May 2015

Criminal Careers in Cyberspace: Examining Website Failure within Child Exploitation Networks

Bryce G. Westlake
Simon Fraser University, bryce.westlake@sjsu.edu

Martin Bouchard
Simon Fraser University

Criminal Careers in Cyberspace: Examining Website Failure within Child Exploitation Networks

Bryce Westlake*, Martin Bouchard

Simon Fraser University

*Corresponding Author: School of Criminology, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6; Tel: +1 778 782 8135 Fax: +1 778 782 4140; bwestlak@sfu.ca.

Bryce Westlake is a PhD candidate of Criminology at Simon Fraser University. His research interests include sexual offending, illegal online networks and cybercrime in general. His work has been published particularly in *Journal of Applied Psychology* and *Policy & Internet.*

Martin Bouchard is an Associate Professor of Criminology and Director of the International Cybercrime Research Centre at Simon Fraser University. In addition to co-editing two books, his work has been published in the *Journal of Quantitative Criminology, Journal of Research in Crime and Delinquency, Justice Quarterly,* and *Social Networks,* among others. In 2013, he received the Award of Excellence from the Dean of Arts and Science at Simon Fraser University emphasizing his quality with mentoring graduate students

Acknowledgement: The authors would like to thank Dr. Richard Frank for his work on creating the web-crawler used for this study as well as the contributions of research assistant Ashleigh Girodat. This work was supported by the Social Sciences and Humanities Research Council under Grant #435-2012-0336.

Word Count: 11,657

Abstract

Publically accessible illegal websites represents an additional challenge for control agencies, but also an opportunity for researchers to monitor, in real time, changes in the their criminal careers. Using a repeated measures design, we examine the evolution in the networks that form around child exploitation (CE) websites over a period of sixty weeks, and determine which criminal career dimensions predict website failure. Network data were collected using a custom-designed web-crawler. Baseline survival rates were compared to networks surrounding (legal) sexuality or sports websites. Websites containing CE material were no more likely to fail than comparisons. Cox regression analyses suggest that increased volumes of CE code-words and images are associated with premature failure. Websites that are more popular have higher odds of survival. We show that traditional criminal career dimensions can be transferred to the context of online CE and constitute some of the key determinants of an interrupted career.

Keywords**:** Criminal careers; Survival analysis; Child pornography; Cybercrime; Web-crawler

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* in 1986 (Blumstein, Cohen, Roth, & Visher) much has been learned about how criminal careers start, change, and end (DeLisi & Piquero, 2011; 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 these lenses 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 the 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 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 therefore may not even matter, to cyber offenders. The process is further complicated by the relative anonymity of online crime (Durkin & Bryant, 1999; Holt, Blevins, & Burkert, 2010), the novelty of the methods some use (Grabosky, Smith, & Dempsey, 2001; Holt & Graves, 2007), 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 use 10 different seed websites with known CE material and launch a custom-designed web-crawling tool that maps the networks around these seed websites up to a limit of approximately 300 websites per network. The web-crawling tool allows us to automatically collect a number of characteristics on each website that can be operationalized under the various dimensions of the criminal career. Our goal is to examine which of these website characteristics are associated with the failure of websites in our sample. We review the prior literature on online persistence and the process of transferring the dimensions of the criminal career to cyberspace, before turning to a more detailed description of the data and methods used.

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). Illegal websites face the same challenges as legal websites but must balance success with maintaining survival in the face of a constant risk of being shut down by control agencies. 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 during their criminal career an offender (or website) may deem that the cost of maintaining that balance exceeds the benefits thereby interrupting its trajectory.

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 financial 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 which includes anonymity, globalization, and the ambiguous legal status of many online behaviors.

Transitioning Criminal Career Dimensions to Cybercrimes

Outside of death, it is difficult to ascertain the cessation of an individual's criminal career (Piquero, Farrington, & Blumstein, 2003). Likewise, the criminal career duration of a website is difficult to quantify. It may be possible to observe websites long enough to see them go offline, but their start dates are unlikely to be available to researchers. As offline offenders may

experience career interruptions due to incarceration or life-events, their ability to avoid detection, or their transition to a predominantly mentorship role, these sorts of conceptualizations are rarely synonymous with absolute desistance. Rather, desistance is 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[1]. That is, the name of a website 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 web address. The websites analyzed for this study were followed for sixty-weeks, with more than 15% going offline during the observation period.

[1] 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 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 content (e.g., movies, music or child exploitation media), offending frequency can be measured in multiple ways. If a website begins distributing ten movies illegally, within a one hour time frame, do we consider that ten offenses or one? If a website continues to distribute those ten movies the following day is it considered a new offense or a continuation of the previous offense? If a website is still distributing those ten 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 (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). 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 (softcore/hardcore), 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 which extend beyond the criminal event (Gallupe & Bouchard, forthcoming; McGloin & Nguyen, 2014; 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, forthcoming; 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 of a website.

The Current Study

Once a solitary crime, with sparse networks, the sexual exploitation of children has gone through a sort of resurgence as a result of 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 content, 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 child exploitation (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?

Methods

Data

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 twenty comparison networks –10 sports and 10 (legal) sexuality. Using a repeated measures design, data were collected 10 times, at an interval of 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 for

their search engine. Following a set of researcher-specified criteria, CENE begins on a *seed website* and collects information on the size of the website and the number of images, videos, and keywords. CENE then follows hyperlinks found on the seed website, and subsequently hyperlinked websites, to other websites and repeats the data collection process. In essence, the hyperlinks act as a form of snowball/chain sampling. Starting from a small subject pool (seed websites) we follow the connections (hyperlinks) to other websites and determine whether those associated websites meet the study criteria and should be included, until we meet the desired sample size. From the websites sampled, a (social) network of websites can be created, connected via hyperlinks.

Seed Websites

Each initial 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[2]. The second source was a list of websites identified, and inspected in previous research, to be involved in the dissemination of CE material. This 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

[2] 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.

subsequent website the hyperlinking webpage had to contain at least 7 of our 82 keywords *or* at least 1 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. Child exploitation networks began with an average of 305.10 (s.d. =2.33) websites and were re-crawled at an interval of 42.14 (s.d. =4.45) days.

As the focus of our study is illegal websites *and* the communities that surround them, we do not discriminate between whether a connecting website is legal or illegal. Given that some of these websites contain thousands of webpages and hidden directories it is beyond the scope of this study to confirm that the primary focus of each website is to disseminate CE material. However, we attempt 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[3] that ranked blogs based on popularity. The site-seed websites were selected using a sports marketing website[4] 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*[5] 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] http://labs.ebuzzing.com/top-blogs/sports

[4] http://www.marketingcharts.com/

[5] 'BDSM' stands for a) bondage and discipline; b) sadomasochism; and c) dominance and submission (Wiseman, 1996).

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[6]). The second were victim identification words (23) not directly linked to CE but typically present (e.g., boy, girl, child). The third were explicit 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 contains 2.25 million hash values[7] classified into three categories. The first (*Child Exploitation*) contains 618,632 images and are classified, under the Canadian Criminal Code, as being CE. The second (*Child Nudity*) contains 652,223 images that would probably be considered CE by a judge. However images in this category are not blatant and, thus, the risk averse nature of law enforcement results in these images being placed into a separate category. The third (Collateral) contains 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 sexually exploited.

---

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

[7] 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}.

Dependent Variable

Duration

Offline, the criminal career is not linear, as it is often filled with interruptions and changes in offending patterns (e.g., Laub & Sampson, 2003). Likewise, a website's criminal career may undergo fluctuations that may include interruptions such as going offline temporarily. Therefore, this study refers to duration as the time between the beginning of data collection and the first, recordable, interruption. We defined an interruption as being a data collection point where the website was not reachable (i.e., offline) [8]. While we assume that some of the interruptions will be due to detection and removal by control agencies, it is 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 are unknown[9]. The start date of the observation period was the same for all websites included within a network. In addition, our data is right-censored because some of our websites remained active at the conclusion of our sixty-week observation period. Although we are unable to accurately determine the full criminal career duration, our primary focus is 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

[8] 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.

[9] 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).

collection, to being active for the full observation period. The relative "age" of a website is controlled for via various indicators of size, as discussed below.

Independent Variables

Volume of illegal content

The cumulative nature of the Internet means that, especially for websites, the career dimension of offending frequency is best measured through the volume of illegal content. We measure volume specifically pertaining to *code keywords* and *known CE images*.

Code Keywords: Of the 82 keywords used by the web-crawler, 27 were code words used to identify CE material specifically. This variable is a count of all code keywords found on a website. While code words may change and new ones evolve, there is some evidence to suggest that many code words are lasting, such as pthc, QWERTY, or similar variants (Steel, 2009).

Child Exploitation Images: The database used by the web-crawler contained information on 618,632 images identified as being CE by the Canadian Criminal Code. This variable is the number of CE images found on a website.

Borderline Images: Although the focus of our study is CE material that we define to be images identified by the RCMP to meet the criteria as set by the Canadian Criminal Code. The presence of *child nudity* or *collateral* images is also indicative of the potential presence of other illegal material. This dichotomized variable identifies whether a website contains any *child nudity* or *collateral* images.

Crime-mix type

Focusing on specialization amongst websites, we created three composite variables that measure a website's content preferences. We focused this investigation in three areas: the *sex of the victim* (boy/girl), the *severity* (hardcore/softcore) of the content descriptions, and the preferred *media* (image/video/story) type chosen for disseminating CE material. 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.

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[10].

Severity (Hardcore/Softcore): The severity focus of a website was determined by the higher relative frequency of hardcore or softcore keywords, selected from the 82 keywords used by CENE. The composite variable *hardcore* consisted of 21 keywords related to severe sexual abuse (e.g., cries, torture, and rape). The composite variable *softcore* 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 hardcore focused.

Media (Video/Image/Story)*:* CE material is most often distributed through three types of media. Distributors may exchange videos or images depicting sexual abuse or they may narrate stories of real or fictional sexual abuse. For both *videos* and *images*, we used the average number of

[10] 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.

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 distribution websites were the most prevalent accounting for 83.7% of websites while 9.2% of websites were video focused and 7.1% were story focused.

Connectivity

The co-offending dimension of criminal careers does not transfer directly to the context of this study. However, the websites we analyze are embedded into a larger network that acts as an online community where recognition and support varies from relatively isolated websites to others acting as popular convergence points for consumers of CE material. The connectivity to and from more websites results in an increased ability to acquire and distribute content. However, theoretically, the more connected a website is, the greater its vulnerability to detection.

Incoming/outgoing hyperlinks*:* 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 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.

Analytic Methods

To determine whether baseline survival rates for websites in CE networks differed from sports websites 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. Some of the individual networks had very few failures and small sample sizes. 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 (domains). 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 was compared to the estimated Nelson-Aalen cumulative hazard (Cox & Snell, 1968). Second, a Harrell's C concordance statistic was calculated (Harrell, Lee, & Mark, 1996)

Results

Do websites derived from legitimate seeds (our comparison groups) survive longer than websites derived from seed hosting child exploitation (CE) material? Figure 1a displays the KM estimates averaged for each genre and seed-type. Results show that websites within networks starting from

a CE seed do not fail at a greater rate than those starting from a legal sexuality or sports seed. More specifically, Figure 1a shows that the survival curves of the sexuality-seed sample of websites 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 1b), 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 1 here]

Selecting only the 1580 unique websites derived from the 10 seeds confirmed to host CE material, Table 1 compares characteristics of websites which survived (n =1303) to those that failed (n =277). The most interesting finding is 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 are 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 act 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'etre* that provides incentives for website owners to maintain the website as active. The results also suggest the existence of a core, static, group of larger websites coexisting along a peripheral group that carries a significantly larger failure rate.

[Table 1 here]

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 include 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 long period prior to the start of the study but happened to fail during the early phases of the study). Of the 277 failures that occurred during the study period, 89 (32.13%) failed between Wave 1 and Wave 2. While this early attrition is expected, we still need to examine whether early failures had different characteristics than subsequent failures. Table 1 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 websites, less likely to be focused on videos, but more likely to be hardcore 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. The fact that we found that so many early failures were focused on presenting hardcore content, or code words used by offenders, is perhaps indicative of the vulnerability of these websites to detection. Despite the similarities between early and subsequent failures, 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 multivariate models presented below.

Predicting Time to Failure

Table 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[11]. Our results show that 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.

[Table 2 here]

The next three models each introduce indicators of three dimensions of the careers of CE websites we introduced in this study. Model 2 removes the general characteristics and introduces our indicators for variations in the volume of illegal content on these websites. The results suggest that volume of illegal content matters. Model 2 (Table 2) shows that for each additional known *CE image* found on a website, odds of survival were decreased by 1.27%. Interestingly, having grey area images (i.e. *child nudity* or *collateral* images) was 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* of the content were shown to be associated with failure. The type of *media* used on a website, however, emerged as a significant factor. Specifically, websites focused on *stories* were shown to have increased odds of failure compared to websites predominantly hosting *images*.

[11] 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.

The importance of the location of a website in the online community is evidenced in Model 4 (Table 2) that examines the effect of the number of *incoming* and *outgoing* 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. For each additional outgoing hyperlink found on a website, odds of survival increased by 2.04%. For each additional incoming hyperlink, odds of survival increased by 1.32%. However, their effects seem to be working independently as the interaction between outgoing and incoming hyperlinks 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 2) simultaneously examines the effects of webpages, volume of illegal content, crime-type mix, and connectivity. There are little changes to the main results: small, story focused websites, as well as those with CE images and websites that are not well connected within their own online communities were shown to fail more quickly than others. Two changes from previous models are 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 of Model 2 that *child nudity* and other child-related images were associated with longer survival no longer holds, when controlling for all of the other relevant factors in Model 5. In the end, it is the volume of illegal content that appears to matter most in predicting website failure in our sample.

Discussion

The fact that an increasing number of crimes are committed in cyberspace, or using it as a tool in

committing crimes, is starting to get the attention of mainstream criminology. Beyond the sheer size of the phenomenon (in costs to society, in number victims and offenders), one reason for the growing attention to cybercrime is because of the impact it has on offline, traditional crime. The rise of the Internet is perceived as one of the plausible explanations in accounting for the crime drop of the 1990s – one of the macro-level explanations for a similarly universal decline (Farrell, Tseloni, Mailley, & Tilley, 2011; Marlow, 2014). Closer to our concerns in this study, 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 in the field will be to test whether the existing conceptual frameworks and theories can be used to study crimes and criminal in cyberspace. Another is to develop new conceptual and methodological tools to account for some of the unique elements of cybercrimes.

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 content. We focused on the dimensions of 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 is uniquely adapted to the fact that we are studying a relatively new unit of analysis (i.e. the illegal website) online, and that this unit of analysis is 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 child exploitation content, and the community of other websites around 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-related seed websites, and 10

legal sexuality seed websites. There are two main findings discussed 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 is survival amongst the confirmed illegal websites that was found to be longer (Figure 1b). While the high demand for child exploitation (CE) material coupled with better detection avoidance tactics for these sites 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 sites. As many (legal or not) websites are quickly abandoned because of the time commitment (and often resources) required to maintain one, coupled with an inability to attract users (Hsu & Lin, 2008; Urboniene, 2014), it may be that the demand for illegal content may motivate an operator to maintain their website. This conclusion is supported by the findings of our multivariate analyses, namely that larger and more popular websites persisted (Table 2), and that early failures were predominantly comprised of small websites with little to no content (Table 1).

Second, a series of findings emerged from the Cox regression analyses focused on the websites that are part of the social structure around our initial 10 seeds hosting illegal content. Here, two of the three website-specific criminal career dimensions emerged as important predictors: volume of illegal content, and connectivity of the websites. Both the volume of illegal content and connectivity tap into the notion of "visibility" online, which 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 sites do not necessary have a high volume of known illegal content. In fact, popularity may stem

from the novelty of the material found on a site, as opposed to older material, albeit material confirmed to be child pornography. Two, popularity may breed motivation for website owners to persist with a site. The more visitors, the more successful, the more incentives to persist and publish new material on a site, or simply keep it online. Recall that website failures are not solely due to the work of control agencies. In all likelihood, many were found to be offline because their owners decided to not pursue the site for personal reasons.

Next, the increased risks of failure associated with higher volumes of illegal content 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 has already classified as illegal naturally increases the vulnerability of a website to detection. We believe that website owners take this risk unknowingly as it is unlikely that they are aware of their website containing 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©, SensualWriter©, or LiveJournal©. When a user creates a blog on one of these websites 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 being taken, or abilities, 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 being a cyclical process. That is, there is very little preventing a user from creating a new blog, on the same or different blog 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 child exploitation video database, similar to image databases. Currently there are no reliable techniques for detecting child exploitation 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).

Limitations

Our study was subject to four primary limitations. First, much like any offline networks include both offenders and non-offenders (e.g. Haynie, 2002), it is likely that contained within our child exploitation networks were websites that did not disseminate CE material (i.e., false positives). However, their connection to websites hosting illegal content 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 website 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 measure websites every six weeks, we do not know the precise date of failure, but rather 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 (Hertz-Picciotto & Rockhill, 1997).

Third, although security measures used by offenders online have been shown to be minimal, those that do use them are probably more likely to persist longer than those that do not. One of the simplest security measures is the use of user registration and/or passwords. The websites analyzed in this study did not use any form of security measure to mask their activities (i.e., publicly accessible). Therefore, our findings are limited to websites with minimal security measures implemented.

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[12], 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

[12] 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.

to examine illegal websites on the Deep Web as well 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. It remains paramount to understand what is available easily and publically as we do in this study.

Future Research

The findings of this study identify at least two 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.

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; Moffitt et al., 2002;). Therefore future research needs to examine, generally, whether the criminal careers of websites can be categorized into specific trajectories and, specifically, whether existing trajectory models can be translated to cyberspace. More generally, future work is needed to conceptualize and measure the trajectories of websites hosting illegal content other than child exploitation material.

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 child exploitation website is associated with increased risks of failure. The paper also provides a framework for future research transitioning the criminal career framework 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 theories are applied in order to identify the differences between offline and online crimes.

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, 2013). 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.

References

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

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

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

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

Bossler, A. M., & Holt, T. J. (2009). On-line activities, guardianship, and malware infection: An examination of Routine Activities Theory. *International Journal of Cyber Criminology, 3,* 400-420.

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

Bouchard, M. (2007). A capture-recapture model to estimate the size of criminal populations and the risk of detection in a marijuana cultivation industry. *Journal of Quantitative Criminology, 23,* 221-241.

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

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

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.

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

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.

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

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

Decary-Hetu, 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 career research, 2000-2011: A state-of-the-art review. *Journal of Criminal Justice, 39,* 289-301.

Durkin, K. F., & Bryant, C. D. (1999). Propagandizing pederasty: A thematic analysis of the online exculpatory accounts of unrepentant pedophiles. *Deviant Behavior, 20,* 103-127.

Farrell, G., Tseloni, A., Mailley, J., & Tilley, N. (2011). The crime drop and the security hypothesis. *Journal of Research in Crime and Delinquency, 48,* 147–175.

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

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

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 hazard tests and diagnostics based on weighted residuals. *Biometrika, 81,* 515-526.

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

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.

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.

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

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

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

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

Holt, T. J., & Graves, D. C. (2007). A qualitative analysis of advanced fee fraud schemes. *The International Journal of Cyber-Criminology, 1,* 137-154.

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

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.

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.

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

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

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: Harvard University Press.

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/.

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.

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., LeBlanc., M., & Proulx, J. (2005). The generality of criminal behavior: A confirmatory factor analysis of the criminal activity of sex offenders in adulthood. *Journal of Criminal Justice, 33,* 177-189.

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.

Marlow, A. (2014). Thinking about the fall in crime. *Safer Communities, 13,* 56-62.

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., & Nguyen, H. (2014). The importance of studying co-offending networks for criminological theory and policy. In C. Morselli (Ed.) *Crime and Networks* (pp. 13-27). New York: Routledge.

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.

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

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., Caspi, A., Harrington, H., & Milne, B. J. (2002). Males on the life-course-persistent and adolescence-limited antisocial pathways: Follow-up at age 26 years. *Development and Psychopathology, 14,* 179-207.

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

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

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.

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.

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

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.

Pyrooz, D. C., Decker, S. H., & Moule Jr, R. K. (2013). Criminal and routine activities in online settings: Gangs, offenders, and the Internet. *Justice Quarterly, (online first),* 1-29.

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

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

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

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.

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

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

Taylor, P. (1999). *Hackers: Crime in the digital sublime.* London: Routledge.

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. (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.

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

Warr, M. (2002). *Companions in crime: The social aspects of criminal conduct*. Cambridge University Press.

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

Westlake, B. G., Bouchard, M., & Frank, R. (2011). Finding the key players in child exploitation networks. *Policy and Internet, 3(2),* Article 6.

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.

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 (NCMEC 06-05-023). Alexandria, VA: National Center for Missing & Exploited Children.

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.

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

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

*p<0.05 **p<0.01

Table 2. Proportional hazard regression models of time to first website failure

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

Figure 1a. Comparing survival rates across genres by seed-type.

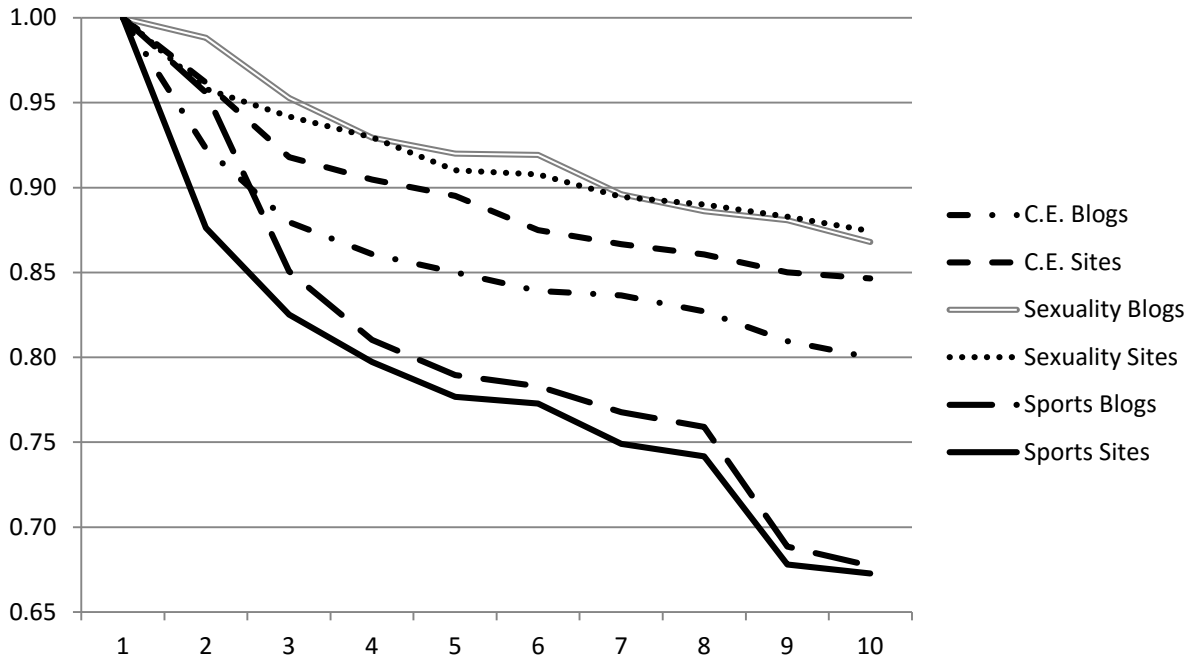

Figure 1b. Comparing survival rates of CE network websites with and without known images to comparison networks.