Adult* to *Boy Images* at Wave 4. However, at Wave 5, all 14 rejoined *Gay Adult.* Finally, some websites oscillated between the two core communities. Within each network, the stability of the two core communities was higher than the surrounding, smaller, communities. Community movement was dominated by websites within the smaller communities transitioning to the core communities. There were some exceptions to this such as *Boy*

[37] For each network, faction analyses were conducted at each wave for four to 18 clusters. The corresponding *final proportion correct* and *q-value* were compared within and between waves. The cluster formation that corresponded to a high final proportion correct and q-value, compared to the other cluster formations, was selected. In situations where the highest final proportion correct and q-value was not selected, these values did not differ from the highest by more than 0.1%.

Images from Sweet Love, where 11 out of the 13 websites remained in the same community throughout the study. *Boy Images* community was also one of only a handful that saw an increase in the quantity of known child exploitation images from Wave 1 to Wave 10. Smaller stable communities might be evidence of some websites insulating themselves from potential infiltration and/or detection.

Figure 4-3: Community stability in the three largest communities in site network Sweet Love and blog network Teddy Bear



4.5.2. Homogeneity between Connected Websites

As entities, websites do not possess any personality characteristics to examine as predictors of websites selecting similar others. In order to determine homophily amongst connected websites, website content characteristics are logical choices for

analyses. Using two of the composite variables[38] created from our keywords –severity (explicit/non-explicit), and medium (image/video/story) – we explored whether websites purposely connected to websites with the same composition (Table 4-4). The results suggest that websites do not show a strong tendency for homophily in terms of severity,

Table 4-4: Observed and expected homophily across ten networks for three composite specialization variables –sex of victim, severity, and medium

		Sex of Victim			*Severity*			*Medium*			
		Boy (%)	*Girl (%)*	*Mix (%)*	*H.C. (%)*	*S.C. (%)*	*Mix (%)*	*Image (%)*	*Story (%)*	*Video (%)*	*Mix (%)*
Site 1	*Obs.*	88.21	0.16	11.63*	4.61**	77.70	17.69	72.39	0.07**	2.28	25.25
	Exp.	93.47	0.11	6.42	0.98	81.09	17.93	64.26	0.30	2.52	32.92
Site 2	*Obs.*	82.03	2.32	15.65*	3.66	62.70	33.62**	41.39	0.67	4.29	53.65
	Exp.	92.19	0.53	7.27	2.84	74.60	22.57	52.05	0.98	4.18	42.78
Site 3	*Obs.*	95.43	0.16	5.87	0.17*	93.42	6.65**	67.21	0.50	0.90*	31.39
	Exp.	94.67	0.10	7.78	0.37	84.64	15.69	64.60	0.41	3.04	31.94
Site 4	*Obs.*	98.59	0.01	1.40**	2.83	63.27	33.90	0.01**	8.22	46.15	45.62
	Exp.	96.74	0.28	2.98	2.20	71.42	26.38	0.11	4.54	67.37	27.97
Site 5	*Obs.*	0.82	84.46	14.72	1.56	11.23	75.95	4.55*	0.18	64.12	31.15
	Exp.	1.42	77.65	20.93	1.68	75.95	22.37	1.72	1.75	61.06	35.47
Blog 1	*Obs.*	72.30	3.30	24.40	1.49**	76.48	21.54	61.98	0.14	1.79	36.10
	Exp.	61.19	4.81	34.00	4.34	62.84	32.82	80.43	0.37	1.01	18.19
Blog 2	*Obs.*	95.14	0.02	4.84	1.26	80.73**	18.01	43.61	1.17	4.86	50.36
	Exp.	95.86	0.43	3.71	1.49	77.15	21.36	55.56	0.78	3.26	40.40
Blog 3	*Obs.*	97.23	0.00*	2.76*	0.08**	95.60	4.31*	57.11	0.42	8.95	33.52
	Exp.	94.13	0.24	5.64	0.49	87.26	12.24	66.84	2.44	3.64	27.08
Blog 4	*Obs.*	95.56	0.08	4.36	2.38	74.83	22.79	0.60	7.61**	53.32	38.47
	Exp.	92.08	0.17	7.75	1.46	76.07	22.27	0.99	2.21	64.58	32.22
Blog 5	*Obs.*	2.30**	67.84	29.86	2.72**	64.95	32.33	0.27*	0.57	71.91	27.26
	Exp.	4.50	37.59	57.91	3.74	39.90	24.34	1.19	0.91	65.84	32.07

*p>0.05 **p>0.01

[38] The boy/girl orientation of a website was not analyzed for homophily because the vast majority of websites examined were boy focused (over 95% in most networks). We did run the sex homophily analysis on the two networks with a majority of girl-oriented websites, and that although not statistically significant girl-oriented websites did cluster at higher than expected rates. In Blog 5, 68% versus 38% while in Site 5, 84% versus 78%.

or the medium predominantly used. Sexually explicit websites were just as likely to connect to non-explicit websites, and rarely did we find the medium used on websites leading to these websites narrowing down their hyperlinks to websites emphasizing similar media. Diversity, rather similarity, appears to drive the hyperlinking among the websites analyzed in this study.

4.6. Discussion

Growth in the prevalence of cybercrime has necessitated increased research into criminal processes conducted online. For many cybercrimes, websites play a crucial role in facilitating criminal activities. While initially produced by individual offenders, the distribution of child sexual exploitation (CE) material is heavily influenced by public websites that host and disseminate images, videos, and other content, and connect individual offenders with each other. The vast amount of data available on the Internet, and public websites especially, provides an opportunity for innovative conceptual and methodological approaches to studying online (and offline) criminal processes. In the current study, we designed an automated tool to collect data on websites involved in the dissemination of CE material and the surrounding hyperlinked websites. From this examination we can assess how criminal-based website communities structure and function within the larger network, and how they evolve over time. We focus on two main findings from this study and conclude with a discussion of the implications for subsequent research into co-offending selection processes.

First, we found that despite having stronger ties (more reciprocated hyperlinks), communities were not comprised of websites focused on specific types of content. This finding extended beyond the macro, community, level to the micro, website, level, as our homophily analysis revealed that websites focused on the same type of content did not connect with homogeneous others at higher than expected rates. Although prior research on these issues is scant, this finding contrasts with those of Burris et al. (2000) who found that the online White Supremacists community showed a clear preference for connecting to websites with similar interests. The lack of content-based clustering found in our study may be indicative of the multiplicative nature of the Internet. In cyberspace CE material can easily be copied from one website to another. This easy access to

content means that websites do not need to be closely connected to similar websites. In fact, the nature of the Internet may mean that multiple connections to similar websites may have negative consequences. While connecting with homogeneous others may provide access to preferred content, the wide distribution of material might mean that the law of diminishing returns applies –being connected to many similar websites may provide access to competitors, leading to lost traffic (e.g., visitors preferring the 'new' website). This conclusion is supported, in part, by Burris, Smith, and Strahm's findings that among commercial (Tremblay, 1993), a website's ability to connect to a variety of CE content types maximizes their potential traffic and partners for exchanging material. A website potentially increases its appeal by providing one type of content directly (via text, images, or videos), and a variety of types indirectly (via hyperlinks). As our analyses were focused on the content distributed, it remains possible that websites do connect to homogeneous others based on other, unmeasured, characteristics, such as the type or quantity of community members, the ability to acquire and distribute new material, the personal relationships between website operators, or the aforementioned age of the victim depicted.

It is important that we also acknowledge the possibility that, outside of a few choice connections, the communities that form around public websites disseminating CE material base their hyperlinking practices on pure availability. Public websites may operate as catch-all marketplaces with their primary purpose being to attract and connect as many people as possible, regardless of their primary material of interest. This possibility is reinforced by the similarities between the distribution mechanisms of piracy (e.g., Drachen and Veitch, 2013; Qu, et al., 2013) and CE material. At some point a website provides a new piece of content. Shortly after, every other website copies that piece of content from one another, ignoring the original source, and integrates it into their community. This method allows for the quickest access of material, through division of labor, and minimizes the risk for the originating source as it becomes nearly impossible to identify the original due to slight modifications during each replication (e.g., resizing of the image). If this is true, that large illicit networks improve efficiency, increase deindividuation, and reduce risk (McGloin and Piquero, 2010), this diffusion method provides a suitable online adaptation. Contrary to many illicit enterprises that strive to confirm the adage that 'small is beautiful' (e.g. Bouchard and Ouellet, 2011;

Reuter, 1983), *public* CE-related communities may function under the premise that any connection is a good connection.

Are online hyperlinking ties so different from offline ties, and should we expect higher numbers than the ones we found in this study? Our second main finding was that community stability varied across networks and no consistent trends could be uncovered. Unlike connections in offline criminal communities, which can easily be severed through inactivity, the removal of a hyperlink between two websites requires a conscious effort on the part of the operator(s) to severe the connection. That is, hyperlinks between two websites do not have natural decay periods. This same phenomenon can be extended to all cyber-relationships. For example, friendship ties on social networks websites (e.g., Facebook©) are only severed when a person consciously 'unfriends' a relationship. With this in mind, it is important to note that the presence of a hyperlink between two websites should not be viewed as evidence of a strong connection between the websites. Rather, the presence of the hyperlink is simply an acknowledgement of awareness and that the connected websites share a common interest.

Given the lack of research on criminal community stability it is difficult to say definitively whether our finding of 48.86% community stability is different from offline criminal communities. While the level of care and nurturing required for a virtual connection may be lower than for an offline connection, it is worth noting that Kreager and colleagues (2011) reported a group stability of 35% between grade 8 and 9 for offline adolescent friendships, slightly lower than our virtual communities followed for 60 weeks. One key difference between offline and online criminal communities is that the lower amount of efforts required to maintain a connection in cyberspace may facilitate the creation of larger criminal communities, with potential implications for the creation of offending opportunities.

While the website communities formed provide individual offenders with the opportunity to acquire criminal capital (McAndrew, 2000; McCarthy and Hagan, 1995; McGloin and Piquero, 2010), they also provide opportunities for non-criminal physical, economic, social, or psychological exchanges which may otherwise be unattainable.

These added benefits have been found in both offline (Morselli, Tremblay, and McCarthy, 2006; Weerman, 2003) and online contexts (Taylor and Quayle, 2003; Tremblay, 2006). As such, it is important that co-offending research goes beyond the immediate criminal event, and examines the role of the broader criminal community in which offenders are embedded, and how its structuring may impact co-offending opportunities. Online, this means examining the communities formed between websites and even within a website, as they provide information on how partnerships materialize, what steps are taken to nurture said partnerships, and the success of the partnership.

4.6.1. Limitations

Our study was subject to three primary limitations that we feel deserved specific attention. First, identifying communities within a network is a subjective process. If tasked with deriving a specific number of communities from any group of people, entities, or objects, a solution can be found. However, it does not necessarily mean that the solution found is the most suitable choice nor that it has any significant meaning. Therefore, it is important that we acknowledge the possibility that the communities we identified were not the most accurate representation of the true nature of the larger network. To best address this possibility, we undertook the community analysis with no preconceived expectations and allowed the *final proportion correct* (goodness of fit) and community densities dictate the number and size of each community. In addition, we tested goodness of fit for between four and 18 communities and conducted our analysis across ten networks, starting from two different website types (blog/site). As our final proportion correct was high (0.80 to 0.85), within-network variance was low (<0.01), and each network was similar in the number of communities, we believe that our interpretation of the community structure of our networks was valid.

Second, faction analysis functions under the notion that nodes, in this case websites, can belong to *only* one community. Research into multiplexity offers plenty of theoretical and empirical support for how individuals belong to multiple, partially overlapping, communities (e.g., Feld, 1981; Krohn Massey, and Zielinksi, 1988; Papachristos and Smith, 2014). Although a nested community detection analysis (see Moody and White, 2003) would have allowed for multiple group memberships, we

elected to focus on each website's 'home base' as allowing 300+ websites to each belong to multiple communities would have made the analysis unnecessarily complicated and removed some of the meaningfulness of a website's primary community attachment.

Third, as the data collection process was automated and not exhaustive, it is unclear whether each website included in a network was directly involved in the dissemination of CE material. As offline delinquent peer groups are interspersed with non-delinquent peer groups (Kreager, Rulison, and Moody, 2011; Haynie, 2002), we would expect that pattern to exist online, with delinquent websites being connected to non-delinquent websites. However, a website does not need to be directly distributing CE material to be CE-related. Websites that connect to other websites distributing CE material can act as gatekeepers to material and perpetuate the distribution process by providing direct access. Although, we included those not directly distributing CE material, to preserve the reality of the larger online social structure, our inclusion criteria (images or keywords) should have minimized false positives, or more colloquially guilty-by-association.

4.7. Conclusion

Using a repeated measures design, to study the communities surrounding websites distributing child sexual exploitation (CE) material, we found that networks were comprised of two, large, core communities surrounded by a series of smaller communities. Known CE images were rarely found once we moved away from the immediate communities surrounding the seed website. Homophily within various content types was no greater than would be expected had websites been connected at random. Findings from this study shed light on the communal nature of CE distribution in cyberspace and how CE-related website connect. It also provides a framework for future research into co-offender selection processes, as the opportunities afforded to individual offenders are influenced by the macro network structure they reside within.

The public nature of the Internet provides an ideal setting for exploring social science questions that may otherwise be difficult to measure offline. The key is whether

the processes identified online can be applied offline, and vice-versa. While outwardly different, online and offline criminal practices have been suggested as overlapping extensively (Grabosky, 2001). Moreover, online practices have been shown to influence offline criminal networks (Moule, Pyrooz, and Decker, 2014; Pyrooz, Decker, and Moule, 2015). It is conceivable that how other social problems manifest online may be indicative of how they function offline. Therefore, the Internet, and the use of automated data collection tools, may provide a huge advantage for accessing hard-to-reach populations, answering research questions that require additional anonymity, and examining how the Internet has modified societal processes.

Chapter 5. Conclusion

The growth of the Internet and the advancements in image and video capturing devices has transformed the previously isolated crime of child sexual exploitation (CE) into a global, community-based, crime. As of 2000, more than three-quarters of CE cases involved a virtual component (Hughes, 2002). Increased prevalence can be partly attributed to the inherent advantages of cyberspace—decentralized, lacking boundaries and laws, anonymous—modifying the structure and processes associated with the crime.

I drew from a longitudinal research design to allow for the measurement of websites involved in the distribution of CE material, at multiple points in time. The data were collected using a type of snowball sampling, through hyperlinks between websites. Starting from a 'seed', subsequent websites were included in the sample through hyperlinks, and meeting specific criteria. Thus, each website was connected to at least one other website within the network. This dissertation differs from existing research as I examined publicly accessible websites, rather than peer-2-peer networks, incorporated the transition of CE to a communal crime, using social network analysis, and began developing a theoretical framework for online CE distribution between entities, based on the criminal career paradigm. Within, I presented three studies to provide evidence that Internet-mediated research (IMR) methods (e.g., automated data collection) can be beneficial for understanding CE material distribution and cybercrime, and that elements of the criminal career framework can be applied to a virtual environment to explain online criminal processes. I outline, and then discuss, three main conclusions from this dissertation.

1) The transitioning of criminal career dimensions to cyberspace open up new avenues for how we conceptualize the evolution of cybercrime processes and the role that non-individual entities play in facilitating crime online and offline.

2) Automated data collection tools are useful for cybercrime research provided that due diligence is conducted in selecting the most appropriate inclusion/exclusion criteria.

3) Websites within CE-seeded networks differ across multiple attributes from websites within non-CE-seeded networks.

5.1.1. A Theoretical Framework for Online Criminal Careers

Growth in the prevalence of cybercrime has necessitated increased responses from criminologists to research the mechanisms that facilitate the online criminal event. Although some offenders intersperse offline and online crimes, others conduct their criminal business solely in cyberspace. As Grabosky (2001) and others (e.g., Moule, Pyrooz, & Decker, 2014) have suggested an overlap in the mechanisms for offline and online crime, it is important that traditional explanations of offline crime are transitioned to cybercrime. An important goal of this dissertation was to begin the process of developing a theoretical framework to explain the evolution of strictly online crimes, based on the concepts and theories examined through the criminal career paradigm lens. Although this dissertation only scratches the surface of building a framework to explain cybercrime evolution, it does provide a base for further discussion and development.

Central to developing an online theoretical framework, based on the concepts examined in the criminal career paradigm, is determining how best to transition existing concepts. Chapter 3 (study 2) explores the transition process, first explaining how to define the criminal career unit of measure. I note that while the difficulties in tracking people in cyberspace suggest that continually modifying one's username is the best way to avoid detection, research shows that offenders often maintain the same username because of the notoriety and respect the username has acquired (Decary-Hetu & Dupont, 2013; Decary-Hetu, Morselli, & Leman-Langlois, 2012). Therefore, the online criminal career is conceptualized, at the micro level, as the development and evolution of a pseudonym. Further adding to the complexity of conceptualizing the online criminal career, I examined cybercrime at the macro, website, level.

The concept of the entity criminal career has yet to be fully developed in criminological research. The closest conceptualization stems from the understanding that the careers of criminal organization members cannot be examined independent from the organization's trajectory (Tremblay, et al., 1989). However, this dissertation takes that concept and goes one step further by arguing that the organizations themselves have their own criminal career. Although not a common way to think about crime, the use of groups as a unit of measure is common in peer and gang research. a new way of thinking about crime, the concept of an entity having its own career path is not unique. For example, while legitimate businesses have Chief Executive Officers (CEO), board members, and share-holders, the long-term evolution of the businesses are not contingent on one person. Even if a CEO operates for an extended period of time, they do so within the context of the company's mission and vision. The CEO may have a strong influence on the direction of the company and may modify the evolution, to more closely align with their vision, however, once a CEO is removed, the company continues to function; and does so similarly to how it functions previously. Therefore, while specific people may influence the goals and direction, the company has a life of its own that continues to operate outside the controls of people. Likewise, a criminal organization's leader may heavily influence the direction of the organization, but they do so within the structure that existed prior to their arrival. Within this dissertation, I developed the idea of an online entity having its own career path. I conceptualize the online entity to be that of a website domain name, and use websites involved in the distribution of CE material as the criminal case study.

While the boundaries that define an online criminal career are important, so too are the application of existing concepts. Again, Chapter 3 outlined the difficulties in transitioning criminal career dimensions to cyberspace, with specific focus on quantifying offending frequency, and translating individual measures to entities. Offline, criminal career research has focused on personal attributes to characterize offending groups and trajectories. However, many characteristics used, such as age, sex, ethnicity, and location, are unknown online. Therefore, an online criminal career framework requires the development of new measures to characterize typologies of offenders. Within this dissertation I suggested, and tested, some website attributes for measuring career duration (specifically career interruptions), offending frequency, crime-type mix, and co-

offending (specifically connectivity). While career interruptions and connectivity proved to be more useful, how best to measure offending frequency complicated its application and subsequent interpretation. Most difficult to operationalize was crime-type mix. Trying to mirror offline sex offender crime types, I developed measures based on sex of victim, 'severity' of victimization, 'type' of victimization. These attributes did not appear to effect survival (Chapter 3) while community structure and connectivity (Chapter 4) were mildly influenced by these measures. The limitations of these measures and suggestions for further refinement in future research are discussed in detail below. Nevertheless, the conceptualization of individual criminal career dimensions to entity attributes provides evidence for their application going forward, and this dissertation act as a starting point for further discussion and refinement.

5.1.2. Automated Data Collection Techniques for Cybercrime Research

The use of IMR methods has grown across disciplines in large part to their cost-effectiveness and efficiency. Further extending those advantages, automated data collection tools have become more commonplace. Although used to examine criminological research questions, their reliability and validity have yet to be determined within criminology. While it may appear that validation research on the accuracy of automated data collection in other disciplines should suffice, the findings of these studies do not necessarily translate to the study of cybercrime. Given the illegal nature of cybercrimes, such as the distribution of CE material, it is possible that offenders are more likely to employ strategies for avoiding detection. Therefore, the secretive nature of criminal activity means that automated data collection techniques need to be investigated specific to cybercrimes. Within Chapter 2 I developed and assessed the ability of an automated webcrawler, known as the Child Exploitation Network Extractor, to distinguish between relevant and irrelevant data based on predefined CE-related criteria.

Key to any automated data collection process, where the data are not manually verified after, is whether the data are actually relevant to the topic being investigated. For online CE, the validity of the data is further in question due to the overlap between

CE material and legal, adult, pornography. This overlap stems from a societal fixation on young, nubile, sex figures within adult pornography, and the descriptions used. Within Chapter 2 I compared networks beginning from 10 known CE websites to those from 10 non-CE sexuality websites and 10 sports websites. I found that CE images, provided by the Royal Canadian Mounted Police and used to convict offenders in Canada, were prominent within networks consisting of CE websites, minimally prominent within non-CE sexuality networks, and non-existent within sports networks. Code keywords—those used by offenders to direct other offenders to CE material—were not effective at distinguishing between the three types of networks. However, thematic and sex-oriented keywords were more frequent within CE-seeded networks than comparisons suggesting the quantity, rather than simply the presence, of keywords may be helpful in identifying websites involved in the distribution of CE material.

Automatic data collection is not without its difficulties no matter the field of study. Within cybercrime research, and CE material distribution specifically, the difficulty stems more from the cross-pollination of CE-related descriptors than from efforts to hide the crime. As words are used by multiple subgroups in different contexts, their ability to act as markers for identifying a specific group becomes more difficult. This is especially true for identifying illegal groups as their efforts to remain hidden may lead to regular changes to vocabulary. However, this change is gradual as modifying them too quickly may result in some groups falling behind and intermittent offenders being unable to reconnect with the subgroup. Additionally, keyword overlap can occur between those committing the crime and those attempting to prevent or control the crime. Bouchard, Joffres, and Frank (2014) note the complication of using keywords to identify terrorist websites as the same keywords used by terrorists are also used by counter-terrorist organizations. The replicative nature of the Internet and society (e.g., what was once old is new again) coupled with those refusing to embrace change, or offenders believing that they will never be caught, means that historical vocabulary still has its place in identifying criminal activities. However, the current landscape needs to be understood and taking into consideration when selecting inclusion criteria focused on vocabulary.

For other criteria, such as the use of specialized database (e.g., image hash values), this dissertation, and previous research (Westlake, Bouchard, & Frank, 2012),

points to their degree of completeness as a key factor in their reliability and subsequent usefulness in automated data collection strategies. For some types of cybercrime, the completeness of a database may be more easily obtained. In some cases, a researcher (or social control agency) may only be interested in specific data and thus can tailor their database to the needs of the data collection. In these scenarios the impediment of an incomplete database may have less of an influence on the conclusions drawn. In sum, the evolving nature of CE distribution, various definitions of CE across countries (and researchers), ease of modifying a hash value, and vast quantity of CE images available impact the reliability of these databases for CE researchers. However, when used in conjunction with other criteria (e.g., keywords), they are effective at conclusively identifying data pertinent to the research question, as they leave little to no argument against the relevancy of the data collected and corresponding findings.

5.1.3. Describing the Characteristics of Websites in CE-seeded Networks

Criminological research focuses on explaining how those who commit crime, or a lot of crime, differ from those that do not. My final objective in this dissertation was to describe and compare websites within CE-seeded networks to the networks surrounding non-CE sexuality and sports websites. These comparison networks were selected to show how CE-seeded networks differed from the similar non-CE sexuality networks and the dissimilar sports networks. Websites within CE-seeded networks contained more webpages than sports websites, were more image-based, and had different hyperlinking properties. More specifically, the hyperlinking properties of CE-seeded websites suggested their networks were comprised of hubs as they were denser than non-CE sexuality and sports networks, but comparable in clustering and reciprocity. Evidence of hubs within CE-seeded networks has implications for law enforcement disruption strategies (see Joffres, Bouchard, Frank, & Westlake, 2011) and how information is diffused throughout the network. Most importantly, while the Internet has transitioned the distribution of CE material from sale to trade (Beech, Elliott, Birgden, & Findlater, 2008; Estes, 2001), evidence of hubs may suggest that there is more competition between CE websites than currently suspected; an area for further research. This is not to suggest that CE distribution does not revolve around exchange. Rather, CE websites are

selective about whom they exchange and may compete to acquire consumers or provide new content.

Adding to the complexity of how CE material is exchanged, based on their crime-type mix, websites within CE-seeded networks did not connect to homogeneous others at higher than expected rates. In addition, faction analysis, based on the hyperlinks between websites, revealed that communities were not comprised of specific crime-type mixes. For example, girl-focused websites did not connect with other girl-focused websites at a higher than expected rate, nor did they cluster together within the larger network. This potentially opposes the above finding that CE websites are selective about who they exchange with, and is contrary to previous terrorist and extremist website communities research (Burris, Smith, & Strahm, 2000; Chau & Xu, 2008; Zhou, et al., 2005); however, another explanation may be that those they select to associate with is not based solely on the type of content provided. Like the different patterns found across offline offenders and offline crime types, these findings reinforce the need to investigate types of cybercrime separately and not assume that patterns will be the same across each.

An offender's ability to persist is impacted by a variety of personal (e.g., marriage) and societal (e.g., incarceration) factors. Within this dissertation, I examined website persistence and the individual factors that contribute to prolonged survival or early failure. Websites within CE-seeded networks, with verified CE images, survived longer than websites across all 30 networks, while early failures were comprised of small websites with minimal content. Within CE-seeded networks, baseline survival was influenced by the type of seed. Networks beginning with a blog-seed had lower survival rates while site-seed networks had higher. A potential explanation is that many blogs are hosted by third-party companies with specific terms of service (TOS). If a blog is found in violation of the TOS, it is removed. Conversely, sites are often hosted by companies without the same rigorous TOS. The removal of a site may be more difficult because of the additional steps needed: identify the hosting company, the location of the data, the laws of the country hosting the data, cooperation from the hosting company, and so on. These may result in fewer being removed, or it taking longer for their removal. Finally, incorporating facets of the (offline) criminal career, the website characteristic volume of

CE images (offending frequency) was linked to decreased survival while connectivity was linked to increased survival. Combined, these findings point to the importance of embeddedness for maintaining survival and the presence of large quantities of CE material increasing risk of failure.

The attributes and structure of websites, and the networks they are a part of, are important to the understanding of how cybercrimes are aided by websites. Not only does distinguishing illegal, and associated, websites from others assist with identification, but understanding how they function, communicate, distribute illegal material, and evolve has implications for creating proactive social control strategies. More broadly, these comparisons reinforce the importance of explaining criminal activities in the context of the surrounding social networks and how specific pathways, or trajectories, can be formulated based on patterns found within the networks.

5.2. Limitations

Within each study, I addressed key limitations specific to that study. However, I feel it is important to outline three general, dissertation-level, limitations. First, the issue of ethics of data collection is always of concern within research, especially in developing fields, such as SNA (Borgatti & Molina, 2005; Kadushin, 2005), and newly studied arenas, such as the Internet (Ess, 2013). Inherent within SNA is the inclusion of data on those being interviewed and those connected to the interviewee. Borgatti and Molina (2005) note that this complicates consent as data is being collected on both the individual and their surrounding network. Therefore, additional risks are taken by participants and the network (e.g., organization) being analyzed. Borgatti and Molina also note that SNA requires some level of individuality, which impedes anonymity at initial data collection. That is, forming the network requires that the members of the network are identified. Kadushin (2005) adds that network data collected by third parties (e.g., criminal surveillance) can contain inaccuracies, while respondents may identify those outside the confines of the subject-pool, whose inclusion would be done without their informed consent. For newly developing domains (e.g., Internet), the appropriate ethical practices are still being determined. Among cyber-researchers, ethical arguments also center on the general topic of consent. More specifically, Hewson (2003) points to

four key ethical concerns for IMR: a) obtaining informed consent; b) ensuring confidentiality for participants during data collection and storage; c) differentiating between public and private data; and d) effective methods for debriefing participants. For the purpose of this dissertation, I focus on the delineation between public and private data, and when is consent necessary?

The data for this dissertation were collected from public websites: websites that did not require registration or personal identification to access the majority of content. The primary reason for this was to ensure that the data were of public record and did not require consent to study. However, Hewson (2003) posit whether it is ever ethically justified to use publicly available data, if the data has not been voluntarily, and deliberately, made available. Included within that discussion is where the line between public and private Internet data is? By agreeing to a website's terms and conditions, does the information you post then become public record? For a publicly accessible website, one could argue that posters are voluntarily making their comments available to all. However, if polled it is quite possible that many posters would disagree with that statement, arguing that while the comment was in a public space, the intended use was not for research purposes and hence consent was not given for that function. Furthermore, if a private company is not allowed to collect data on a blog they host, even if viewing it requires no special permissions (i.e., publicly accessible), then when is data collected by a researcher ethically or unethically justified? Although important questions, the contextual application of these, with regards to ethical considerations, is still unclear. Given the alternative of accessing a much smaller sample or websites requiring special permissions, the scanning and use of public websites seemed like a suitable compromise to this debate. As Wolak, Finkelhor, and Mitchell (2005) found that only 20% of online CE offenders used any type of security to hide content, I felt that public websites still provided a representative view of the current CE distribution landscape. Nevertheless, the use of publicly accessible websites is a limitation in itself, as it is unclear what type of data is missing when excluding private websites. Therefore, the mechanisms that result in identification (Chapter 2), failure (Chapter 3) or community structure (Chapter 4) may differ for private websites, and networks.

The findings of any study are dependent on the reliability and validity of the data analyzed. The second limitation of this dissertation are the criteria used to include websites into the CE-seeded networks. A focal point of Chapter 2 was accessing the frequency of CE images and CE-related keywords in non-CE related networks (sports and sexuality). While CE images were found minimally within non-CE networks, and non-code keywords were more frequent in CE-seeded networks, the validity and reliability of the inclusion criteria are unclear. This lack of clarity is with regard to both false positives and false negatives. The studies within this dissertation were focused on describing the networks *surrounding* known CE websites because of the limitation of the inclusion criteria. Although some of the websites surrounding CE-seeds would fit most general criteria for being CE-focused, others would not. Equally important is that some websites were, most likely, examined by the webcrawler and deemed off-topic because they did not meet the criteria and yet were CE-related. Therefore, it is unclear how many websites were included or excluded, as a result of too broad or too narrow of inclusion criteria. Nevertheless, the findings within this dissertation are important to understanding CE distribution on the WWW. Like an accomplice offline does not need to be directly involved in a crime (Warr, 1996), a website does not need to contain CE material for it to be a key contributor. By connecting to websites with CE material, they are an accomplice and facilitator of CE distribution. Additionally, as the Internet is one large network (Wellman, Boase, & Chen, 2002), the data collected are only a representation of part of the global network. Therefore, the findings of this dissertation are limited to CE-seeds and their surrounding networks, up to a specific size (~300 websites), and should not be viewed as applying to all public CE websites. That is, one of the goals of this dissertation was to improve on the criteria used for identifying CE websites.

A central objective for this dissertation was to begin to develop a conceptual framework for analyzing the evolution of CE websites, built on the offline criminal career paradigm (Blumstein, Cohen, & Nagin, 1978). As the criminal career has yet to be applied to a virtual environment, an objective within this dissertation was to determine the best ways to operationalize criminal career dimensions for cyberspace, and for websites (entities). Adding to the complexity of this transition was the inherent anonymity of the Internet. Therefore, a third limitation of this dissertation is the operationalization of criminal career dimensions.

Discussed in Chapter 3, transitioning dimensions such as offending frequency to cyberspace, and further to websites, poses a problem as the multiplicity of illegal content on the Internet clouds the definition of frequency. Offline, an offender can only fence an item to one person. In cyberspace, an offender can advertise, and sell, their stolen credit card information to multiple people and websites. Also, that information can be maintained or reused (e.g., distribution of copyrighted music) multiple times. Therefore, offending frequency can be interpreted and operationalized a variety of ways. This same issue exists with transitioning other career dimensions such as co-offending (e.g., what criteria is necessary for a hyperlink to be considered a co-offending partnership?), and duration (e.g., when does the career of a website end?).

Adding to the complexity of interpreting criminal career concepts for websites is that many of the attributes used within criminal career research are unknown, or do not exist, online. For example, offline co-offenders typically share similarities in age, sex, ethnicity, residency, and other personal attributes (Sarnecki, 2001; Warr, 2002; Weerman, 2003). As these attributes are unknown online, it is unclear how co-offending partnerships are formed. This complicates inferences about what connects websites to one another. That is whether other attributes replace typical offline attributes, or whether connections are based on pure availability. Within this dissertation, I made decisions about how each criminal career dimension would be operationalized and with those decisions come potential limitations to the conclusions drawn. Although important to the findings within this dissertation, the overall goal is to spur the discussion of how best to operationalize offline concepts for a virtual environment. From these discussions, those best practices can be determined and the resulting limitations from the definitions used within this dissertation can be addressed.

5.3. Policy Implications

Two central themes emerged from this dissertation that have important practical and theoretical implications. First, the development and use of automated data collection tools, such as webcrawlers. Second, the development of a criminal career based theoretical framework for analyzing the distribution of CE material through website

entities. Implications for law enforcement, private organizations, CE researchers, and cybercrime researchers are discussed.

The graphic nature and quantity of CE material being distributed has been shown to increase psychological harm and burnout suffered by investigators within Integrated Child Exploitation (ICE) units. Research suggests that ICE officers suffer from secondary-traumatic stress disorder, emotional exhaustion, intrusive thoughts, and interpersonal/marriage problems (Bourke & Craun, 2014; Burns, Morley, Bradshaw, & Domene, 2008; Craun, Bourke, & Coulson, 2015; Krause, 2009; Perez, Jones, Englert, & Sachau, 2010). However, the impact of the graphic material regularly observed by officers is only part of the problem. Powell, Cassematis, Benson, Smallbone, and Wortley (2014) interviewed 32 ICE investigators and found that large workloads and insufficient time and resources were stressors resulting in burnout. The validation of automated data collection tools and the continued improvement and refinement to their criteria reliability, features, and overall performance, can aid in reducing investigator health costs. By automating the scanning and data collection process, officers could spend more time investigating cases rather than searching the Internet for material. Through automation, officers would also view less material for verification purposes and be able to focus website searches on, for example, previously catalogued material. Combined with social network analyses, the data collected can be used to improve prioritization strategies by finding key players/targets (Joffres, Bouchard, Frank, & Westlake, 2011; Westlake, Bouchard, & Frank, 2011).

For private companies, such as data hosting services, and watchdog organizations, such as National Center for Missing and Exploited Children, the development of reliable and valid automated search tool can be beneficial for monitoring services and/or websites. For data hosting companies, automated tools can allow a company to regularly scan websites they are hosting to ensure that the client is adhering to their terms of service. As with law enforcement, watchdog organizations could use webcrawlers to improve efficiency of detection and minimize personal contact with CE-related material. Of course, for both there are issues regarding privacy. Although arguments can be made, and have been made (Gibbs, 2014a), for the anonymous, automated, scanning and collecting of data by private companies, the manual process of

verifying an account believed to be in violation of their service agreement is a more complicated issue[39]. For watchdog organizations, privacy becomes more of an issue due to laws regarding accessing a person's private information without consent. Nevertheless, if issues pertaining to privacy can be overcome, webcrawlers can have important implications for non-law enforcement agencies looking to contribute to combatting CE material distribution.

The lack of academic research into the mechanisms of CE material distribution in cyberspace may be partly attributed to the laws surrounding possession (for researchers) and the process of viewing content. The development of automated data collection tools has implications for the study of online CE material distribution. First, the webcrawler (potentially) eliminates the element of possession in research, as it does not download the illegal content onto a hard drive. Instead it downloads the image into memory, calculates the hash value, checks it against the database, and then discards the content. Second, automation eliminates the need to ever view CE content as the webcrawler provides detailed information about webpages searched. This includes the URL of an image and the hash value associated with that image. By having this information, without viewing the image, researchers can plot the CE distribution chain of a network, and the steps taken by individual websites to hide content.

While this dissertation focused on the crime of CE distribution, my findings have implications for other forms of cybercrime. Webcrawlers have been suggested as effective tools for exploring and understanding different types of cybercrime, especially terrorism (Bouchard, Joffres, & Frank, 2014; Chen, 2012; Zhou et al., 2005). As public websites are used to distribute pirated material, drugs, fraudulently acquired data, and other services, automated data collection tools can aid in understanding online criminal processes and determine the similarities and differences across cybercrimes.

[39] One potential option for data hosting companies is to automatically search accounts and automatically remove those that meet a specific criteria threshold. However, automatic removal may result in too many false positives and loss of customers and credibility. A second option is to outline, in the terms of service, the threshold for subsequent manual verification. The concern here is that a vague outline of the minimal threshold would frustrate customers, resulting in mistrust, while a more detailed outline would provide a blueprint for how customers avoid detection. Ergo, despite being useful for online data hosting companies, current privacy concerns may limit the effectiveness of these tools for commercial ventures.

The findings in this dissertation have implications for the second main theme covered (criminal career) for law enforcement, child sexual exploitation researchers, and criminological researchers, more generally. Law enforcement strategies for combating the distribution of CE material in cyberspace focus on identifying and removing content. Although automated search tools assists with identification, the process is still retroactive. The findings within this dissertation, and subsequent research regarding how websites survive and communicate, assist with creating proactive policing strategies. More specifically, understanding the characteristics and evolution of websites, strategies used for detection avoidance, and how CE material is stored and distributed can aid in identifying future key players and prioritizations.

The transition of the criminal career paradigm to a virtual environment also has implications for all criminological research. First, the growth in cybercrime has necessitated the need for theories explaining crimes committed partly or solely online. Although theories have been developed to specifically explain cybercrime (e.g., Jaishankar, 2008), Holt and Bossler (2014) highlight the need for the application of traditional criminological theories to cybercrime. Within this dissertation I began the application of the criminal career framework to cybercrime. This application helps facilitate discussion surrounding how best to transition the criminal career, and life course theories, to cyberspace. It also has implications for understanding how criminal processes differ between offline and online crime and how IMR methods can be used to research dimensions of the criminal career that are difficult to study offline (e.g., co-offender selection processes). Second, there has been little conceptual development surrounding the collective/entity criminal career. The (re)introduction of the entity-based criminal career has implications for offline and online criminal career research. Conceptualizing (some) individual crime as being embedded within an entity and therefore influenced by the entity can modify the way we frame criminal processes and the factors that we consider when explaining crime. In addition, the entity criminal career opens up new avenues for research and for the application of new principles and dimensions to the criminal career paradigm. Online, the introduction of the entity into the understanding of cybercrime emphasizes the global nature of cybercrime and how their presence can heavily dictate the flow of goods, services, and information on the World Wide Web. It also underlines the arguments made offline that individual offenders cannot

be examined outside of the context of the entities they engage. Likewise, cybercriminals cannot be examined independently of the networks they create in cyberspace. Third, I provide a framework and baseline for understanding how CE material is distributed on public (and private) websites. This has implications for understanding distribution at both the macro and micro level. At the macro level, it provides information for how the larger network is structured and communicates; precursors to distribution. At the micro level, it provides a context for investigating individual concepts, For example, from website connectivity, researchers can begin to define the boundaries of the co-offender pool for individual offenders and pose research regarding how co-offenders are selected or not selected.

5.4. Future Research

Improvement to validity and reliability of webcrawlers for data collection require that addition criteria be included and the existing criteria is refined. Within this dissertation, I used an image hash value database provided by the Royal Canadian Mounted Police, based on cases investigated in Canada. Despite including more than 2.25 million hash values, the database used cannot be considered exhaustive of the existing material being distributed. Therefore, future research is required that incorporate additional databases from other sources. Further enhancement can come through those databases using some type of categorization (e.g., COPINE or SAP scale).

The findings from Chapter 2 showed that refinement to the keywords used are needed. Future research replicating the methods used within this dissertation but with refinement to the keywords used (including controlling for context) is required. The refinement process needs to include consultation with law enforcement and social control agencies to ensure that modifications reflect the current landscape. Comparing the revised findings to those from this dissertation, and investigating additional research questions, will provide insight into CE material distribution.

The advancement of video-capture technology suggests a potential future increase in the trafficking of CE videos. The process of building a video hash value database would address current and future growth in video distribution, and investigator

psychological health (secondary-traumatic stress disorder). To further minimize trauma, a form of digital video fingerprinting can be added to automated tools. However, the creation and use of any video databases as criteria require validation. Future research needs to compare the different criterion independently and jointly and examine the process of CE video distribution.

Detecting websites distributing CE material is only part of the issue. Another part is shutting down said websites. The process of removal is influenced by the physical location of the illegal material. Hosted websites are governed by the country where they reside. Therefore, subsequent improvements to the webcrawler are required, including identifying the hosting location of the material and any addition information about the website, such as who is the register. This additional information can aid law enforcement with targeting websites they have jurisdiction over, and forwarding information about websites to other law enforcement agencies, where they do not have jurisdiction.

As the webcrawler developed for this dissertation is revised, additional research into the structure and function of CE websites and their corresponding networks are required. First, a qualitative study of the structure of CE websites would be beneficial. This includes determining how obvious websites are with their intention. For example, do websites broadcast their CE focus on their homepage or do they attempt to hide their intentions and, if they do hide, what methods are used to hide the true nature. Building from this, another important area of research is to identify where CE material is located on a website and how that material is distributed throughout the network. For example, following a series of hash values found on a website and seeing the rate at which those hash values are found on connected websites. A key component of each of these research objectives is to collect a complete website structure and plot the evolution over time. Finally, while it is important to research public (and private) websites, the growth in public prominence of the Deep Web has increased it as a location where offenders venture to find CE material. It is very likely that in this dynamic, anonymous, environment, that the behaviours of websites differ considerably from websites found on the Surface Web. However, the dynamic nature of the Deep Web means that changes to automated data collection tools are required to deal with webpages that do not have a fixed location.

Along with the assessment of automated data collection techniques (inclusion criteria) and tools (webcrawler), a main objective within this dissertation was to begin the transition of the criminal career paradigm to cyberspace. This process included formulating how to operationalize offline concepts within a virtual framework. While online characteristics were incorporated additional operationalization is required, including the proposal of alternative definitions, and subsequent analyses, to the dimensions discussed and to other criminal career parameters. For example, I defined the duration of a website's career to be the time spent at one domain. However, a) I was unable to properly identify the onset of the criminal career (only estimating it); b) I only followed websites for a maximum of 60 weeks, measuring them every six weeks. Subsequent longitudinal research examining the duration of the entity criminal career would be useful as it could indicate any turning points (Sampson & Laub, 1993) early in a website's career that dictate its future trajectory. Adding to this, a more detailed understanding of how the initial operator influenced the subsequent career can be obtained, parsing out the role of the operator from the role of the website and/or community. Finally, as legal issues can result in the temporary removal of a website and relocation to a new address, extending the concept of the entity criminal career beyond a singular url would provide a more well-rounded understanding of the entity career and how 'life' events dictate, for example, crime-type mix. While these objectives can be accomplished through large data collection, similar to that within this dissertation, an alternative proposal is to follow several websites more in-depth, providing richer data. From this patterns can be formulated to be applied, or tested, on other websites, across types of cybercrime.

Introduced through the term 'entity criminal career', further research exploring the evolution of criminal organizations online and offline could provide insight into criminal patterns at the micro and macro levels. For example, the Islamic State of Iraq and ash-Sham (ISIS) was once known as al-Qaeda in Iraq. This represents a criminal entity (al-Qaeda) experiencing a form of fracturing resulting in offshoots with different agendas. Presumably this process was heavily influenced by the associated members. Therefore, as it is naïve to interpret the evolution of an offender without considering their associations, so to would it be naïve to ignore the role of influential members in the trajectory of a criminal entity, or, more importantly, how disagreements to the trajectory

of a criminal entity can facilitate the creation of new entities. Like the individual criminal career, the entity criminal career is complex, requiring different levels of analysis and incorporating a variety of factors.

While the focus within this dissertation was introducing the entity criminal career and developing said framework for a virtual environment, the transition of the criminal career paradigm to individual cybercriminals follows a similar path. Many of the conceptualizations of criminal career dimensions incorporated can be directly applied to individual offenders. In fact, this process may be simpler as the criminal career paradigm is targeted to explaining the evolution of individual offenders (e.g., co-offending). Therefore, additional research exploring the virtual criminal career of specific pseudonyms would aid in developing both the entity and singular cyber-criminal career framework. This research can incorporate social network analysis to explain how individual offenders interact with one another as well as with websites. More focused, research into the structure and interaction between offenders on a specific website and the associations of said website can provide the context required to better understand the interplay between entities and individuals and how each influences the other.

Finally, this dissertation focused on CE websites and their surrounding networks. Although the sexual exploitation of children on the Internet garners extensive attention from the media and policy makers, the criminal processes of other cybercrimes are also conducted using public and private websites. While there will be overlap in practices, the factors that govern the evolution of websites will be influenced by the cybercrime being investigated. For example, cybercrimes that garner less attention from law enforcement (e.g. phishing) may modify the security tactics employed, while cybercrimes that involve the trafficking of physical items (e.g., drugs or people) may function differently. For each, the concept of the entity criminal career may be more or less applicable while the incorporating of an offline component (i.e., exchanging a physical item) may also modify how the career evolves. In sum, the online criminal careers of entities involved in different types of cybercrimes need to be compared and contrasted to develop a well-rounded understanding of crimes committed using the Internet.

5.5. Summary

The distribution of CE material on the Internet can be investigated using automated data collection techniques and explained using concepts of the criminal career paradigm and social network analysis measures. The use of automated tools has benefits for social control and research, however, their effectiveness and accuracy is impacted by the ability to select appropriate inclusion criteria (Chapter 2). Provided due diligence is conducted when choosing identification criteria, the information collected can assist with creating typologies (e.g., crime-type mix), determining survival (Chapter 3), and identifying how sub-groups form and interaction within a larger network (Chapter 4).

Online CE distribution is a complex network of players that include individuals and entities, and the interactions between the two. The use of social network analysis to explain online criminal behaviour is necessary to completely understand the online criminal process (Chapters 2 and 4). Moreover, social network analysis techniques can be effective at distinguishing criminal websites from non-criminal websites as the two types appear to function differently on several key measures. While comparisons provide insight into how criminal-based websites compete and diffuse information throughout the entire network, and whether that differs from non-criminal networks (Chapter 2).

The studies presented within this dissertation open up multiple avenues for further exploration into the criminal processes of offenders and entities conducting activities solely in cyberspace. The methods used provide a framework for investigation into the evolution of websites facilitating any type of cybercrime. As such, future research applying the data collection techniques used here are necessary. In addition, the transition of criminal career concepts to CE distributions between websites highlight not only the ability to, but the importance of, transitioning other central criminological theories and frameworks to cybercrime. Further incorporation of criminal career dimensions to all types of cybercrime will aid in understanding online criminal processes and how they differ from offline processes. Moreover, it will provide insight into criminal processes that are more difficult to study in an offline setting, such as co-offender selection. From additional research into cybercrime, public policies and law enforcement

techniques can be further refined to address the ever-present and growing threat of cybercrime.

References

Ackland, R., & Shorish, J. (2009). Network formation in the political blogosphere: An application of agent based simulation and e-research tools. *Computational Economics, 34,* 383-398.

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery (pp. 36-43)*. New York, NY: DCM.

Akdeniz, Y. (2013). *Internet child pornography and the law: National and international responses.* Farnham, Surrey: Ashgate Publishing.

Albini, J. (1971). *The American Mafia: Genesis of a legend.* New York, NY: Meredith.

Alexy, E. M., Burgess, A. W., & Baker, T. (2005). Internet offenders: Traders, travelers, and combination trader-travelers. *Journal of Interpersonal Violence, 20,* 804-812.

Almutairi, A., Parish, D., & Phan, R. (2012). Survey of high interaction honeypot tools: Merits and short-comings. Retrieved from: http://www.cms.livjm.ac.uk/pgnet2012/Proceedings/Papers/1569604821.pdf.

Amirault, J., & Bouchard, M. (2015). A group-based recidivist sentencing premium? The role of context and cohort effects in the sentencing of terrorist offenders. *International Journal of Law, Crime, and Justice, online first.*

Andresen, M. A., & Felson, M. (2010). The impact of co-offending. *British Journal of Criminology, 50,* 66-81.

Andresen, M. A., & Felson, M. (2012). Co-offending and the diversification of crime types. *International Journal of Offender Therapy and Comparative Criminology 56,* 811 - 829.

Armstrong, H. L., & Forde, P. J. (2003). Internet anonymity practices in computer crime. *Information Management & Computer Security, 11,* 209-215.

Babchishin, K. M., Hanson, R. K., & Hermann, C. A. (2011). The characteristics of online sex offenders: A meta-analysis. *Sexual Abuse: A Journal of Research and Treatment, 23,* 92-123

Babchishin, K. M., Hanson, R. K., & van Zuylen, H. (2015). Online child pornography offenders are different: A meta-analysis of the characteristics of online and offline sex offenders against children. *Archives of Sexual Behavior, 44,* 45-66.

Bachmann, M. (2007). Lesson spurned? Reactions of online media music pirates to legal prosecution by the RIAA. *International Journal of Cyber Criminology, 2,* 213-227.

Ball, L. (2013). Automating social network analysis: A power tool for counter-terrorism. *Security Journal, online first*. doi: 10.1057/sj.2013.3.

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286,* 509-512.

Basamanowicz, J., & Bouchard, M. (2011). Overcoming the Warez paradox: Online piracy groups and situational crime prevention. *Policy & Internet, 3(2),* Article 4. doi: 10.2202/1944-2866.1125.

Basu, A. (2014). Social network analysis: A methodology for studying terrorism. In M. Panda, S. Dehuri, & G. Wang (Eds.), *Social networking: Mining visualization, and security (pp. 215-242).* Cham, Switzerland: Springer International Publishing.

Beech, A. R., Elliott, I. A., Birgden, A., & Findlater, D. (2008). The Internet and child sexual offending: A criminological review. *Aggression and Violent Behavior, 13,* 216-228.

Bergman, M. K. (2001). The Deep Web: Surfacing hidden value [White paper]. *Journal of Electronic Publishing,* 7(1), doi: http://dx.doi.org/10.3998/3336451.0007.104.

Blumstein, A., & Cohen, J. (1979). Estimation of individual crime rates from arrest records. *The Journal of Criminal Law and Criminology, 70,* 561-585.

Blumstein, A., Cohen, J., Nagin, D. (1978). *Deterrence and incapacitation: Estimating the effects of criminal sanctions on crime rates.* Washington, DC: National Academy of Sciences.

Blumstein, A., Cohen, J., Roth, J. A., & Visher, C. A. (1986). *Criminal careers and 'career criminals'.* Washington, DC: National Academy Press.

Borgatti, S. P., & Halgin, D. S. (2011). On network theory. *Organization Science*, *22*, 1168-1181.

Borgatti, S. P., & Molina, J. (2005). Toward ethical guidelines for network research in organizations. *Social Networks, 27,* 107-117.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for Windows: Software for social network analysis.* Harvard, MA: Analytic Technologies.

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

Bossler, A. M., & Burruss, G. W. (2011). The General Theory of Crime and computer hacking: Low self-control hackers? In T. Holt & B. H. Schell (Eds.), *Corporate hacking and technology-driven crime (pp. 38-67).* Hershey, PA: IGI Global.

Bossler, A. M., & Holt, T. J. (2010). The effect of self-control on victimization in the cyberworld. *Journal of Criminal Justice, 38,* 227-236.

Bossler, A. M., Holt, T. J., May, D. C. (2012). Predicting online harassment among a juvenile population. *Youth and Society, 44,* 500-523.

Bouchard, M., & Konarski, R. (2014). Assessing the core membership of a youth gang from its co-offending network. In C. Morselli (Ed.), *Crime and networks (pp. 81-93).* New York, NY: Routledge.

Bouchard, M., & Morselli, C. (2014). Opportunistic structures of organized crime. In L. Paoli (Ed.), *The Oxford handbook of organized crime (pp. 288-302).* New York, NY: Oxford University Press.

Bouchard, M., & Nash, R. (2014). *Researching terrorism and counter-terrorism through a network lens.* Retrieved from: http://library.tsas.ca/media/TSASWP14-01_Bouchard-Nash.pdf

Bouchard, M., & Ouellet, F. (2011). Is small beautiful? The link between risks and size in illegal drug markets. *Global Crime, 12,* 70-86.

Bouchard, M., & Spindler, A. (2010). Gangs, groups, and delinquency: Does organization matter? *Journal of Criminal Justice, 38,* 921-933.

Bouchard, M., Joffres, K., & Frank, R. (2014). Preliminary analytical consideration in designing a terrorism and extremism online network extractor. In V. K. Mago & V. Dabbaghian (Eds.), *Computational models of complex systems (pp. 171-184).* Cham, Switzerland: Springer International Publishing.

Bourke, M. L., & Craun, S. W. (2014). Secondary traumatic stress among Internet Crimes Against Children task force personnel. *Sexual Abuse: A Journal of Research and Treatment, 26,* 586-609.

Bovenkerk, F., Siegel, D., & Zaitch, D. (2003). Organized crime and ethnic reputation manipulation. *Crime, Law, and Social Change, 39,* 23-28.

Briggs, P., Simon, W. T., & Simonsen, S. (2011). An exploratory study of Internet-initiated sexual offenses and the chat room sex offender: Has the Internet enabled a new typology of sex offender? *Sexual Abuse: A Journal of Research and Treatment, 23,* 72-91.

Bright, D. A., & Delaney, J. J. (2013). Evolution of a drug trafficking network: Mapping changes in network structure and function across time. *Global Crime, 14,* 238-260.

Bruinsma, G., & Bernasco, W. (2004). Criminal groups and transnational illegal markets. *Crime, Law, and Social Change, 41,* 79-94.

Burns, C. M., Morley, J., Bradshaw, R., & Domene, J. (2008). The emotional impact on coping strategies employed by police teams investigating internet child exploitation. *Traumatology, 14,* 20-31.

Burris, V., Smith, E., & Strahm, A. (2000). White supremacist networks on the Internet. *Sociological Focus, 33,* 215-235.

Burruss, G. W., Bossler, A. M., & Holt, T. J. (2012). Assessing the mediation of a fuller social learning model on low self-control's influence on software piracy. *Crime and Delinquency, 59,* 1157-1184.

Burt, R. (1992). *Structural holes: The social structure of competition.* Cambridge, MA: Harvard University Press.

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology, 110,* 349-399.

Cain, K. C., Harlow, S. D., Little, R. J., Nan, B., Yosef, M., Taffe, J. R., & Eilliot, M. R. (2011). Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. *American Journal of Epidemiology, 173,* 1078-1084.

Callanan, C., Gercke, M., De Marco, E., & Dries-Ziekenheiner, H. (2009). *Internet blocking*. Aconite Internet Solutions.

Canadian Centre for Justice Statistics (1999). *Organized crime activity in Canada, 1998: Results of a 'pilot' survey of 16 police services.* Ottawa, ON: Statistics Canada.

Canadian Criminal Code, RSC. (1985). c C-46 s163.

Carolan, B. V. (2013). *Social network analysis and education.* Thousand Oaks, CA: SAGE Publication, Inc.

Carr, A. (2004). *Internet traders of child pornography and other censorship offenders in New Zealand.* Wellington, NZ: Department of Internal Affairs. Retrieved from http://www.dia.govt.nz/Pubforms.nsf/URL/entirereport.pdf

Carrington, P. J. (2002). Group crime in Canada. *Canadian Journal of Criminology and Criminal Justice, 44,* 277-315.

Carrington, P. J. (2009). Co-offending and the development of the delinquent career. *Criminology, 47,* 1295-1329.

Carrington, P. J. (2014). The structure of age homophily in co-offending groups. *Journal of Contemporary Criminal Justice, online first.* doi: 10.1177/1043986214553376.

Carrington, P. J., Brennan, S., Matarazzo, A., & Radulescu, M. (2013). Co-offending in Canada, 2011. *Juristat 33(3).* Statistics Canada Catalogue no. 85-002-X. Ottawa, ON; Statistics Canada.

Chang, L., & Krosnick, J. A. (2009). National surveys via rdd telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly, 73,* 641-678. doi: 10.1093/poq/nfp075.

Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *Information Security in the Knowledge Economy, 65,* 57-70.

Chau, M., & Xu, J. (2008). Using web mining and social network analysis to study the emergence of cyber communities in blogs. *Terrorism Informatics, 18,* 473-494.

Chau, M., Shiu, B., Chan, I., & Chen, H. (2007). Redips: Backlink search and analysis on the Web for business intelligence analysis. *Journal of the American Society for Information Science and Technology, 58,* 351-365.

Chen, H. (2012). *Dark Web: Exploring and data mining the dark side of the web.* New York, NY: Springer.

Choi, K. C. (2008). Computer crime victimization and integrated theory: An empirical assessment. *International Journal of Cyber Criminology, 2,* 308-333.

Chow-White, P. A. (2006). Race, gender, and sex on the net: Semantic networks of selling and storytelling sex tourism. *Media Culture Society, 28,* 883-905. doi: 10.1177/0163443706068922.

Christin, N. (2013). Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd International Conference on World Wide Web (pp. 213-224).* Rio de Janeiro, Brazil.

Chu, B., Holt, T. J., Ahn, G. J. (2010). Examining the creation, distribution, and function of malware on-line. Washington, DC: National Institute of Justice. NIJ Grant No. 2007-IJ-CX-0018.

Clarke, R. V. G. (1997). *Situational crime prevention.* Monsey, NY: Criminal Justice Press.

Cleves, M., Gould, W., Gutierrez, R., & Marchenko, Y. (2010*). An introduction to survival analysis using Stata* (3rd ed.). College Station, TX: Stata Press.

Cohen, J. (1986). Research on criminal careers: Individual frequency rates and offense seriousness. In A. Blumstein, J. Cohen, J. A. Roth, & C. Visher (Eds.), *Criminal careers and 'career' criminals, Vol. 1, (pp.292-418).* Washington, DC: National Academy Press.

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, 30,* 248-275.

Craun, S. W., Bourke, M. L., & Coulson, F. N. (2015). The impact of Internet crimes against children work on relationships with families and friends: An exploratory study. *Journal of Family Violence, 30,* 393-402.

D'Alessio, S. J., & Stolzenberg, L. (2008). Do cities influence co-offending? *Journal of Criminal Justice, 38,* 711-719.

de Amorim, S. G., Barthelemy, J. P., & Ribeiro, C. C. (1992). Clustering and clique partitioning: Simulated annealing and Tabu search approaches. *Journal of Classification, 9*, 17-41.

de la Cruz, I. P., Aller, C. F., Garcia, S. S., & Gallardo, J. C. (2010). A careful design for a tool to detect child pornography in P2P networks. In *Proceedings of the 2010 IEEE International Symposium on Technology and Society (pp. 227-233).* New South Wales, Australia.

De Maeyer, J. (2013). Towards a hyperlinked society: A critical review of link studies. *New Media & Society, 15,* 737-751.

Décary-Hétu, D., & Dupont, B. (2012). The social network of hackers. *Global Crime, 13,* 160-175.

Décary-Hétu, D., & Dupont, B. (2013). Reputation in a dark network of online criminals. *Global Crime, 14*, 175-196.

Décary-Hétu, D., Morselli, C., & Leman-Langlois, S. (2012). Welcome to the scene: A study of social organization and recognition among warez hackers. *Journal of Research in Crime and Delinquency, 49*, 359–382.

DeLisi, M., & Piquero, A. R. (2011). New frontiers in criminal careers research, 2000-2011: A state-of-the-art review. *Journal of Criminal Justice, 39,* 289-301.

Denney, A. S., & Tewksbury, R. (2013). Characteristics of successful personal ads in a BDSM on-line community. *Deviant Behavior, 34,* 153-168.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method* (2nd ed.). New York, NY: John Wiley and Sons.

Dolliver, D. S. (2015). Evaluating drug trafficking on the TOR network: Silk Road 2, the sequel. *The International Journal of Drug Policy, online first.* doi: 10.1016/j.drugpo.2015.01.008.

Drachen, A., & Veitch, R. W. D. (2013). Patterns in the distribution of digital games via BitTorrent. *International Journal of Advanced Media and Communication, 5,* 80-99.

Dupont, B. (2013). Skills and trust: A tour inside the hard drives of computer hackers. In C. Morselli (Ed.), *Crime and networks (pp. 195-217).* New York, NY: Routledge.

Durkin, K., Forsyth, C. J., & Quinn, J. F. (2006). Pathological Internet communities: A new direction for sexual deviance research in a post modern era. *Sociological Spectrum, 26,* 595-606.

Dykstra, J., & Sherman, A. T. (2013). Design and implementation of FROST: Digital forensic tools for OpenStack cloud computing platform. *Digital Investigation, 10,* S87-S95. doi: 10.1016/j.diin.2013.06.010.

Elliott, D. S. (1994). Serious violent offenders: Onset, developmental course, and termination-the American Society of Criminology 1993 Presidential Address. *Criminology, 32,* 1-21.

Elliott, I. A., Beech, A. R., Mandeville-Norden, R. (2013). The psychological profiles of Internet, contact, and mixed Internet/contact sex offenders. *Sexual Abuse: A Journal of Research and Treatment, 25,* 3-20.

Elliott, I. A., Beech, A. R., Mandeville-Norden, R., & Hayes, E. (2009). Psychological profiles of Internet sexual offenders: Comparison with contact sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 21,* 76-92.

Ess, C. (2013). *Digital media ethics (2nd Edition).* Cambridge, UK: Polity Press.

Estes, R. J. (2001). *The sexual exploitation of children: A working guide to the empirical literature.* Philadelphia, PA: National Institute of Justice.

Evans, R. D., Forsyth, C. J., & Wooddell, G. (2000). Macro and micro views of erotic tourism. *Deviant Behavior, 21,* 537-550.

Faber, W., Mostert, S., Faber, J., & Vrolijk, N. (2010). *Phising, kinderporno en advance-fee Internet fraud.* Faber Organisatievernieuwing.

Farrell, D., & Peterson, J. C. (2010). The growth of Internet research methods. *Sociological Inquiry, 80,* 114-125.

Feld, S. L. (1981). The focused organization of social ties. *American Journal of Sociology, 86,* 1015-1035.

Felson, M. (2003). The process of co-offending. In M. J. Smith & D. B. Cornish (Eds.), *Theory for practice in situational crime prevention (pp. 149-167).* Monsey, NY: Criminal Justice Press.

Finckenaueur, J. O., & Waring, E. J. (1998). *Russian Mafia in America: Immigration, culture, and crime.* Boston, MA: Northeastern University Press.

Foot, K. A., Schneider, S. M., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 U.S. electoral web sphere. *Journal of Computer-Mediated Communities, 8.* doi: 10.1111/j.1083-6101.2003.tb00220.x.

Fortin, F., & Corriveau, P. (2015). *Who is Bob_34?* Vancouver BC: UBC Press.

Fournier, R., Cholez, T., Latapy, M., Chrisment, I., Magnien, C., Festor, O., & Daniloff, I. (2014). Comparing Pedophile Activity in Different P2P Systems. *Social Sciences, 3,* 314-325.

Frank, R., Westlake, B. G., & Bouchard, M. (2010). The structure and content of online child exploitation. In *Proceedings of the 16th ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD 2010)*, Article 3. Washington, DC.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarifications. *Social Networks, 1,* 215-239.

Freiburger, T., & Crane, J. S. (2008). A systematic examination of terrorist use of the internet. *International Journal of Cyber Criminology*, *2*, 309-319.

Fu, T., Abbasi, A., & Chen, H. (2010). A focused crawler for Dark Web forums. *Journal of the American Society for Information Science and Technology*, *61*, 1213-1231.

Gallupe, O., & Bouchard, M. (in press). The influence of positional and experienced social benefits on the relationship between peers and alcohol use. *Rationality and Society.*

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, *3*(1). doi: 10.1111/j.1083-6101.1997.tb00062.x

Gibbs, S. (2014a, April 15). Gmail does scan all emails, new Google terms clarify. *The Guardian.* Retrieved from: http://www.theguardian.com/technology/2014/apr/15/gmail-scans-all-emails-new-google-terms-clarify.

Gibbs, S. (2014b, December 10). Swedish police raid sinks The Pirate Bay. *The Guardian*. Retrieved from: http://www.theguardian.com/technology/2014/dec/10/swedish-police-raid-pirate-bay.

Gillespie, A. A. (2011). *Child pornography: Law and policy.* New York, NY: Routledge.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99,* 7821-7826.

Glover, F., & Laguna, M. (1998). Tabu search. In D.Z. Du and P.M. Pardalos (Eds.), *Handbook of Combinatorial Optimization (pp. 621-757).* Dordrecht, NL: Kluwer Academic Publishers.

Godson, R., & Olson, W. J. (1995). International organized crime. *Society, 32,* 18-29.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66,* 877-902.

Grabosky, P. N. (2001). Virtual criminality: old wine in new bottles? *Social & Legal Studies, 10,* 243-249.

Grabosky, P., Smith, R. G., & Dempsey, G. (2001). *Electronic theft: Unlawful acquisition in cyberspace.* Cambridge, UK: Cambridge University Press.

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika, 81,* 515-526.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78,* 1360-1380.

Haller, M. H. (1990). Illegal enterprise: a theoretical and historical interpretation. *Criminology, 28,* 207–235.

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods.* Riverside, CA: University of Riverside.

Hardy, R. L., & Kreston, S. S. (2004). Geeks with guns, or how I stopped worrying and learned to love computer evidence. Paper presented at the *South African Professional Society on the Abuse of Children National Conference*, Pretoria, South Africa. Retrieved from: http://www. sapsac. co. za/geeks. pdf.

Hargittai, E., Gallo, J., & Kane, M. (2008). Cross-ideological discussion among conservative and liberal bloggers. *Public Choice, 134,* 67-86.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in Medicine, 15,* 361-387.

Harris, D. A., Smallbone, S., Dennison, S., & Knight, R. A. (2009). Specialization and versatility in sexual offenders referred for civil commitment. *Journal of Criminal Justice, 37,* 37-44.

Hartman, C. R., Burgess, A. W., & Lanning, K. V. (1984). Typology of collectors. In A. W. Burgess & M. L. Clark (Eds.), *Child pornography and sex rings (pp. 93-109).* Toronto, ON: Lexington Books.

Haynie, D. L. (2001). Delinquent peers revisited: Does network structure matter? *American Journal of Sociology, 106,* 1013-1057.

Haynie, D. L. (2002). Friendship networks and delinquency: The relative nature of peer delinquency. *Journal of Quantitative Criminology, 18,* 99-134.

Heckathorn, D. D. (2007). Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology, 37,* 151-207.

Hertz-Picciotto, I., & Rockhill, B. (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics, 53,* 1151-1156.

Hewson, C. (2003). Is the Internet a viable research tool? In C. Hewson (Ed.), *Internet research methods (pp. 26-56).* London, UK: SAGE Publications, Ltd.

Hewson, C., & Laurent, D. (2008). Research design and tools for Internet Research. In N. Fielding, M. Raymond, & G. Blank (Eds.), *The SAGE Handbook of Online Research Methods* (pp. 58-79). London, UK: SAGE Publications, Ltd.

Hewson, C., Laurent, D., & Vogel, C. M. (1996). Proper methodologies for psychological and sociological studies conducting via the Internet. *Behavior Research Methods, Instruments, and Computers, 32,* 186-191.

Hewson, C., Yule, P., Laurent, D., & Vogel, C. M. (2003). *Internet research methods: A practical guide for the social and behavioral sciences.* London, UK: Sage.

Higgins, G. E., & Marcum, C. D. (2011). *Digital piracy: An integrated theoretical approach.* Raleigh, NC: Carolina Academic Press.

Hillman, H., Hooper, C., & Choo, K. R. (2014). Online child exploitation: Challenges and future research directions. *Computer Law & Security Review, 30,* 687-698.

Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, *14*, 206-221.

Hoechle, D. (2007). Robust standard errors for panel regressions with cross-sectional dependence. *Stata Journal, 7,* 281-312.

Hogan, B. (2008). Analyzing social networks via the Internet. In N. Fielding, R. Lee, & G. Blank (Eds.), *The SAGE handbook of online research methods* (pp. 141-161). London, UK: Sage.

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

Holt, T. J. (2009). Lone hacks or group cracks: Examining the social organization of computer hackers. In F. Smalleger & M. Pittaro (Eds.), *Crimes of the Internet (pp. 336-355).* Upper Saddle River, NJ: Pearson Prentice Hall.

Holt, T. J. (2012). Exploring the intersections of technology, crime, and terror. *Terrorism and Political Violence, 24,* 337-354.

Holt, T. J. (2013). Examining the forces shaping cybercrime markets online. *Social Science Computer Review, 31,* 165-177.

Holt, T. J., Blevins, K. R. (2007). Examining sex work from the client's perspective: Assessing johns using online data. *Deviant Behavior, 28,* 333-354.

Holt, T. J., Blevins K. R., & Burkert, N. (2010). Considering the pedophile subculture on-line. *Sexual Abuse: Journal of Research and Treatment, 22,* 3-24.

Holt, T. J., & Bossler, A. M. (2014). An assessment of the current state of cybercrime scholarship. *Deviant Behavior, 35,* 20-40.

Holt, T. J., Bossler, A. M., & May, D. C. (2012). Low self-control deviant peer associations and juvenile cyberdeviance. *American Journal of Criminal Justice, 37,* 378-395.

Holt, T. J., Burruss. G. W., Bossler, A. M. (2010). Social learning and cyber deviance: Examining the importance of a full social learning model in the virtual world. *Journal of Crime and Justice, 33,* 15-30.

Holt, T. J., & Lampke, E. (2010). Exploring stolen data markets on-line: Products and market forces. *Criminal Justice Studies, 23,* 33-50.

Holt, T. J., Strumsky, D., Smirnova, O., & Kilger, M. (2012). Examining the social networks of malware writers and hackers, *International Journal of Cyber Criminology, 6,* 891-903.

Holt, T. J., & Turner, M. G. (2012). Examining risks and protective factors of online identify theft. *Deviant Behavior, 33,* 308-323.

Howard, P. N., & Jones, S. (2004). *Society online: The Internet in context*. Thousand Oaks, CA: Sage Publication, Inc.

Hsu, C., & Lin, J. (2008). Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation. *Information & Management, 45,* 65-74.

Hughes, D. M. (2002). The use of new communications and information technologies for sexual exploitation of women and children. *Hastings Women's LJ*, *13*, 127-325.

Hughes, J. (2012). Authenticity and identity Internet contexts. In J. Hughes (Ed.), *SAGE Internet research methods* (pp. 95-127). London, UK: SAGE Publications, Ltd.

Hurley, R., Swagatika, P., Soroush, H., Walls, R. J., Albrecht, J., Cecchet, E., … Wolak, J. (2013). Measurement and analysis of child pornography trafficking on P2P networks. In *Proceedings of the 22nd International Conference on World Wide Web (pp. 631-642).* Geneva, Switzerland.

Ingram, J. R., Hinduja, S. (2008). Neutralizing music piracy: An empirical examination. *Deviant Behavior, 29,* 334-366.

Internet Watch Foundation (2014). *IWF Operational Trends 2014.* Retrieved from https://www.iwf.org.uk/resources/trends.

Iqbal, F., Fung, B. C. M., & Debbabi, M. (2012). Mining criminal networks from chat log. *2012 IEEE/WIC/ACM International Conferences on* (Vol. 1, pp. 332-337). doi: 10.1109/WI-IAT.2012.68.

Jackson, M. H. (1997). Assessing the structure of communication on the World Wide Web. *Journal of Computer-Mediated Communication, 3(1),* 0. doi: 10.1111/j.1083-6101.1997.tb00063.x

Jaishankar, K. (2008). Space transition theory of cyber crimes. In F. Schmalleger & M. Pittaro (Eds.), *Crimes of the Internet (pp. 283-301).* Upper Saddle River, NJ: Prentice Hall.

Jenkins, P. (2003). *Beyond tolerance: Child pornography on the Internet.* New York, NY: New York University Press.

Joffres, K., Bouchard, M., Frank, R., & Westlake, B.G. (2011). Strategies to disrupt online child pornography networks. In *Proceedings of the 2011 EISIC - European Intelligence and Security Informatics* (pp. 163-170). Athens, Greece.

Jordan, T., & Taylor, P. (1998). A sociology of hackers. *Sociological Review, 46*, 757-780.

Kadushin, C. (2005). Who benefits from network analysis: Ethics of social network research. *Social Networks, 27,* 139-153.

Kanich, C., Chachra, N., McCoy, D., Grier, C., Wang, D., Motoyama, M., Levchenko, K., Savage, S., & Voelker, G. (2011). No plan survives contact: Experience with cybercrime measurement. In *CSET'11 Proceedings of the 4th Conference on Cyber Security Experimentation and Test, Article 2.* Berkley, CA.

Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication, & Society, 15*, 639-661. doi: 10.1080/1369118X.2012.665468.

Kenney, M. (2007). The architecture of drug trafficking: network forms of organisation in the Colombian cocaine trade. *Global Crime, 8*, 233-259.

Kontostathis, A., Edwards, L., & Leatherman, A. (2010). Text mining and cybercrime. In M. Berry & J. Kogan (Eds.), *Text mining: Applications and theory* (pp. 149-164). Chichester, UK: John Wiley & Sons.

Kotrla, K. (2010). Domestic minor sex trafficking in the United States. *Social work*, *55*, 181-187.

Krause, M. (2009). Identifying and managing stress in child pornography and child exploitation investigators. *Journal of Police and Criminal Psychology, 24,* 22-29.

Kreager, D. A., Rullison, K., & Moody, J. (2011). Delinquency and the structure of adolescent peer groups. *Criminology, 49,* 95-127.

Krohn, M. D., Massey, J. L., & Zielinski, M. (1988). Role overlap, network multiplexity, and adolescent deviant behavior. *Social Psychology Quarterly, 51,* 346-356.

Krone, T. (2004). A typology of online child pornography offending. *Trends and Issues in Crime and Criminal Justice, 279,* 1-6.

Kumar, R. A., & Kaliyaperumal, G. (2012). Optimal fingerprint scheme for video on demand using block designs. *Multimedia Tools and Applications, 61,* 389-418.

Kwan, G. C. E., & Skoric, M. M. (2013). Facebook bullying: An extension of battles in school. *Computers in Human Behavior, 29,* 16-25.

Lanning, K. V. (2001). *Child molesters: A behavioral analysis.* Alexandria, VA: National Center for Missing and Exploited Children.

Latapy, M., Magnien, C., & Fournier, R. (2013). Quantifying paedophile activity in a large P2P system. *Information Processing & Management, 49,* 248-263.

Laub, J. H., & Sampson, R. J. (2003). *Shared beginnings, divergent lives: Delinquent boys to age 70.* Harvard, MA: Harvard University Press.

Layton, R., Watters, P., & Dazeley, R. (2011). Automatically determining phishing campaigns using the USCAP methodology. In *eCrime Researchers Summit, 2010* (pp. 1-8). Los Alamitos, CA.

Leary, M.G. (2007) Self-produced child pornography: The appropriate societal response to juvenile self-sexual exploitation. *Virginia Journal of Social Policy & the Law 15*(1), 1-50.

Leary, M.G. (2009) Sexting or self-produced child pornography? The dialog continues - Structured prosecutorial dis cretion within a multi-disciplinary response. *Virginia Journal of Social Policy & the Law 17*(3), 486-566.

LeBlanc, M., & Loeber, R. (1998). Developmental criminology update. *Crime and Justice, 23,* 115-198.

LeGrand, B., Guillaume, J., Latapy, M., & Magnien, C. (2009). Technical report on Dynamics of Paedophile Keywords in eDonkey Queries. Measurement and analysis of P2P activity against paedophile content project. Retrieved from: http://antipaedo.lib6.fr/.

Leheny, D. (1995). A political economy of Asian sex tourism. *Annals of Tourism Research*, *22*, 367-384.

Lewandowski, J. L. (2003). Stepping off the sidewalk. *Journal of School Violence, 2,* 19-63.

Lijoi, A., & Nipoti, B. (2014). A class of hazard rate mixtures for combining survival data from different experiments. *Journal of the American Statistical Association, 109,* 802-814.

Loeber, R., & Farrington, D. P. (1998). *Serious and violent juvenile offenders: Risk factors and successful interventions*. Thousand Oaks, CA: Sage Publications.

Loeber, R., Wung, P., Keenan, K., Giroux, B., Stouthamer-Loeber, M., Van Kammen, W. B., & Maugham, B. (1993). Developmental pathways in disruptive child behavior. *Development and Psychopathology, 5*, 103-133.

Lussier, P. (2005). The criminal activity of sexual offenders in adulthood: Revisiting the specialization debate. *Sexual Abuse: A Journal of Research and Treatment, 17,* 269-292.

Lussier, P., Bouchard, M., Beauregard, E. (2011). Patterns of criminal achievement in sexual offending: Unravelling the 'successful' sex offender. *Journal of Criminal Justice, 39,* 433-444.

Lussier, P., Tzoumakis, S., Cale, J., & Amirault, J. (2010). Criminal trajectories of adult sex offenders and the age effect: Examining the dynamic aspect of offending in adulthood. *International Criminal Justice Review, 20,* 147-168.

Maimon, D., Alper, M., Sobesto, B., & Cukier, M. (2014). Restrictive deterrent effects of a warning banner in an attacked computer system. *Criminology, 52,* 33-59.

Malm, A., & Bichler, G. (2011). Networks of collaborating criminals: Assessing the structural vulnerability of drug markets. *Journal of Research in Crime and Delinquency*, *48*, 271-297.

Malm, A., Bichler, G., & Nash, R. (2011). Co-offending between criminal enterprise groups. *Global Crime, 12,* 112-128.

Mann, C., & Stewart, F. (2012). Power issues in Internet Research. In J. Hughes (Ed.), *SAGE Internet research methods* (pp.49-73). London, UK: SAGE Publications, Ltd.

Marín, J. M. F., Naranjo, J. Á. M., & Casado, L. G. (2015). Honeypots and Honeynets: Analysis and Case Study. In M. M. Cuz-Cunha & R. M. Portela (Eds.), *Handbook of research on digital crime, cyberspace security, and information assurance* (pp. 452-482). Hershey, PA: IGI Global.

Martinez-Torres, M. R., Toral, S. L., Palacios, B., & Barrero, F. (2011). Web site structure mining using social network analysis. *Internet Research, 21,* 104-123. doi: http://dx.doi.org/10.1108/10662241111123711.

Mazumdar, T. (2 January 2015). "The Pirate Bay Homepage Shows A Countdown Timer Set To February 1, 2015". *International Business Times*. Retrieved from: http://au.ibtimes.com/pirate-bay-homepage-shows-countdown-timer-set-february-1-2015-1405451.

McAndrew, M. A. (2000). *Presidential campaigns, political messages and the Internet: An analysis of communication and design strategies on Web sites with political content* (Doctoral dissertation). Georgetown University: Washington D.C.

McCarthy, B., & Hagan, J. (1995). Getting into street crime: The structure and process of criminal embeddedness. *Social Science Research, 24,* 63-95.

McCarthy, B., Hagan, J., & Cohen, L. E. (1998). Uncertainty, cooperation and crime: Understanding the decision to co-offend. *Social Forces, 77,* 155-184.

McCord, J., & Conway, K.P. (2002). Patterns of juvenile delinquency and co-offending. In R. Waring & D. Weisburd (Eds.), *Crime and social organization (pp. 15-30).* New Brunswick, NJ: Transaction Publishers.

McGloin, J. (2005). Policy and intervention consideration of a network analysis of street gangs. *Criminology & Public Policy, 4,* 607-635.

McGloin, J. M., & Nguyen, H. (2013). The importance of studying co-offending networks for criminological theory and policy. In C. Morselli (Ed.), *Crime and networks (pp. 13-27).* New York, NY: Routledge.

McGloin, J. M., & Piquero, A. R. (2010). On the relationship between co-offending network redundancy and offending versatility. *Journal of Research in Crime and Delinquency, 47,* 63-90.

McGloin, J. M., & Stickle, W. P. (2011). Influence or convenience? Rethinking the role of peers for chronic offenders. *Journal of Research in Crime and Delinquency, 48,* 419-447.

McGloin, J. M., Sullivan, C. J., Piquero, A. R., & Bacon, S. (2008). Investigating the stability of co-offending and co-offenders among a sample of youth offenders. *Criminology, 46,* 155-188.

McGuire, M., & Dowling, S. (2013). Cyber-dependent crimes. In *Cyber crime: A review of the evidence.* Home Office Research Report 75.

McPherson, J. M., & Smith-Lovin, L. (1987). Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociology Review, 52*, 370-379.

Medina, A., Matta, I., & Byers, J. (2000). On the origin of power laws in Internet topologies. *ACM SIGCOMM Computer Communication Review*, *30*(2), 18-28.

Miethe, T. D., Olson, J., & Mitchell, O. (2006). Specialization and persistence in the arrest histories of sex offenders: A comparative analysis of alternative measures and offense types. *Journal of Research in Crime and Delinquency, 43*, 204-229.

Milrod, C., & Monto, M. A. (2012). The hobbyist and the girlfriend experience: Behaviors and preferences of male customers of Internet sexual service providers. *Deviant Behavior, 33,* 792-810. doi: 10.1080/01639625.2012.707502.

Milrod, C., & Weitzer, R. (2012). The intimacy prism: Emotion management among the clients of escorts. *Men and Masculinities, 15,* 447-467.

Miranda Gonzalez, F. J., & Banegil Palacios, T. M. (2004). Quantitative evaluation of commercial web sites. *International Journal of Information Management, 24,* 313-328.

Mitchell, K. J., Finkelhor, D., Jones, L. M., & Wolak, J. (2010). Use of social networking sites in online sex crimes against minors: An examination of national incidence and means of utilization. *Journal of Adolescent Health, 47,* 183-190.

Mitchell, K. J., Finkelhor, D., & Wolak, J. (2003). The exposure of youth to unwanted sexual material on the Internet. *Youth & Society, 34,* 330-358.

Mitchell, K. J., Wolak, J., & Finkelhor, D. (2008). Are blogs putting youth at risk for online sexual solicitation or harassment? *Child Abuse & Neglect, 32,* 277-294.

Mitchell, K. J., Wolak, J., Finkelhor, D., & Jones, L. (2011). Investigators using the Internet to apprehend sex offenders: Findings from the Second National Juvenile Online Victimization Study. *Police Practice and Research: An International Journal, 13,* 267-281.

Moffitt, T. E. (1993). Life-course persistent and adolescent limited antisocial behavior: A developmental taxonomy. *Psychological Review, 100,* 674-701.

Monto, M. A., & Milrod, C. (2014). Ordinary or peculiar men? Comparing the customers of prostitutes with a nationally representative sample of men. *International Journal of Offender Therapy and Comparative Criminology, 58,* 802-820.

Moody, J., & White, D. R. (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review, 68,* 103-127.

Moore, R., Guntupalli, N. T., & Lee, T. (2010). Parental regulation and online activities: Examining factors that influence a youth's potential to become a victim of online harassment. *International Journal of Cyber Criminology, 4,* 685-698.

Moores, T. T., & Esichaikul, V. (2011). Socialization and software piracy: A study. *Journal of Computer Information Systems*, *51*(3), 1-9.

Morselli, C. (2009). *Inside criminal networks.* New York, NY: Springer.

Morselli, C. (2010). Assessing vulnerable and strategic positions in a criminal network. *Journal of Contemporary Criminal Justice*, *26*, 382-392.

Morselli, C., & Roy, J. (2008). Brokerage qualifications in ringing operations. *Criminology, 46,* 71-98.

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

Morselli, C., Tremblay, P, & McCarthy, B. (2006). Mentors and criminal achievement. *Criminology, 44,* 17-43.

Motoyama, M., McCoy, D., Levchenko, K., Savage, S., Voelker, G. M. (2011). An analysis of underground forums. *Internet Measurement Conference,* 71-79.

Moule, R. K., Pyrooz, D. C., & Decker, S. H. (2014). Internet adoption and online behaviour among American street gangs. *British Journal of Criminology, 54,* 1186-1206.

Nagin, D. S. (2008). Thoughts on the broader implications of the 'miracle of the cells'. *Criminology & Public Policy, 7,* 37-42.

Newman, G. R., & Clark, R. V. (2003). *Superhighway Robbery.* New York, NY: Routledge.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structures in networks. *Physical Review, 69.* doi: .1103/PhysRevE.69.026.113.

Nhan, J. (2013). The evolution of online piracy: Challenge and response. In T. J. Holt (Ed.), *Crime on-line: Causes correlates, and context, (pp. 61-80).* Raleigh, NC: Carolina Academic Press.

O'Connell, R. (2001). Paedophiles networking on the Internet. In C. A. Arnaldo (Ed.), *Child abuse on the Internet: Ending the silence (pp. 65-80).* Oxford, UK: Berghahn.

O'Halloran, E., & Quayle, E. (2010). A content analysis of a 'boy love' support forum: Revisiting Durkin and Bryant. *Journal of Sexual Aggression, 16,* 71-85.

O'Neill, R. A. (2001). *International trafficking in women to the United States: A contemporary manifestation of slavery and organized crime.* Retrieved from: https://www.cia.gov/library/center-for-thestudy-of-intelligence/csi-publications/books-and-monographs/trafficking.pdf.

Palfrey, J., Boyd, D., & Sacco, D. (2010). *Enhancing child safety and online technologies.* Durham, NC: Carolina Academic Press.

Pandit, S. J., Kulkarni, M. V., & Dhore, M. L. (2009). Image comparison search engine based on traditional and improved fractal encoding techniques. *International Journal of Recent Trends in Engineering, 2,* 108-110.

Papachristos, A. V. (2009). Murder by structure: Dominance relations and the social structure of gang homicide. *American Journal of Sociology, 115*, 74-128.

Papachristos, A. V. (2011). The coming of a networked criminology. *Measuring Crime & Criminality: Advances in Criminological Theory, 17,* 101-140.

Papachristos, A. V., & Smith, C. M. (2014). The embedded and multiplex nature of Al Capone. In C. Morselli (Ed.), *Crime and networks (pp.97-115).* New York, NY: Routledge.

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections, 25,* 49-61.

Park, H. W., & Thelwall, M. (2003). Hyperlink analyses of the World Wide Web: A review. *Journal of Computer-Mediated Communication, 8(4),* 0. doi: 10.1111/j.1083-6101.2003.tb00223.x

Park, H. W., Kim, C. S., Barnett, G. (2004). Socio-communicational structure among political actors on the Web. *New Media and Society, 6,* 403-423.

Park, H. W., Thelwall, M., & Kluver, R. (2005). Political hyperlinking in South Korea: Technical indicators of ideology and content. *Sociological Research Online, 10(3).*

Patchin, J. W., & Hinduja, S. (2011). Traditional and non-traditional bullying among youth: A test of General Strain Theory. *Youth and Society, 43*, 727-751.

Patterson, G. R., & Yoerger, K. (1999). Intraindividual growth in covert antisocial behaviour: A necessary precursor to chronic juvenile and adult arrests? *Criminal Behaviour and Mental Health, 9,* 24-38.

Peersman, C., Schulze, C., Rashid, A., Brennan, M., & Fischer, C. (2014). iCOP: Automatically identifying new child abuse media in P2P networks. *2014 IEEE Security and Privacy Workshops.*

Perez, L. M., Jones, J., Englert, D. R., & Sachau, D. (2010). Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media images. *Journal of Police and Criminal Psychology, 25,* 113-124.

Perliger, A., & Pedahzur, A. (2011). Social network analysis in the study of terrorism and political violence. *PS: Political Science & Politics*, *44*, 45-50.

Piquero, A. R., Brame, R., & Lynam, D. (2004).Studying criminal career length through early adulthood among serious offenders. *Crime & Delinquency, 50,* 412-435.

Piquero, A. R., Farrington, D. P., & Blumstein, A. (2003). The criminal career paradigm. *Crime and Justice, 30,* 359-506.

Potter, G. (1994). *Criminal organizations: Vice, racketeering and politics in an American city*. Prospect Heights, IL: Waveland Press.

Powell, M., Cassematis, P., Benson, M., Smallbone, S., & Wortley, R. (2014). Police officers' perceptions of their reactions to viewing Internet child exploitation material. *Journal of Police and Criminal Psychology, online first.* doi: 10.1007/s11896-014-9148-z.

Pratt, T. C., Holtfreter, K., Reisig, M. D. (2010). Routine on-line activity and Internet fraud targeting: Extending the generality of Routine Activity theory. *Journal of Research in Crime and Delinquency, 47,* 267-296.

Prichard, J., Watters, P. A., & Spiranovic, C. (2011). Internet subcultures and pathways to the use of child pornography. *Computer Law & Security Review, 27,* 585-600.

Provos, N., & Holz, T. (2007). *Virtual honeypots: from botnet tracking to intrusion detection*. Boston, MA: Pearson Education.

Pyrooz, D. C., Decker, S. H., & Moule Jr, R. K. (2015). Criminal and routine activities in online settings: Gangs, offenders, and the Internet*. Justice Quarterly, 32,* 471-499.

Pyrooz, D. C., Sweeten, G., & Piquero, A. R. (2013). Continuity and change in gang membership and gang embeddedness. *Journal of Research in Crime and Delinquency*, *50,* 239-271.

Qu, B., Niu, W., Zhu, T., Wu, L., Liu, S., & Wang, N. (2013). Dynamic user behavior-based piracy propagation monitoring in wireless peer-to-peer networks. *Behavior and Social Computing: Lecture Notes in Computer Science, 8178,* 44-55.

Quayle, E., & Taylor, M. (2002). Child pornography and the Internet: Perpetuating a cycle of abuse. *Deviant Behavior, 23,* 331-361.

Quayle, E., & Taylor, M. (2011). Social networking as a nexus for engagement and exploitation of young people. *Information Security Technical Report, 16,* 44-50.

Quinn, J. F., & Forsyth, C. J. (2005). Describing sexual behavior in the era of the Internet: A typology for empirical research. *Deviant Behavior, 26,* 191-207.

Rasmussen, K. B. (2008). General approaches to data quality and Internet-generated data. In N. Fielding, M. Raymond, & G. Blank (Eds.), *The SAGE handbook of online research methods* (pp.79-98). London, UK: SAGE Publications, Ltd.

Ray, J. V., Kimonis, E. R., & Seto, M. C. (2014). Correlates and moderators of child pornography consumption in a community sample. *Sexual Abuse: A Journal of Research and Treatment, 26,* 523-545.

Reips, U., & Birnbaum, M. H. (2011). Behavior research and data collection via the Internet. In K. Vu & R. Proctor (Eds.), *Handbook of human factors in web design* (pp.564-585). Boca Raton, FL: CRC Press.

Reips, U., Buchanan, T., Krantz, J., & McGrawn, K. (2011). Methodological challenges in the use of the Internet for scientific research: Ten solutions and recommendations. *Studia Psychologica, 11,* 5-18.

Reiss, A. J. (1986). Co-offending and criminal careers. In A. Blumstein, J. Cohen, J. A. Roth, & C. A. Visher (Eds.), *Criminal careers and career criminals (pp. 121-160).* Washington, DC: National Academy Press.

Reiss, A. J. (1988). Co-offending and criminal careers. *Crime and Justice, 10,* 117-170.

Reiss, A. J., & Farrington, D. P. (1991). Advancing knowledge about co-offending: Results from a prospective longitudinal survey of London males. *Journal of Criminal Law & Criminology, 82,* 360-395.

Ressler, S. (2006). Social network analysis as an approach to combat terrorism: Past, present, and future research. *Homeland Security Affairs, 2,* 1-10.

Reuter, P. (1983). *Disorganized crime: The economics of the visible hand.* Cambridge, MA: MIT Press.

Rheingold, H. (2006). Social networks and the nature of communities. In P. Purcell (Ed.), *Networked Neighbourhoods (pp. 47-75).* London, UK: Springer.

Rice, S. R., & Ross, M. W. (2014). Differential processes of Internet versus real life sexual filtering and contact among men who have sex with men. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 8,* Article 1. doi: 10.5817/CP2014-1-6.

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

Roberts, J. W., & Hunt, S. A. (2012). Social control in a sexually deviant cybercommunity: A cappers' code of conduct. *Deviant Behavior, 33,* 757-773. doi: 10.1080/01639625.2012.679894.

Rodriguez, M. G., Leskovec, J., & Scholkopf, B. (2013). Structure and dynamics of information pathways in online media. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 23-32). New York, NY.

Rosenmann, A., & Safir, M. P. (2006). Forced online: Pushed factors of Internet sexuality: A preliminary study of paraphilic empowerment. *Journal of Homosexuality, 51,* 71-92.

Rutgaizer, M., Shavitt, Y., Vertman, O., & Zilberman, N. (2012). Detecting pedophile activity in BitTorrent networks. *Lecture Notes in Computer Science, 7192,* 106-115.

Saari, E., & Jantan, A. (2013). E-Cyborg: The cybercrime evidence finder. In *Information Technology in Asia (CITA), 2013 8th International Conference on* (pp. 1-6). Kota Samarahan.

Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology, 34,* 193-240.

Sampson, R. J., & Laub, J. H. (1993). *Crime in the making: Pathways and turning points through life.* Cambridge, MA: Harvard University Press.

Sampson, R. J., & Laub, J. H. (1997). A life-course theory of cumulative disadvantage and the stability of delinquency. In T. P. Thornberry (Ed.), *Developmental theories of crime and delinquency, vol. 7 (pp. 133-161).* New Brunswick, NJ: Transaction Publishers.

Sarnecki, J. (2001). *Delinquent networks: Youth co-offending in Stockholm.* New York, NY: Cambridge University Press.

Schaefer, D. R., (2012). Youth co-offending networks: An investigation of social and spatial effects. *Social Networks, 34,* 141-149.

Schell, B. H., Dodge, J. L. (2002). *The hacking of America: Who's doing it, why, and how*. Westport, CT: Quorum Books.

Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research, 37,* 291-318. doi: 10.1177/0049124108327128.

Seigfried, K. C., Lovely, R. W., & Rogers, M. K. (2008). Self-reported online child pornography behavior: A psychological analysis. *International Journal of Cyber Criminology, 2,* 286-297.

Seto, M. C., Hanson, R. K., & Babchishin, K. M. (2011). Contact sexual offending by men with online sexual offenses. *Sexual Abuse: A Journal of Research and Treatment, 23,* 124-145.

Seto, M. C., Hermann, C. A., Kjellgren, C., Priebe, G., Svedin, C. G., & Langstrom, N. (2015). Viewing child pornography: Prevalence and correlates in a representative community sample of young Swedish men. *Archives of Sexual Behavior, 44,* 67-79.

Sharp, K., & Earle, S. (2003). Cyberpunters and cyberwhores: Prostitution on the Internet. In Y. Jewkes (Ed.), *Dot cons. Crime, deviance and identity on the Internet (pp. 33-89).* Portland, OR: Willan Publishing.

Shropshire, K. O., Hawdon, J. E., & Witte, J. C. (2009). Web survey design: Balancing measurement, response, and topical interest. *Sociological Methods & Research, 37,* 344-370. doi: 10.1177/0049124108327130.

Slonje, R., Smith, P. K., & Frisén, A. (2013). The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior, 29*, 26-32.

Spink, A., Ozmutlu, HC, & Lorence, DP (2004). Web searching for sexual information. An exploratory study. *Information Processing and Management: An International Journal, 40,*113-123.

Spitzner, L. (2003). *Honeypots: tracking hackers* (Vol. 1). Reading, UK: Addison-Wesley.

Steel, C. M. S. (2009). Child pornography in peer-to-peer networks. *Child Abuse & Neglect, 33,* 560-568.

Steinmetz, K. F., & Tunnell, K. D. (2013). Under the pixelated jolly roger: A study of on-line pirates. *Deviant Behavior, 34,* 53-67.

Sullivan, J., & Beech, A. R. (2004). Assessing Internet sex offenders. In M. C. Calder (Ed.), *Child sexual abuse and the Internet: Tackling the new frontier (pp. 69-83).* Lyme Regis, UK: Russell House.

Taylor, M., Holland, G., & Quayle, E. (2001). Typology of paedophile picture collections. *The Police Journal, 74,* 97-107.

Taylor, M., & Quayle, E. (2003). *Child pornography: An Internet crime.* New York, NY: Brunner-Routledge.

Taylor, P. (1999). *Hackers.* London, UK: Routledge.

Taylor, R. W., Fritsch, E. J., & Liederbach, J. (2014). *Digital crime and digital terrorism*. Upper Saddle River, NJ: Prentice Hall Press.

Technical Analysis Group. (November, 2013). *Examining the cyber capabilities of Islamic terrorist groups.* Retrieved from: http://www.ists.dartmouth.edu/library/164.pdf.

Tenti, V., & Morselli, C. (2014). Group co-offending networks in Italy's illegal drug trade. *Crime, Law, and Social Change, 62,* 21-44.

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology*, *57*, 60-68.

Tita, G. E., & Radil, S. M. (2011). Spatializing the social networks of gangs to explore patterns of violence. *Journal of Quantitative Criminology, 27,* 521-545.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry, 32,* 425-443.

Tremblay, P. (1993). Searching for suitable co-offenders. In R. V. Clarke & M. Felson (Eds.), *Routine activity and rational choice (pp. 17-36).* New Brunswick, NJ: Transaction Publishers.

Tremblay, P. (2002). Social interactions among paedophiles*. Les Cahiers de Recherches Criminologiques, 36,* 1-48.

Tremblay, P. (2006). Convergence settings for nonpredatory 'Boy Lovers'. In R. Wortley & S. Smallbone (Eds.), *Situational prevention of child sexual abuse (pp. 145-168).* Monsey, NY: Criminal Justice Press.

Tremblay, P., Bouchard, M., & Petit, S. (2009). The size and influence of a criminal organization: A criminal achievement perspective. *Global Crime, 10,* 24-40.

Tremblay, P., Laisne, G., Cordeau, G., MacLean, B., & Shewshuck, A. (1989). Carrieres criminelles collectives: Evolution d'une population delinquante. *Criminologie, 22,* 65-94.

Tretyakov, K., Laur, S., Smant, G., Vilo, J., Prins, P. (2013). Fast probabilistic file fingerprinting for big data. *BMC Genomics, 14, S2-S8.* doi: 10.1186/1471-2164-14-S2-S8.

Urboniene, A. (2014). Motivation for blogging: A qualitative approach*. International Journal of Global Business Management and Research, 2*(2), Paper 2 (1-14).

van Hout, M. C., & Bingham, T. (2013). 'Silk Road', the virtual drug marketplace: A single case study of user experiences. *International Journal of Drug Policy*, *24*, 385-391.

van Mastrigt, S. B., & Carrington, P. J. (2014). Sex and age homophily in co-offending networks: Opportunity or preference? In C. Morselli (Ed.), *Crime and networks (pp. 28-51).* New York, NY: Routledge.

van Mastrigt, S. B., & Farrington, D. P. (2011). Prevalence and characteristics of co-offending recruiters. *Justice Quarterly, 28,* 325-359.

van Wijk, A., Nieuwenhuis, A., & Smeltink, A. (2009). *Behind the scenes: An exploratory investigation into the downloaders of child pornography.* Arnhem, NL: Bureau Beke.

Vehovar, V., Ziberna, A., Kovacic, M., Mrvar, A., & Dousak, M. (2009). Technical report on An Empirical Investigation of Paedophile Keywords in eDonkey P2P Network. Measurement and analysis of P2P activity against paedophile content project. Retrieved from: http://antipaedo.lib6.fr/.

Warr, M. (1996). Organization and instigation in delinquent groups. *Criminology, 34,* 11-36.

Warr, M. (2002). *Companions in crime*. New York, NY: Cambridge University Press.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* Cambridge, UK: Cambridge University Press.

Webb, L., Craissati, J., & Keen, S. (2007). Characteristics of Internet child pornography offenders: A comparison with child molesters. *Sexual Abuse: A Journal of Research and Treatment, 19,* 449-465.

Weerman, F. M. (2003). Co-offending as social exchange. Explaining characteristics of co-offending. *British Journal of Criminology, 43,* 398-416.

Weimann, G. (2005). How modern terrorism uses the Internet. *The Journal of International Security Affairs, 8,* 91-105.

Wellman, B. (1983). Network analysis: Some basic principles. *Sociological Theory, 1,* 155-200.

Wellman, B., Boase, J., & Chen, W. (2002). The networks nature of community: Online and offline. *IT & Society, 1,* 151-165.

Wellman, B., & Haythornwaite, C. (Eds.). (2002). *The Internet in everyday life.* Malden, MA: Blackwell Publishers Ltd.

Westlake, B. G., Bouchard, M., & Frank, R. (2011). Finding the key players in child exploitation networks. *Policy & Internet, 3(2)*, Article 6. doi: 10.2202/1944-2866.1126.

Westlake, B. G., Bouchard, M., & Frank, R. (2012). Comparing methods for detecting child exploitation content online. Paper presented at the *European Intelligence and Security Informatics Conference 2012,* Odense, Denmark.

Whittle, H. C., Hamilton-Giachritsis, C. E., & Beech, A. R. (2014a). “Under His Spell”: victims’ perspectives of being groomed online. *Social Sciences*, *3*, 404-426.

Whittle, H. C., Hamilton-Giachritsis, C. E., & Beech, A. R. (2014b). In their own words: young peoples’ vulnerabilities to being groomed and sexually abused online. *Psychology, 5(10),* 1-12. doi: 10.4236/psych.2014.510131.

Whittle, H. C., Hamilton-Giachritsis, C., Beech, A., & Collings, G. (2013). A review of online grooming: Characteristics and concerns. *Aggression and Violent Behavior*, *18*, 62-70.

Williams, P. (2001). Organizing transnational crime: Networks, markets and hierarchies. In P. Williams & D. Vlassis (Eds.), *Combating transnational crime: Concepts, activities, and responses (pp. 57-87).* London, UK: Frank Cass Publishers.

Williams, R., Elliott, I. A., & Beech, A. R. (2013). Identifying sexual grooming themes used by Internet sex offenders. *Deviant Behavior, 34,* 135-152. doi: 10.1080/01639625.2012.707550.

Wiseman, J. (1996). *SM 101: A realistic introduction.* San Francisco, CA: Greenery Press.

Wolak, J., Finkelhor, D., & Mitchell, K. J. (2005). *Child pornography possessors arrested in Internet-related crimes: Findings from the National Juvenile Online Victimization Study.* Alexandria, VA: National Center for Missing & Exploited Children.

Wolak, J., Finkelhor, D., & Mitchell, K. J. (2012). *Trends in law enforcement responses to technology-facilitated child sexual exploitation crimes: The Third National Juvenile Online Victimization Study (NJOV-3).* Durham, NH: Crimes against Children Research Center.

Wolak, J., Liberatore, M., & Levine, B. N. (2014). Measuring a year of child pornography trafficking by U.S. computers on a peer-to-peer network. *Child Abuse & Neglect, 38,* 347-356.

Wortley, R. K. (2012). Situational prevention of child abuse in the new technologies. In E. Quayle & K. M. Ribisl (Eds.), *Understanding and preventing online sexual exploitation of children (pp. 1-28).* New York, NY: Routledge.

Wortley, R. K., & Smallbone, S. (2006). *Child pornography on the Internet*. Retrieved from: www.cops.usdoj.gov/publications/e04062000.pdf.

Wortley, R. K., & Smallbone, S. (2012). *Internet child pornography: Causes, investigation, and prevention.* Santa Barbara, CA: ABC-CLIO.

Wurtele, S. K., Simons, D. A., & Moreno, T. (2014). Sexual interest in children among an online sample of men and women: Prevalence and correlates. *Sexual Abuse: A Journal of Research and Treatment, 26,* 546-568.

Yang, C. C., & Ng, T. D. (2007). Terrorism and crime related weblog social network: Link, content analysis and information visualization. In *Intelligence and Security Informatics, 2007, 55-58*.

Yasrebi, H., Sperisen, P., Praz, V., & Bucher, P. (2009). Can survival prediction be improved by merging gene expression datasets? *PLOS One, 4*(10), 1-14. doi: 10.1371/journal.pone.0007431.

Young, K. (2005). Profiling online sex offenders, cyber-predators, and pedophiles. *Journal of Behavioral Profiling, 5,* 1-18.

Zhang, S. X. (2009). Beyond the 'Natasha' story–a review and critique of current research on sex trafficking. *Global Crime*, *10*, 178-195.

Zhao, Y., Levina, E., & Zhu, J. (2011). Community extraction for social networks. *Proceedings of the National Academy of Science, 108,* 7321-7326.

Zhou, Y., Reid, E., Qin, J., Chen, H., & Lai, G. (2005). US domestic extremist groups on the Web: Link and content analysis. *Intelligent Systems, IEEE, 20,* 44-51.