Shuffle Up and Deal:
An Application of Capture-Recapture Methods to Estimate the Size of Stolen Data Markets

by

Mitchell Clark Macdonald

B.A. (Honours, Criminology), Saint Mary's University, 2013

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

in the
School of Criminology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Summer 2016

Approval

Name:	**Mitchell Clark Macdonald**
Degree:	**Master of Arts (Criminology)**
Title:	***Shuffle Up and Deal: An Application of Capture-Recapture Methods to Estimate the Size of Stolen Data Markets***
Examining Committee:	**Chair:** Martin Andresen Professor

Richard Frank, PhD
Senior Supervisor
Assistant Professor _____

Martin Bouchard, PhD
Supervisor
Associate Professor _____

Thomas Holt, PhD
External Examiner
Associate Professor
School of Criminal Justice
Michigan State University _____

Date Defended/Approved: August 24, 2016

Abstract

Often overlooked in the measurement of crime is the underlying size of offender populations. This holds true for online property crimes involving the sale, purchase, and use of stolen financial data. Despite available data suggesting that such frauds are steadily increasing, the number of actors comprising stolen data markets has yet to be determined. The current study addresses this issue using two related capture-recapture methods—Zelterman's estimator and its extended covariate adjusted model—to estimate the population sizes of buyers, vendors, money launderers, and facilitators who are active within online marketplaces in a calendar year. Data analysis consisted of samples collected from 3 websites that facilitate financial crimes and frauds. While the observed overlap between marketplaces was rare, results indicate that websites are perhaps not distinct entities, but are better conceptualized as a collective marketplace that is much larger in size than what can otherwise be observed.

Keywords: Cybercrime; Carding; Stolen Data Markets; Online Property Crimes; Size of Criminal Populations; Capture-Recapture Methods

Dedication

To three good men. My grandfathers, Jim and Tom, and my great uncle, Allan.

Acknowledgements

I would first like to acknowledge my senior supervisor Richard Frank. Thanks for being patient with me, while allowing me to do my own thing. Your hands off approach has allowed me to make mistakes, figure things out on my own, and gain a true understanding of what research is all about. You've also given me the opportunity to work on two different research projects, which has granted me plenty of opportunities to grow as a researcher. Thank you for everything.

To the remainder of my supervisory committee. Foremost, I would like to acknowledge the support of Martin Bouchard. Your 'networks' class was my first experience in graduate school. It was an intense and (at times) intimidating experience that I was not ready for, but it nonetheless set the tone for the remainder of my Masters degree. Your continued support has since helped to shape my thesis, conceptually and analytically. Thanks to Tom Holt for taking on the role of my External Examiner. Your input is very much appreciated and improved the quality of my thesis. I would also like to thank Martin Andresen for not only taking the time to Chair my defence, but also for being good natured and supportive since the time that I've walked through the door at SFU.

To Russell (Augie) Westhaver. Thank you for taking an interest in me as an undergraduate student just a few short years ago. Your mentorship was key to the development of my writing, critical thinking skills, and grit—qualities that have enabled me to have success in graduate school. I am sure that you have had a similar effect on many others, regardless of their career paths.

To the friends I've made along the way. Julianna Mitchell, Monica Ly, Allison Campbell, Oliva Ha, Bryan Monk, Nick Bordignon, Walt Works, Tarah Hodgkinson, Evan McCuish, Jeff Mathesius, Kyle Sutherland, and Ryan Scrivens. Your friendship means a great deal to me and I am fortunate to have met each and every one of you.

To my family. To my niece, Halle. You are my favourite person on Earth. Its been fun watching you grow up, even if much of it has been from afar. To my 'little sister', Al. I caused you a lot of grief throughout much of our upbringing, but you took it standing up. While not close like we once were, I hope that will change in time. With your smarts and

strong-mindedness, you can achieve anything and I truly hope that you do. To my sister, Amanda. You're one of the smartest people I know, and I truly admire your gritty, unapologetic, yet caring personality. You've also done a bang up job at being a 'big sister'. You were always good to me as a kid and you continue to show that same kindness. Thanks for the constant love and support. And to my Mom and Dad. You've raised me to be kind, respectful, and confident—hopefully some of those qualities have stuck. The two of you have truly afforded me every opportunity in life, supported me along the way, and have never given me any grief throughout the pursuit of my goals. Thanks for being great parents.

This study was supported by the Cyber Security Cooperation Program (CSCP), Public Safety Canada.

Table of Contents

List of Tables

List of Figures

Chapter 1.

Introduction

Through the expansion of online financial services, including Internet banking, money transfer, and payment systems, millions of financial transactions are now conducted online each day. But the proliferation of these services has also created boundless opportunities for financially motivated hacking and phishing, which has become pervasive. The proceeds of data breaches are estimated to surpass USD$1 trillion per year, affecting millions of people, corporations, and governments worldwide (United Nations, 2015). Data theft most frequently affects the financial sector (IBM, 2014; Trend Micro, 2015), as banking institutions and related financial systems are three times more likely to be targeted than IT firms or government agencies (Ponemon Institute, 2015). Those responsible for data breaches seek access to large repositories of personally identifiable information: full and common names; physical addresses; national identification numbers (e.g., social insurance numbers); passports; driver's licenses; date of births; telephone numbers; email addresses; and, perhaps the most sought after pieces of information, bank account, credit, and debit card numbers. By extension, frauds that occur subsequent to data theft, such as the sale, purchase, and monetization of stolen financial data are steadily increasing; however, a great deal of speculation surrounds their actual *prevalence* (Tcherni, Davies, Lopes, & Lizotte, 2006; Williams & Levi, 2012), as measured by the number of offences, victims, and/or the direct financial losses incurred in a given year.

Central to the reliability issues associated with the data sources used to derive prevalence estimates of online fraud, is the fact that these crimes go largely unnoticed. Compared to traditional frauds, online frauds are arguably harder to detect and impede (Filipkowski, 2008)—an issue which is directly tied to the hidden and elusive nature of the underlying offender population. Further amplifying this issue is that the size of offender

populations is a focal, but often overlooked, indicator of prevalence. Bouchard and Lussier (2015) put forward that at the root of this oversight is the assumption that the number of offenders within a population reflect patterns and frequency of offending. For instance, it is perhaps intuitive for one to assume that increases in crime correspond with increases in the number of offenders, or vice versa. Though the size of offender populations and frequency of offending are not as interrelated as what researchers may assume. Unforeseen factors, such as more lucrative criminal opportunities, may indeed contribute to an upward progression in offender frequencies in a population that is stable (Bouchard & Lussier, 2015). In contrast, increases in the size of offender populations may not coincide with increases in the frequency of offending due to the insufficient human and social capital needed to sustain a successful criminal career (Loughran, Nguyen, Piquero, & Fagan, 2013; Tremblay, Bouchard, & Petit, 2009). By merely accepting these fallacies, the size of criminal populations will remain obscure.

But estimating the number of offenders perpetrating online fraud poses particular challenges, the upmost of which concerns accessing suitable data from which to derive estimates. The rates of detection (e.g., number of arrests and rearrests) for these offences are so few that official crime statistics are not a viable option (Tcherni et al., 2016). To circumvent such limitations, criminologists now consider alternative data sources like online crime forums to study populations of online offenders. Because crime forums archive exchanges between participants that occur before, during, and/or after the crime commission process, criminal intent and activity is recorded through open or public dialogue. These depositories of 'naturally occurring' user-generated data are able to be collected, coded, and analyzed to measure patterns and frequencies of online offending, and scope populations that would otherwise remain hidden (Décary-Hétu & Aldridge, 2015; Housely et al., 2014; Hutchings & Holt, 2015; Williams & Burnap, 2016; Williams, Burnap, & Sloan, 2016). For online property crimes specifically, carding forums—websites that facilitate financial crimes and frauds, also colloquially referred to as stolen data markets—provide rich sources of data that are suitable to study the prevalence of offenders involved in the sale, purchase, or monetization of stolen financial data.

Despite legitimate interest in the social dynamics and organization of stolen data markets (Afroz, Garg, McCoy, & Greenstadt, 2013; Décary-Hétu & Leppänen, 2013; Holt,

2011, 2013b; Holt & Lampke, 2010; Hutchings & Holt, 2015; Motoyama et al., 2011; Soudijn & Zegers, 2012; Yip, Shadbolt, & Webber, 2012; Yip, Webber, & Shadbolt, 2013), it remains unknown just how many offenders, who are active, but remain hidden, comprise these online marketplaces (Décary-Hétu & Laferrière, 2015). Part of the issue is feasibility of data collection. Previous analyses have relied on manual data collection techniques (Holt, 2013b; Holt & Lampke, 2010; Holt, Smirnova, Chua, & Copes, 2015; Holt, Smirnova, & Chua, 2016), which are time intensive and not practical for the collection of voluminous amounts of data needed to undertake large-scale analysis. But perhaps the more pressing issue is the absence of a strong conceptual framework from which to explain and measure the population size of online marketplaces (Ablon, Libicki, & Golay, 2014; Décary-Hétu & Laferrière, 2015; Holt & Lampke, 2010). Prior studies have shown capture-recapture methods to be useful in meeting the challenges associated with scoping criminal populations, particularly the number of offenders comprising illicit markets (Bouchard, 2007a, 2008; Bouchard & Tremblay, 2005; Bouchard et al., 2012).

The current study uses two related capture-recapture methods—Zelterman's truncated Poisson estimator and its extended covariate adjusted model—to estimate the population sizes of buyers, vendors, money launderers, and facilitators, active in stolen data markets over one calendar year. Data analysis involved samples collected from 3 websites representing different online marketplaces for stolen data. Analysis consisted of two stepwise procedures. First considered was the size of each market, after which the population size of offenders frequenting *any* marketplace was extrapolated by taking the overlapping distribution across the 3 independent samples and applying Zelterman's models. While the observed overlap between marketplaces was rare, results indicate that markets are perhaps not distinct entities, but may be better conceptualized as singular components of a collective marketplace that is much larger in size than what can otherwise be observed. Findings have numerous implications towards a broader understanding of dynamics and social organization within these online market settings.

Chapter 2.

Conceptual Framework and Literature Review

2.1. Prevalence of Online Fraud

2.1.1. Prevalence of Offending

Prevalence of offending is typically measured by counts of known offences that are derived from official statistics reported by police to the Uniform Crime Report (UCR) survey, which are then aggregated to arrive at national figures.[1] In Canada, like most other of the world's developed countries, aggregate trends in property crimes as measured by the UCR have steadily decreased along with the larger crime drop, beginning in the early 1990s. Fraud is no exception to this trend. As shown in Figure 2.1, over the 10 years spanning 2005 – 2014, there has been an observable 24% drop in the crime rate (Boyce, 2015). Researchers have raised the possibility that this drop is the direct result of an incremental increase in the number of frauds that are perpetrated online (Kong, 2006; Tcherni et al., 2016); however, more recent statistics show a reversal in this trend. Following a steep decline after 2011, the number of fraud-related offences incrementally increased from 2012 – 2015 (Boyce, 2015; Boyce, Cotter, & Perreault, 2014). But why the decline and rise in the trend? Beginning in 2011, modifications to the reporting procedures for official crime statistics mandated the recording of cybercrimes, including frauds perpetrated online. Prior to this change in reporting, fraud-related offences perpetrated online would have been recorded as a property crime—in many cases without recording specifics about the 'online component'. Analysis of the crime trend spanning 2011 – 2015 indicates that this change in reporting almost certainly impacted the reported prevalence of fraud—reflected by the steepness of the drop following 2011 (Brennan & Dauvergne,

[1] But the interpretation of counts varies depending on the crime being measured. For violent crime, offences represent the number of victims, but for (online) property crimes, such as fraud, offences represent the number of incidents (Kong, 2006). For instance, a data breach exposing 10,000 unique records can potentially lead to the victimization of 10,000 individuals, although the underlying offence, the data breach, is recorded as a single incident.

2011). In other words, rather than reflect a crime drop, official statistics simply reflect the changing nature of fraud and crime reporting (Kong, 2006; Tcherni et al., 2016).

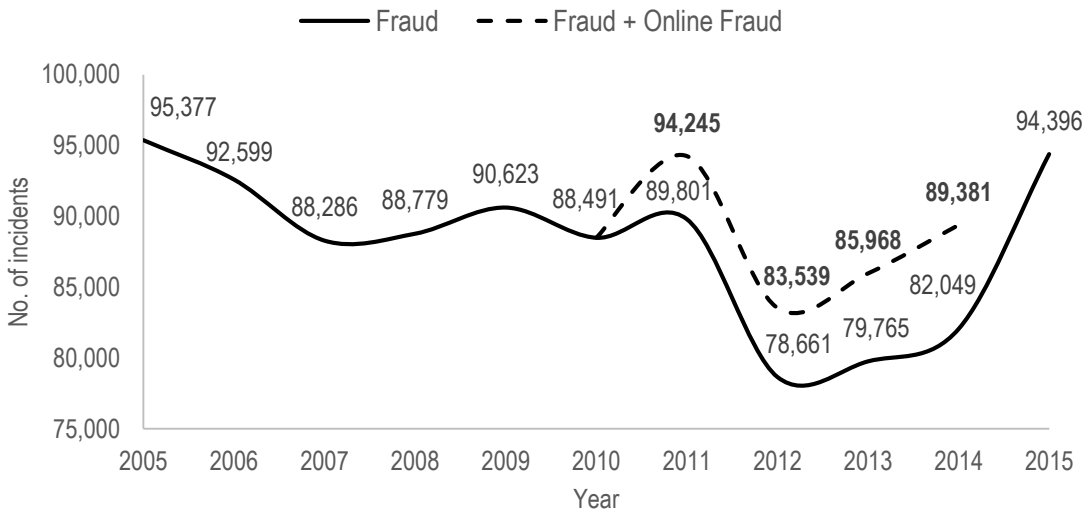

Figure 2.1 Police-reported fraud statistics, 2005 – 2015
Note: Statistics on fraud (2005 – 2014) were derived from crime reports published by the Canadian Centre for Justice Statistics (Allen, 2016; Brennan, 2012; Brennan & Dauvergne, 2011; Dauvergne, 2008; Dauvergne & Turner, 2010; Silver, 2007; Wallace, 2009). Data on cybercrimes (2011 – 2014) was accessed through written request to Statistics Canada.

Considering the trend in online fraud alone, the number of police-reported incidents increased by nearly 65% from 4,444 offences in 2011 to 7,332 in 2014. While this short term trend provides evidence that the number of police-reported incidents of online fraud are steadily increasing, the actual prevalence of these crimes and the degree in which they are increasing remains unclear. The speculation surrounding estimates is attributed to reliability issues associated with official statistics, including varying definitions of fraud and inconsistencies in the manner in which frauds are counted by local police detachments and in the UCR survey, leading to different estimates of offending, which renders official statistics unreliable (Kong, 2006; Williams & Levi, 2012). Furthermore, despite increasing specification in reporting procedures, as discussed above, there is still an unknown degree of confounding between annually reported incidents of fraud (property crime) and online fraud (cybercrime) that cannot be separated in aggregate crime statistics. As Boyce et al. (2014) rightly argue, part of the difficulty in determining the number of frauds that are perpetrated online is attributed to the blurring of online-offline financial activity. The

concurrent and interchangeable use of paper currency and online payment systems has obscured the boundaries between traditional and online commerce, which makes it difficult to distinguish whether or not frauds actually involve an online component, and for this reason, many incidents are obscurely recorded (Williams & Levi, 2012).

But the underlying issue with official statistics concerns under-reporting bias on the part of victims. Many victims of frauds, perpetrated online or otherwise, do not report their victimization to police, thereby underestimating the prevalence of online frauds reported to police. Non-reporting is the result of either a conscious choice made by the victim, confusion surrounding proper reporting procedures or a lack of reporting requirements, or that victims simply remain oblivious to their victimization (Bossler & Holt, 2013). Victims who choose not to disclose their victimization to police do so for a multitude of reasons, including the belief that the police lack the ability to adequately investigate these crimes, the meager value of their financial losses, or that financial institutions or credit card companies flagged the fraud(s) and remitted the funds, so victims do not feel as though they were victimized. Likewise, victims are often unsure as to whom exactly to report incidents. That is, it is unclear for many whether to report incidents to the police, their financial institution or credit card company, or anti-fraud organizations like the Canadian Anti-Fraud Centre. If left unaddressed, the many issues associated with non-reporting ensure that the 'true' prevalence of online fraud will remain underestimated.

2.1.2. Prevalence of Victimization and Direct Financial Losses

The under-reporting of online fraud to police is illustrated by the discrepancy between figures reported by the UCR and data compiled by victimization surveys, which indicate that prevalence far exceeds what is reported by official statistics. According to the 2009 General Social Survey – Victimization (GSS), 4% of Internet users in Canada were victims of online fraud (Perreault, 2011). Considering that 80.3% of the 2009 Canadian population (33.63 million) had access to the Internet (World Bank, 2016), this 4% figure suggests that upwards to 1.08 million people, or 3.2% of the Canadian population, were victims of online fraud. More recent figures disclosed by the Canadian Bankers Association (CBA) indicated that some forms of online payment card fraud are on the

decline, whereas others are increasing (CBA, 2015).[2] In 2008, the year that chip and personal identification number (PIN) technology was introduced to the Canadian market, 148,000 accounts were flagged for debit card fraud, compared to 24,795 accounts in 2015. The 83% decrease over this period suggests that these security measures have resulted in their intended effect of reducing debit card fraud; however, during this same period, there was a sharp increase in victimization rates of credit card fraud. A reported 450,322 accounts were flagged for flagrant use in 2008, compared to 961,851 accounts in 2015, representing nearly a 114% increase in the number of compromised accounts.[3] Interestingly, these figures are also similar to the proportion of victims reported in the 2009 GSS. The combined incidents of debit and credit card fraud in 2015 (986,646 accounts), suggest that just 0.74% – 9.06% of incidents are actually captured by official statistics.[4]

Related to prevalence of victimization is the severity of the offence. The severity associated with online fraud has two defining characteristics: 1) a single incident can affect a grossly disproportionate number of people and 2) the direct financial losses accrued by victims range from negligible to large sums. In Canada, victims of phishing and online payment card fraud incurred losses averaging CAD$940 in 2012, CAD$569 in 2013, and CAD$508 in 2014 (Canadian Anti-Fraud Centre, 2014). Data from the CBA (2015) also indicated that the average direct loss from debit card fraud dropped almost 33% from

[2] Data sources from financial institutions and corporate merchants are amongst the most complete and reliable sources from which to assess the prevalence of victimization and the corresponding extent of severity within a given population; however, cooperation from the financial sector with regards to data sharing for measuring the prevalence of fraud has proved challenging (Kong, 2006; Taylor-Butts & Perreault, 2009). Corporate entities are inclined to withhold data out of privacy concerns for their clients, but perhaps more so to preserve corporate reputation and customer confidence in the anti-fraud measures associated with their products and services.

[3] Comparisons between data sources illustrates the large discrepancies that can be found in reported victimization. For instance, using 2007 – 2008 fraud data disclosed by 29% of all individual branches of the major Canadian financial institutions, Taylor-Butts and Perreault (2009) found that fraudulent use of debit cards account for nearly 50% of all flagged incidents, whereas credit card frauds did not account for even 1% of the total number of incidents. But according to 2008 figures reported by the CBA (2015), credit card fraud accounted for 67% of incidents of online bank fraud. Given the low level of cooperation from financial institutions with regards to Taylor-Butts and Perreault's survey, the statistics released by CBA are likely to contain less bias and, thus, are preferred for reference.

[4] This range is given as due to the discussed unknown degree of confounding between frauds that are counted as property crimes compared to frauds that are recorded as cybercrimes in official crime statistics.

CAD$706 in 2008, to CAD$476 in 2015, whereas the average direct loss from credit card fraud dropped nearly 17% from CAD$905 to CAD$755, over the same time period. Though the accuracy of per victim estimates are not reliable, as outliers that skew estimates upwards are generally not accounted for (Florêncio, & Herley, 2013), these figures nonetheless suggest an interesting development—at the domestic level, the average financial losses incurred by victims are trending in the opposite direction of the crime rate. In other words, though the direct financial losses experienced by victims are, on average, less severe, the number of incidents are indeed increasing.

CBA (2015) data also shows that national trends in the total volume of direct losses incurred by victims vary by type of fraud. From 2008 to 2015, the cumulative losses of debit card fraud reduced by almost 89% from CAD$104.5 million to CAD$11.8 million—the direct result of the introduction of chip and PIN security measures. In contrast, the direct cumulative losses resulting from credit card fraud rose 78% from CAD$407.7 million in 2008 to CAD$726.5 million in 2015. To get a global estimate of the direct costs incurred by victims, Anderson and colleagues (2013) extrapolated figures from the United Kingdom, assuming that domestic estimates reflected 5%, a figure equivalent to the United Kingdom's share of the global gross domestic product, of global victimization. Under these parameters, online payment card fraud was estimated to cost victims USD$4.2 billion annually. But in addition to the rather arbitrary metric used to derive this estimate, the statistics used for calculation were published through annual 'threat reports' produced by security vendors, such as McAfee®, Microsoft®, and Symantec™. Estimates from such data sources raise concerns about validity regarding potential over-reporting bias, as vendors have a vested interest in inflating statistics to create heighten demand for their products (Anderson et al., 2013; Williams & Levi, 2012).

2.2. Estimating Prevalence of Offender Populations

One means to address the measurement issues posed by inconsistencies across data samples and the dark figure of crime is to estimate the size of offender populations (van der Heijden, Cruyff, & Böhning, 2014). The size of offender populations concerns the underlying prevalence of offending or the proportion of a population committing one or more crime types, who are active during an observation period (Blumstein, Cohen, Roth,

& Visher, 1986). Though the size of offender populations is often overlooked in favour of measuring the frequency or severity of offences (Blumstein & Cohen, 1987; Blumstein, Cohen, & Farrington, 1988; Farrington, 1992; Rossmo & Routledge, 1990). This holds true for offenders involved in online property crimes, including financial crimes, frauds, and market crimes, despite the evidence presented which suggests that such activity is steadily increasing. This may be due to belief in a widely held fallacy: that the number of offenders within a population reflects patterns and frequency of offending. Perhaps it is intuitive to assume that increases in crime correspond with increases in the number of offenders active in a population, or vice versa. Though the size of offender populations and frequency of offending may not be as interrelated as what would otherwise be assumed. Factors that impact criminal opportunities, such as changes in technology (Brennan & Dauvergne, 2011), may contribute to an upward progression in offender frequencies in an otherwise stable population of offenders (Bouchard & Lussier, 2015). Likewise, increases in the size of offender populations may not coincide with increases in frequency of offending if, for example, offenders lack the necessary human and social capital needed to sustain a successful criminal career (Loughran et al., 2013; Tremblay et al., 2009). By merely accepting that aggregate crime trends, whether in the short or long term, reflect the underlying number of offenders, the true prevalence of offender populations will remain obscure.

Capture-recapture methodologies have proven to be effective methods for producing accurate estimates of offender populations. Capture-recapture is based on the simple logic that reoccurring patterns within data collected on open or closed populations can be modeled to derive an estimate of the population that is active, but remains hidden. Initially used to estimate the incompleteness of census data (Darroch, 1958, 1961; Fienberg, 1972; Jolly, 1965; Seber, 1965), capture-recapture methods were later adopted to study stability in fisheries (Newman & Waters, 1989) and wildlife populations (Pollock, Nichols, Brownie, & Hines, 1990). In a criminological context, Greene and Stollmack (1981) first used capture-recapture methods to estimate the general population of offenders in Washington, DC, who were active over a 2-year period. Since this initial application, capture-recapture methods have been used to estimate the prevalence of auto thefts (Collins & Wilson, 1990), burglars (Riccio & Finkelstein, 1985), incidents of domestic violence (van der Heijden et al., 2014), drug users (Brecht & Wickens, 1993; Hay

et al., 2008; Hser, 1993; Wickens, 1993) and dealers (Bouchard & Tremblay, 2005; Bouchard et al., 2012), forced labour and human trafficking (van der Heijden, de Vries, Böhning, & Cruyff, 2015), illegal gun owners (van der Heijden, Cruyff, & van Houwelingen, 2003), marijuana cultivation and production (Bouchard, 2007a, 2008), phishing (Weaver & Collins, 2007), prostitutes (Leyland, Barnard, & McKeganey, 1993; Rossmo & Routledge, 1990) and johns (Roberts & Brewer, 2006), and sex offenders (Bouchard & Lussier, 2015).

2.2.1. Virtues and Challenges of Estimating Offender Prevalence

Estimating the size of criminal populations provides researchers with substantive context about underlying patterns of offending. This has broad implications, spanning theory development to informing legislation and the scope of interventions or controls, while effectively allocating limited resources (Rossmo & Routledge, 1990). The most extensive body of research in this area has sought to expose the size of drug user populations (e.g., Böhning, Suppawattanabodee, Kusolvisitkul, & Viwatwongkasem, 2004; Brecht & Wickens, 1993; Hay et al., 2008; Hser, 1993; Mastro et al., 1994; Rhodes, 1993; Wickens, 1993) and illegal drug markets more generally (Bouchard, 2007a, 2008; Bouchard & Tremblay, 2005; Bouchard et al., 2012). Such knowledge is necessary for developing adequate responses to address the unique issues associated with the offender populations on both the demand and supply side of drug markets. Estimating the size of drug user populations has direct implications for the successful development of treatment programs, such as informing staffing requirements, the size of facilities, and the funding needed to properly subsidize programs. Knowing the number of active suppliers and dealers in a marketplace is needed to assess whether the disruption strategies of law enforcement are effective at curbing the supply side of drug markets (Bouchard, 2007b). Likewise, knowledge of the number of offenders partaking in online frauds as the potential to inform whether the disruption strategies of law enforcement cause enough of a disruption effect to be considered a good use of scarce resources (see Glenny, 2011). By comparison, population estimates may suggest that crime prevention, particularly the development of cyber security policy that reduces opportunities for crime and impacts the economics influencing the market (Jones, 2007; Nagurney, 2015), is the better strategy to address issues with offender prevalence.

But estimating the size of criminal populations poses inherent challenges, many of which are especially relevant for estimating the number of offenders partaking in online property crimes, such as fraud. For starters, capture-recapture methods require at least one, but usually two or three complementary datasets to derive suitable estimates of hidden populations. For especially vulnerable populations like drug users, it is possible to triangulate and exhaust data sources (e.g., hospital, court, and police records, etc.) to saturate the recorded activity of the population(s) under study (Hser, 1993). While the inclusion of multiple, mutually exclusive datasets is ideal, for the analysis of online offenders, this luxury is not necessarily practical from a data collection standpoint, as data on offenders is scarce (Kong, 2006; Tcherni et al., 2016). Aside from official crime statistics there are no other obvious complementary data sources that capture the recorded history of online offenders. Victimization surveys and incident reports are not useful in this context, as these statistics do not reflect the prevalence of the offender population. The biggest issue with these statistics for deriving offender estimates is that victimization rates do not consider the impact of repeated offenders, the 5% offenders responsible for 50% – 60% of all crime (Wolfgang, Figilo, & Sellin, 1972), or control for 'susceptible' victims who are victimized on more than one occasion.[5]

Despite being among the best options, official crime statistics also pose particular issues for estimating the size of offender populations. First of which, official crime statistics are more reflective of the total number of incidents, a measurement which is also not necessarily representative of offender populations. As previous studies have concluded, a substantial proportion of offending involves two or more co-offenders. Andresen and Felson (2012) found that co-offending is most common among serious offenders, including those involved in financial and market crimes. These findings were subsequently supported by Morselli, Grund, and Boivin (2015), who found that financial and market crimes were amongst the most frequent and central crimes within a large network comprised of all known co-offenders in the Canadian province of Québec. Though official statistics portray a different story for online property crimes. Carrington, Brennan,

[5] Individuals at increased risk of victimization of online fraud share different characteristics than the 'average' victim. Those especially susceptible to online fraud are more likely to have post-secondary education, earn more than CAD$60,000, and use the Internet on a daily basis, compared to individuals at less risk (Perreault, 2011).

Matarazzo, and Radulescu (2013) found that 88% of online frauds were perpetrated by solo offenders, compared to 12% involving two or more co-offenders, although recent statistics indicate that less than 5% of offenders involved in online property crimes, such as online payment card fraud and identity theft, are identified by police (Mazowita & Vézina, 2014); therefore, the incompleteness of official statistics limits one's ability to draw satisfactory conclusions about such patterns of offending. But more importantly, the incompleteness of these data sources also negates their applicability for estimating the size of offender populations perpetrating online property crimes. The rates of detection (e.g., arrest and rearrests) for online offenders are so few that, if limited to official statistics, the application of capture-recapture methods to estimate the size of this offender population is not even a viable option.

2.3. Stolen Data Markets

Data collected from websites that facilitate financial crimes and fraud, known as carding forums, afford researchers opportunities to compile archived data on otherwise invisible populations (Burnap & Williams, 2015, 2016; Williams & Burnap, 2016). Two obvious reasons come to mind as to why forums are exploited for use by online offenders: 1) forums follow an inherent market design structure (Afroz et al., 2013), such as Amazon™ and eBay™ and 2) forums facilitate social networking. In a market context, the social networking features of forums connect networks of buyers, vendors, and other market actors to exchange money for prohibited products and services, such as drugs (Aldridge & Décary-Hétu, 2014; Dolliver, 2015), malware and vulnerabilities (Chu, Holt, & Ahn, 2010; Holt, 2013a), and botnet services (Décary-Hétu & Dupont, 2012, 2013). Although the underlying financial motives of the contemporary hacker community (Holt & Lampke, 2010; Kilger, Arkin, & Strutzman, 2004) has led to the persistent issue of carding—hacker jargon encompassing the theft, sale, purchase, and monetization of stolen financial data (Décary-Hétu & Leppänen, 2013; Holt, 2013b; Holt & Lampke, 2010; Hutchings & Holt, 2015; Motoyama et al., 2011; Peretti, 2008; Yip et al., 2012, 2013). Interactions that occur through carding forums, colloquially referred to as stolen data markets, commonly involve vendors posting threads advertising goods and services to which prospective buyers respond. But as with any other profit-oriented illicit activity,

avoiding detection involves obscuring the proceeds of data theft, so that they are inconspicuous to audit and law enforcement agencies. Given the implications of stolen financial data for subsequent money laundering activity, carding forums also support the profit motive through the advertisement and procurement of services that facilitate the money laundering process, so that funds can be freely spent (Holt, 2013b; Hutchings & Holt, 2015; Richet, 2013; Soudijn & Zegers, 2012).

2.3.1. Applicability of Online Data for Capture-Recapture

The social dynamics and organization of stolen data markets have garnered the legitimate interest of researchers (Afroz et al., 2013; Décary-Hétu & Leppänen, 2013; Holt, 2011, 2013b; Holt & Lampke, 2010; Hutchings & Holt, 2015; Motoyama et al., 2011; Soudijn & Zegers, 2012; Yip et al., 2012, 2013). Cybercrime researchers have advocated for future research to undertake comparative analysis of online marketplaces to determine whether social and market forces are ubiquitous or merely reflective of particular markets (Décary-Hétu & Leppänen, 2013; Holt, 2013b; Holt & Lampke, 2010). Although qualitative and quantitative analyses have revealed that there are more commonalities than differences among marketplaces, despite variations in size, products, goods, and services, and primary spoken language (Afroz et al., 2013; Holt & Lampke, 2010; Holt, 2013b; Motoyama et al., 2011; Zhao et al., 2016). An analytic framework consolidating data collected from multiple marketplaces into a single dataset increases generalizability, which has implications for the exploration of trends across marketplaces (Décary-Hétu & Aldridge, 2015). Furthermore, as there is empirical evidence to suggest that online offenders frequent multiple markets (Motoyama et al., 2011; Zhao et al., 2016), consolidating data in this manner enables mobility patterns in the known populations of online offenders to be modeled using capture-recapture methodologies to determine the prevalence of offenders lurking across various marketplaces where stolen data is bought, sold, and obscured for fraudulent use.

Although online market offenders remain arguable even more hidden than actors within illegal drug markets, for example, due to the anonymous nature of online exchanges, which transcend physical time and space (Hutchings, 2014; Hutchings & Holt, 2015). Nonetheless, online offenders are identifiable by their online pseudonyms, which

links an identity to illicit activity (Holt, Strumsky, Smirnova, & Kilger, 2012), whether that consists of buying, selling, or monetizing the proceeds of stolen financial data. As such, online pseudonyms provide a unique identifying feature that enables offenders to be observed or *captured* within and across multiple markets, which makes capture-recapture methods suitable for estimating the populations frequenting these online settings. But it is also plausible and likely that market actors mitigate the risks of heightened status and reputation, such as increased monitoring from law enforcement, through the use of multiple pseudonyms to obscure the extent of their activity; however, qualitative interviews have revealed that individuals who frequent hacking and carding forums are generally not inclined to alter their online pseudonyms (Lusthaus, 2012), as the social capital accumulated through one's reputation is directly linked to a specific online pseudonym (Afroz et al., 2013). Thus, it is logical to assume that the majority of market actors within stolen data markets, particularly those on the supply side, are willing to trade increased security for greater efficiency (Morselli, Giguère, & Petit, 2007). That is, it is likely that the majority of market offenders adhere to a single online pseudonym to reap the benefits of the social capital (e.g., reputation) that they have accumulated in the marketplace.

2.4. Dynamics and Organization of Stolen Data Markets

2.4.1. Role Occupancy and Specialization of Labour

Many have argued that the role occupancy and specialization of labour evident in stolen data markets demonstrates a level of sophistication reflective of traditional 'hierarchical' crime networks (Afroz et al., 2013; Broadhurst, Grabosky, Alazab, Bouhours, & Chon, 2014; Hutchings, 2014; Mcguire, 2012; Moore, Clayton, & Anderson, 2009).[6] But as Reuter (1983) argued, market actors are organized only insofar that they can effectively carry out their illicit activities. Stolen data markets are more appropriately described as a

[6] Implicit in this argument is that stolen data markets are analogous to criminal organizations. While this is not the argument made here, a United States Department of Homeland Security – Secret Service joint investigation led to the 2014 arrest and conviction of 57 individuals connected to an online marketplace where stolen data is bought and sold through an unprecedented use of the *Racketeer Influenced and Corrupt Organizations* (RICO) *Act*. RICO legislation grants authority for extended sentencing in cases where the offences were committed on behalf or to facilitate the activity of a criminal organization.

weakly connected criminal supply chain of suppliers, vendors, and buyers (Peretti, 2008), as well as other auxiliary actors like money launderers and facilitators (Holt, 2011, 2013b; Hutchings & Holt, 2015; Soudijn & Zegers, 2012). But these roles do not necessarily exist in a vacuum. Market actors may occupy dual roles, such as suppliers and vendors, buyers and vendors, or vendors and facilitators (Franklin, Perring, Paxson, & Savage, 2007; Holt, 2013b; Hutchings & Holt, 2015).

Suppliers are the offenders responsible for harvesting stolen data. Most commonly, data is obtained through phishing attacks, which deceive victims into disclosing their financial and/or personal credentials to a seemingly trustworthy source, such as a representative from a financial institution (Leukfeldt, 2014, 2015; Leukfeldt, Kleemans, & Stol, 2016). More sophisticated data theft involves targeted malware attacks on computer networks and databases operated by financial institutions and credit card suppliers, as well as related financial systems like automated teller machines and point of sale terminals.[7] Suppliers distribute stolen data to vendors (Holt, 2013b; Peretti, 2008), market offenders synonymous to street level dealers in illegal drug markets.

Vendors operate in a marketplace where advertised products are detailed via thread postings in which the breadth of description is argued to be the direct result of competition (Holt, 2011; Holt & Lampke, 2010). Vendors emphasize the quality and validity of their data, in which higher rates of useable, quality data garner higher demand on the open market (Holt, 2011, 2013b; Holt et al., 2016). The most common forms of data exchanged in the marketplace in bulk lots include credit card and bank account credentials that include corresponding client information and account verification numbers ('*dumps*', *track 1 and 2*). The extent of corresponding information that accompanies advertised data varies depending on quality and price (Holt, 2013b; Holt & Lampke, 2010). The highest quality data advertised in online marketplaces are full disclosures of financial credentials ('*fullz*') that include the accountholder's name, date of birth, and address, as well as the

[7] The major variants of these malware were developed in Eastern Europe. The most notorious banking Trojan, *Zeus*, is known for its pervasiveness and long lifespan, compared to similar forms of malware. Custom 'plugin' tools also coincide with these pre-existing malware to cater to the specific needs of the end user(s) and widen the scope of potential targets.

issuing bank, PIN, and national identification number (e.g., social insurance number)—all of which facilitate fraud and identity theft (Holt et al., 2015).

Motoyama et al. (2011) found that the top 10% of vendors accounted for upwards of 50% of the market's inventoried products and services, as measured by the number of threads created by vendors. Vendors with the largest market shares may also be those with the highest quality products (Herley & Florêncio, 2010; Holt & Lampke, 2010; Yip et al., 2013). These core vendors are fixtures within their markets, ensuring marketplace stability despite uncertain market conditions (Afroz et al., 2013) and the transient nature of the majority actors (Broadhurst et al., 2014). As relatively few core vendors service the market, competition decreases amongst vendors, resulting in higher prices for advertised goods and services (Eck, 1995). But according to Herley and Florêncio (2010), advertised prices often exceed market demand, regardless of the quality of data. In support of this argument, subsequent studies have found that vendors with the lowest advertised prices accrued the most contacts (Décary-Hétu & Laferrière, 2015; Holt, 2013b). Apart from setting competitive pricing structures, vendors use other strategies to attract and maintain clientele over time, including extending customer service post-purchase and discount pricing for buyers making bulk purchases (Franklin et al., 2007; Holt & Lampke, 2010; Hutchings & Holt, 2015). Research by Holt, Smirnova, and Chua (2013) found that the impact of extending customer service post-purchase, through mechanisms such as refunds for unusable data, on potential revenue yields only modest returns for vendors, though the authors note that any impact that customer service has on ensuing transactions with buyers may ensure that the additional output is worthwhile. In contrast, the economies of scale associated with bulk sales through discount prices ensures that both buyers and vendors benefit from this arrangement—higher quantity of data sold at a lower price per unit allows vendors to maximize profit, whereas buyers satisfy their demand at a lower cost (Décary-Hétu & Laferrière, 2015).

Buyers are perhaps the most transient subgroup frequenting stolen data markets. Within any marketplace, there are argued to be limited populations of buyers for whom vendors must compete (Décary-Hétu & Laferrière, 2015). But despite competition amongst vendors, it is often more efficient for the many buyers to make contact with the relatively few vendors supplying the market, than for vendors to pursue buyers (Eck,

1995). Buyers make contact with vendors through direct response to thread advertisements and/or through the forum's private messaging system.[8] The basic laws of supply and demand would suggest that the marketplace favours buyers, though buyers are in the precarious position of incurring the risk in any given transaction. As the quality and validity of the advertised products and/or services is really only truly known to the vendor, buyers assume all risk involved with encountering unscrupulous actors. In short, the asymmetry in the buyer-vendor relationship caused by the dynamics of the marketplace favours vendors (Holt et al., 2015; Holt et al., 2016; Yip et al., 2013).

In addition to the complications imposed by assuming all risk associated with transactions, many buyers do not often understand how to transform their newly acquired funds into a legitimate, useable form of currency (Dupont, 2012). Like other profit-oriented illicit activity, avoiding detection involves obscuring the proceeds of stolen financial data. Market actors referred to as facilitators seize these opportunities by providing the marketplace with specialized labour for the monetization of stolen funds (Bouchard, 2007b; Holt, 2011, 2013b; Hutchings & Holt, 2015; Morselli & Giguère, 2006; Zhang & Chin, 2002). Motoyama et al. (2011) found that approximately 5% of threads posted to carding forums advertised money laundering services, although recent studies suggest that the extent and frequency in which these activities occur through these settings is much greater (Richet, 2013). Proceeds accrued through stolen data markets are most commonly laundered through money wire transfers or online payment systems like MoneyGram®, PayPal™, Ukash®, WebMoney©, and Western Union®, with cashiers taking a percentage of funds as a service fee (Holt, 2011; Wehinger, 2011). Funds may also be recouped from compromised accounts through informal cash couriers, referred to as drops or money mules, who receive transactions at a disclosed location, withdraw the funds into physical cash, and relay the money directly to an account controlled by the money launderer in exchange for a commission (Richet, 2013).[9]

[8] Researchers have noted that the actual terms transactions are most frequently negotiated in real time via online chat rooms (Holt & Lampke, 2010; Holt, 2013b; Holt et al., 2015; Motoyama et al., 2011).

[9] Alternatively, some offenders in the possession of stolen financial data also make purchases for items, most commonly electronics, after which the value of the goods are remitted through ecommerce or reshipping scams, or are fenced through online auction sites like eBay™; however, the details of these schemes are not typically disclosed through stolen data markets.

2.4.2. The Financial Lure of Stolen Data Markets

Why does a marketplace exist for stolen data and related fraud services? Why don't data thieves and/or vendors make personal use of stolen financial data, instead of selling it for a fraction of its true value? Few researchers have even pondered this question, let alone attempt to arrive at a satisfactory answer, as just two hypotheses have been forwarded within the cybercrime literature to explain the emergence of stolen data markets. One hypothesis is that stolen data markets are 'lemon markets' that exist only for fraudsters to defraud other fraudsters (see Herley and Florêncio, 2010; Holt et al., 2013; Yip et al., 2013). Another hypothesis forwarded by Herley and Florêncio (2010) is that suppliers to the marketplace either do not understand how to monetize data or are unable to do so due to its poor quality (see Holt, 2011 and Holt et al., 2016 for discussions on differences in data quality).[10] Otherwise, data thieves would not be incentivized to sell their earnings at discount prices.

Though a third and more likely reason for the emergence of stolen data markets is feasibility issues associated with monetizing the proceeds data theft, as the volume of data exposed through hacking or phishing is simply too voluminous for small groups of actors to use in a reasonable length of time. Once system administrators become aware of a network breach, it must be reported to the proper authorities and affected accountholders in a timely fashion.[11] Suppose, for instance, that detection occurs rather quickly. For data breaches resulting in tens, if not hundreds of thousands of records, monetizing such a large volume of data in a short period of time is not feasible, leaving thieves to forfeit opportunities to use the funds. Thus, it is much more likely the case that a marketplace for

[10] As noted by Holt et al. (2016) stolen data of lower quality consisting of credit and debit card numbers does not include information pertaining to the account holder, but does contain the corresponding credit verification value that enables buyers to immediately use the funds for purchase or exchange through online payment systems. Thus, purchase of lower quality data does not (necessarily) need to be laundered to obscure the identity or location of the buyer. But higher quality data, consisting of credit and debit card numbers along with credentials of the account holder, needs to be obscured through money laundering services following purchase.

[11] In June, 2015, the Canadian Parliament amended the *Personal Information Protection and Electronic Documents Act* (PIPEDA) by passing Bill S-4, *The Digital Privacy Act*. Bill S-4 mandates that organizations comply with notification, disclosure, and recording requirements for all security breaches involving personal data under their control to the proper authorities and individuals affected.

stolen data exists so that hackers and the like are ensured to capitalize on data theft by selling the data to other fraudsters, who are then responsible for the funds.

Recent research examining the impact of economic forces on outcomes within stolen data markets support this hypothesis. Nagurney's (2015) research, for instance, found that the economics of stolen data markets are similar to those impacting the global drug market. As with the costs associated with global shipping and distribution that impact prices in domestic drug markets, Nagurney's findings indicates that the associated costs of data theft and its distribution through the supply chain in stolen data markets impacts the pricing structures set by vendors. But the set prices are not fixed, as stolen data is a perishable commodity or has a 'best before' date. The time lapse spanning data theft and its delivery to the market for sale has a negative affect on advertised price, which continues to drop over time if a vendor does not find a buyer. Although Nagurney's study was the first to apply stringent economic models to data gathered from stolen data markets, Franklin et al. (2007) were amongst the first researchers to recognize that the economic principles of supply and demand impacted the volume of activity that occurred within stolen data markets. By observing trends in the price for advertised products and services, Franklin and colleagues were able to infer the quantity of goods available within markets to produce crude estimates of the total revenues that could be generated from the advertised products.

Recent studies have developed more sophisticated methods to further this line of inquiry. Holt and colleagues (2016) investigated the relationship between price and potential revenues that could be earned by vendors through exchanges with prospective buyers. Conservative estimates revealed that vendors had the potential to earn revenues in the hundreds of thousands of dollars, while upper bound estimates ranged from USD$1 – USD$2 million. Because vendors sell financial data at only a fraction of it's net value, Holt and colleagues found that buyers stand to substantially profit much more from these exchanges. For purchases of smaller lots of data, buyers could accrue revenues ranging from an estimated USD$500,000 to more than USD$2.5 million, but surpassed net gains of USD$10 million assuming 100% reliability in the data. For purchases of large lots of data, revenues may exceed an estimated USD$20 million.

Although Holt et al.'s analysis did not account for temporal trends in generated revenues. Instead they calculated estimates of potential earnings 'over the lifetime of observed activity' for which there is no timeframe indicated. Using a similar analytic approach, Bulkah and Gupta (2015) estimated that monthly revenues generated through stolen data markets range from USD$180,000 to USD$270,000 per month, projecting USD$2 to USD$3 million per year, consistent with the lower bound estimates derived by Holt et al. Though Reuter (2013) acknowledges that estimates of illicit revenues are not feasible due to the dark figure of the underlying predicate crimes and the covert nature of illicit financial transactions, these figures nonetheless suggest that opportunities for market offenders are potentially lucrative and outweigh the associated risks of detection and arrest or being scammed by unscrupulous actors.[12]

2.4.3. Trust and Uncertainty

Recent research indicates that market principles other than supply and demand bear influence on potential revenues that may be earned in stolen data markets. Research by Holt and colleagues (2013) suggests that a vendor's reputation and trust had the most impact on advertised prices across various markets, as trusted vendors were able to fetch greater returns on their products, compared to vendors yet to establish trust. The emphasis on trust reflects the hostility of illicit marketplaces. But rather than the added threat of physical harm that characterizes illicit drug markets, for example (Eck & Weisburd, 1995), hostility and risk in stolen data markets is limited to the financial losses incurred by market actors. Financial losses are the result of unscrupulous market actors, colloquially referred to as *rippers*, who pose as vendors or facilitators to connect with prospective buyers and money launderers, and accept transfer of payment without delivering the advertised product(s) or service(s).

Due to the high level of ambiguity associated with fraudulent exchanges, these financial repercussions are relatively commonplace (Décary-Hétu & Laferrière, 2015; Yip et al., 2013). The exploits of alleged rippers are exposed to the larger marketplace through public denouncements posted to specified 'ripper threads' (Holt, 2013b; Holt et al., 2015;

[12] Implied in Reuter's (2013) argument is that the estimated revenues of shadow economies like stolen data markets either under/overestimate the scope of illegal revenues.

Yip et al., 2013). The acknowledgement of shady actors arouses distrust and uncertainties within the marketplace that can have a substantial impact on both advertised price and the volume of market activity—threatening the underlying profit motives of market actors (Herley & Florêncio, 2010). Holt and colleagues (2013) found that in marketplaces plagued with ripping, advertised prices for stolen data were lower, although Décary-Hétu and Laferrière (2015) found that extent to which the market activity is affected depends on the reputation of the accused vendors.[13]

Eck (1995) highlighted issues associated with trust and uncertainty in illicit marketplaces, in what he called the 'dilemma of mutually accessibility'—the dynamics connecting buyers and vendors when the nature of the exchange is very risky. Repeated contact between buyers and vendors fosters trust and mitigates risks (Afroz et al., 2013), though contact in illegal markets is often short lived, often consisting of just a single transaction (Morselli et al., 2015). Such highly volatile, risky relationships necessitate other entrenched security mechanisms that serve to regulate behaviour and maximize the marketplace's performance (Holt et al., 2016). For this reason, the trust endowed by one's reputation may be the most important factor impacting the dynamics and organization of stolen data markets (Décary-Hétu & Laferrière, 2015; Dupont, Côté, Savine, & Décary-Hétu, 2016). To establish favourable reputation within the larger marketplace, vendors and facilitators undergo a vetting process that includes multiple layers meant to ensure the trust of buyers and those seeking to launder newly acquired funds. This vetting process is managed by market administrators and/or moderators, who perform duties that make them analogous to regulators or place managers within legitimate marketplaces (Eck, 1995).

The first phase of this vetting process involves testing advertised products and services to ensure they are in fact legitimate, such that they are of good quality and are useable. Upon completion of these quality assessments, administrators and/or moderators post a public review to specified 'reviewer threads' that detail the tested

[13] Décary-Hétu and Laferrière note that banishment of reputable vendors from the marketplace for fraudulent conduct affected the volume of market activity in the short term, although the market was resilient to these internal shocks. In contrast, the removal of vendors yet to acquire trust within the marketplace occurred almost daily without impacting the volume of market activity.

product(s) or service(s) and their recommendations for purchase (Holt, 2011). Similar to the vetting process in illegal drug markets, buyers are also expected to provide public reviews that vouch for or denounce vendors and the quality of their products and services. Regardless of positive or negative feedback, research indicates that the reputation and longevity of buyers within the marketplace influences the perceived validity of their claims (Holt et al., 2015). By publically reviewing the products or services offered by vendors, reputable buyers play an active role in cleansing the marketplace of unscrupulous actors. Any disputes that arise between market actors through these public exchanges are acknowledged by market administrators and a temporary status is assigned to the vendor and/or facilitator until the issues are resolved, serving as a caution to other prospective buyers (Holt, 2011; Holt et al., 2015).

Vendors and facilitators that accrue endorsements and positive feedback through these processes, while avoiding negative feedback, are accredited with verified status.[14] As in other illicit marketplaces (Morselli, 2003), reputable status as a verified vendor in stolen data markets is acquired over the course of a career. Motoyama and colleagues (2011) found that prior to receiving their first ratings, roughly 50% of vendors posted 60 advertisements in more than 50 threads and were involved in more than 30 private exchanges with an average of 13 prospective buyers—a substantial amount of activity considering the high population displacement in online illicit marketplaces (Hutchings & Holt, 2015). Thus, verification status is not distributed freely or randomly, but rather, obtaining verification requires serious commitment on the part of vendors and facilitators; however, the return on social capital following acquisition of verified status may reduce barriers that would otherwise impede exchange (Moore et al., 2009). Research by Holt and Lampke (2010) found that the majority of buyers evade unnecessary risks by limiting contact to only trusted vendors and Motoyama et al. (2011) found that verified vendors are solicited 2 – 3 times more than unverified vendors. In a marketplace otherwise void of indicators of trust, verification status appears to be a legitimate mechanism to regulate market activity and ensue trust between actors.

[14] Negative feedback can lead to verified vendors surrendering this status due to reduced confidence in the products and services offered and/or poor level of customer service (Holt, 2011).

But in the absence or breakdown of these strong internal regulations, it is difficult to foster trust in the marketplace (Franklin et al., 2007; Holt & Lampke, 2010; Holt et al., 2015). Under such circumstances, or for buyers seeking increased security, facilitators who offer escrow services, referred to as guarantors, are included in transactions to serve as financial intermediaries (Herley & Florêncio, 2010; Holt, 2013b; Holt & Lampke, 2010; Holt et al., 2015), who are entrusted to ensure that both buyers and vendors fulfil their respective roles in the transaction. Similar to the use of third party handovers in illicit drug markets (Eck, 1995), guarantors receive payments from buyers and hold it until vendors deliver the purchased product(s). As soon as buyers confirm the products or goods have been received, guarantors distribute payments to vendors. As triadic relationships reduce uncertainties associated with risky transactions (Burt, 2000; Eck, 1995), guarantors provide an added layer of security and trust to risky exchanges—especially for buyers connecting with unverified vendors; however, the added complexity that results from involving guarantors in transactions also has its drawbacks, including increased transaction times and costs (Holt et al., 2013). As price increases serve to offset the commission set by guarantors and because the increase in price is passed off to buyers, buyers remain tempted to engage directly with (unverified) vendors to minimize transactions costs and maximize profits (Zhang & Chin, 2003).

2.5. The Current Study

Two simple research questions form the basis of the current study: 1) what is the population size frequenting stolen data markets that is active, but remains hidden? and 2) what is the composition of markets with regards to buyer, vendor, money launderer, and facilitator populations? The current study used two related capture-recapture methods, Zelterman's truncated Poisson estimator and its extended covariate adjusted model (Zelterman's regression), to estimate the size of these offender populations. Data analysis involved samples collected from 3 websites that facilitate financial crimes and frauds using a custom web-crawler. Analysis consisted of a number of stepwise procedures. First considered was the size of each market, after which the population size of offenders frequenting any marketplace was extrapolated by taking the overlapping distribution across the 3 independent samples and applying Zelterman's models. While the observed

overlap between marketplaces was rare, results indicated that markets are perhaps not distinct entities, but may be better conceptualized as single components of a collective marketplace that is much larger in size than what can otherwise be observed. Estimates were then calculated by subpopulation to compare market composition within and across marketplaces. Findings have numerous implications towards a broader understanding of offender prevalence and how prevalence impacts the way in which researchers understand and study the dynamics and social organization in illegal online marketplaces.

Chapter 3.

Data and Methods

3.1. Data and Collection Methods

Data for analysis was collected from carding forums. As with other online forums, carding forums consist of a series of *sub-forums* that are dedicated to various interrelated topics of discussion related to the marketplace for stolen data, such as the products, goods, and services that are bought and sold, as well as software and vulnerabilities that enable data theft via the hacking of financial networks and related systems. Within sub-forums, participants begin new *threads* of discussion through which other participants can then reply through *posts*. Threads are public discussions to which all participants are privy; however, designated sub-forums, along with the threads and posts therein, may also be restricted to only those participants who meet specific criteria. To gain access to these restricted areas, forum participants must typically meet thresholds of forum-related activity, including exceeding a 'milestone' number of posts, or are invited or vouched for by other 'privileged' forum participants. Forums also have built in private messaging systems to facilitate private dialogue between two users. In carding forums specifically, these covert communication features are used to further arrange the terms of sale between market actors. Thus, an undisclosed number of exchanges between participants remain hidden.

The first step in the data collection procedures was to generate a list of potential candidates from which data was to be collected. As in previous research, keywords common to the jargon used in stolen data markets, such as *carding*, *ccv*, and *dump* were entered into Google© search engine to set the search parameters (Holt, 2011, 2013b; Holt & Lampke, 2010; Holt et al., 2013, 2015, 2016). Each of the 25 pages retrieved through the Google© search were manually revised, with 38 forums indexed on the public Internet that catered to English-speaking fraudsters. As this search was limited to websites indexed by Google©, these 38 forums are likely a gross underestimation of the true prevalence of

these websites.[15] Furthermore, as the Internet is constantly growing, new websites emerge, whereas others go offline or are shutdown due to legality issues. Thus, the number of websites identified was reflective of the carding scene only at the time of this study.

Characteristics of each of the indexed 38 carding forums were manually catalogued, including its URL address, list of sub-forums, and the then-current number of registrants, threads, and posts. Of the 38 forums, 13 were closed models that required either registration or invitation to access the website and/or its content. The other 25 forums were open models that did not restrict accessibility to content. Put another way, non-registrants or guests cannot access and view the content of closed model forums, but these same restrictions do not apply to open model forums, such that they are freely and publically accessible websites. Only open model forums were considered for data collection to comply with Simon Fraser University's (SFU) Policy R20.01. In accordance with paragraph 7.3, this study is exempt from SFU's Research Ethics Board review process, because of the public nature, availability, and accessibility of the data. To address any outstanding privacy concerns, research considerations forwarded by Holt (2010, 2015) were adopted. Particularly, the URL addresses or domain names of the forums subject to analysis and other potential identifying features, including user pseudonyms, are not disclosed in this study.

Of the 25 candidate forums, 6 were selected for data collection. But the sheer volume of content archived on carding forums poses analytic challenges concerning how best to collect the data to be analyzed. Because the data is user-generated, it should be collected in such a fashion that ensures the researcher does not contaminate it (Holt, 2010, 2015). Researchers have previously used manual data collection techniques to extract data from online forums (Holt, 2011, 2013a, 2013b; Holt & Lampke, 2010; Holt et al., 2013, 2015, 2016), though these methods are labour-intensive and not practical for the collection of large quantities of data. By comparison, automated methods streamline

[15] Not identified are websites that are hosted on anonymity networks, or 'darknets', such as the Tor anonymity network or even those that are not indexed by Google©, but nonetheless leverage the infrastructure of the public Internet.

data collection efforts and ensure that data is compiled in a systematic, structured way (Décary-Hétu & Aldridge, 2015). For this study, a custom designed web-crawler was used to scrape the data posted to the 6 selected carding forums.[16] The web-crawler began its process by indexing the many webpages comprising each website. Due to the thematic grouping of sub-forums and the chronological listing of threads, each webpage follows more or less a ubiquitous design. Each thread is titled and contains a denoted number of replies (the user-generated content). This design is consistent across threads and forums more generally. After all webpages were catalogued for each of the 6 websites, the sought after data, including thread titles, user pseudonyms, user date of registration, user-generated text, and date of the user-generated text was sequentially downloaded and stored in a corresponding database from which it could be extracted into raw text format.

Due to time constraints imposed on this study, among the first considerations to website selection were the technical challenges posed by the underlying design of each, as websites with poor and/or complex design structure may have potentially impeded data capture. As the design of each website is unique, websites that demanded few technical adjustments to the software were preferred. Outside of these technical concerns, the other important consideration for website selection was its size—measured by the cumulative population of known registrants and the volume of user-generated content (e.g., threads and posts). Collecting large volumes of data was necessary to derive adequate population estimates, so larger websites were preferred.

But the forums that were selected for analysis are also particularly relevant for several reasons. First, due to accessibility issues to data, previous studies have largely analyzed data extracted from dated forums (Afroz et al., 2013; Décary-Hétu & Laferrière, 2015; Holt et al., 2015, 2016; Hutchings & Holt, 2015; Motoyama et al., 2011; Yip et al., 2012, 2013), whereas those targeted for data collection for this study were active online at the time of analysis. As a result, the results reflect the dynamics of the current

[16] This software was developed by Dr. Richard Frank, faculty of the School of Criminology at SFU, with expertise in computational criminology. The web-crawler used in this study is a modification of a previous design used to identify networks of hyperlinked websites that collectively distribute child sexual exploitation material (Frank, Westlake, & Bouchard, 2010; Westlake & Bouchard, 2016; Westlake, Bouchard, & Frank, 2011) and extremist propaganda (Bouchard, Joffres, & Frank, 2014).

marketplace for stolen data. Second, to saturate the market with their products and services, vendors and facilitators may be motivated to conduct commerce in open, public settings. The forums selected for this study are public marketplaces and while not as secure as closed or private markets, public marketplaces have the advantage of reaching the majority of fraudsters who are likely less technically inclined or do not have the prerequisite social capital needed to participate in more insular markets. Third, the selected forums operated on a partial closed model, meaning that parts of the websites and their content were restricted and not able to be collected by the software. Websites with content restrictions are argued to be more insular than fully open model forums (Holt, 2013a; Holt & Lampke, 2010; Holt et al., 2013, 2015, 2016). By targeting websites with these privacy features, some of the virtues of closed model forums were thought to be obtained in the data samples. Finally, there were no qualitative differences across the websites from which data was sampled or those subjected to other research (e.g., Afroz et al., 2013; Décary-Hétu & Leppänen, 2013; Décary-Hétu & Lefarrière, 2015; Holt, 2013b; Holt & Lampke, 2010; Hutchings & Holt, 2015; Motoyama et al., 2011; Soudijn & Zegers, 2012; Yip et al., 2012, 2013); therefore, there are no reasons to suspect that the results of this study would drastically differ by using data sampled from other stolen data markets.

3.2. Sampling Procedures and Data Processing

Once the data had been collected, a number of procedures ensued to process and sample the data. The first procedure concerned sampling the large volumes of data that were collected across the 6 forums.[17] To arrive at manageable, yet meaningful data samples, a number of parameters had to be set. First and foremost, a logical timespan had to be selected from which to estimate the population sizes of offenders frequenting stolen data markets. For drug dealing, marijuana cultivation, and sex offending, 3 year timeframes produced suitable estimates of the respective populations, as there was a long enough observation period for criminals to be arrested and/or imprisoned, released from custody or incarceration, and reoffend (Bouchard, 2007a; 2008; Bouchard & Lussier, 2015; Bouchard & Tremblay, 2005); however, with regards to the timespan for measurement in the current study, 1 calendar year makes sense from a conceptual and

[17] Data collection retrieved a total of 121,572 threads and 824,838 posts.

methodological standpoint, as crime is measured on a per year basis. One calendar year was also deemed to be a reasonable and sufficient timespan to observe actors frequenting multiple markets.

Of the 6 forums from which data was collected, 3 overlapped for the entirety of the 2015 calendar year (January 1 – December 31). From each of these 3 concurrent forums, only threads posted to relevant sub-forums, or those dedicated to market activity, during 2015 were extracted from the database to comprise the samples for analysis. This meant that threads posted prior to 2015, but that contained posts from 2015 were also discarded. Using data generated during the most recent overlapping year between forums was considered advantageous for two reasons. First and foremost, to construct datasets suitable for capture-recapture analysis, it must be possible for market actors to be found in multiple samples. By using mutually exclusive, yet concurrent data sources, the probability of capturing market offenders who frequent multiple online marketplaces was theoretically possible. Additionally, this specific timeframe provided the most current snapshot of the online marketplace for stolen data.

The next set of procedures involved filtering noise from this data. For previous studies which sought to distinguish and classify relevant from irrelevant content, the presence of noise within the data was necessary to address the research objectives (see Burnap & Williams, 2015, 2016; Williams & Burnap, 2016; Williams et al., 2016). But with regards to the objectives of the current study, excessive noise within the data would merely skew estimates one way or the other, resulting in much higher degrees of measurement error. This nose had to be filtered from the data samples to avoid artificially inflating the size and scope of the data, and by doing so, more precise estimates of market populations could be derived. Noise consisted of mundane and redundant discussion, such as items purchased with stolen funds, tutorials on how to defraud legitimate online marketplaces, such as Amazon™ and eBay™, and spammed advertisements posted to the forums message boards. Such grievances are guarded against by the population at large and are typically regulated, and corrected for, by forum administrators (Holt, 2011). But despite these persistent efforts, a certain degree of noise is sure to ensue.

Table 3.1 Samples used for analysis

Forum statistics	**Marketplace 1**	**Marketplace 2**	**Marketplace 3**	**Across Marketplaces**
Threads	308	358	225	891
Posts	971	2,854	1,441	5,266
N	407	552	393	1,325

Duplicate threads were deleted, followed by the removal of non-alphanumeric characters from the user-generated text data, after which all text was converted to lowercase format. As the data needed to be manually revised and coded, these latter procedures rendered the text data easier to read and decipher. During revision, any and all threads not involving the sale, purchase, and monetization of financial data—bank account numbers and login credentials, all types of credit card (*cvv* and *cvv2*, *dumps, track 1* and *2*, and *fullz*), and debit card data—or money transfers involving bank, MoneyGram®, PayPal™, Ukash®, WebMoney©, and Western Union® accounts and the like were removed from the samples. The volumes of data that ensued following data processing comprised the final samples for analysis, shown in Table 3.1. The resulting samples were mutually exclusive and represented the volume of activity generated by market actors openly operating in each of the 3 marketplaces, which served as the basis for capture-recapture analysis for each of the independent marketplaces. These samples were also consolidated into a single dataset, as the cross-sample distribution reflected a generalization of the number of market actors frequenting any marketplace where stolen financial data is bought, sold, or laundered during the observation period.

3.3. Capture-Recapture: Methods and Assumptions

The size stolen data markets are best assessed through a set of methodologies that have been used to estimate the size of other illegal markets (Bouchard, 2007a, 2008; Bouchard & Tremblay, 2005; Bouchard et al., 2012). Market actors with 1, 2, …*n* contacts occur, but those with 0 contacts are not observable, which necessitates the application of

capture-recapture methodologies to estimate the size of the hidden population that shares similar characteristics to the known population (Böhning & van der Heijden, 2009). In the context of online forums, the '0s' represent the *lurkers*—registered forum participants that do not engage in or contribute to open dialogue, but are nonetheless active for an unknown duration of time. The reasons for lurking are largely unclear, yet lurkers represent a considerable proportion of the population frequenting online forums (Nonnecke & Preece, 2001; Preece, Nonnecke, & Andrews, 2004; Rafaeli, Ravid, & Soroka, 2004). With regards to the research objectives of the current study, the lurkers in carding forums represent the 'dark figure' of offenders participating in stolen data markets that are to be estimated using capture-recapture methods.

Capture-recapture methods are based on the underlying assumption that reoccurring patterns within data reflective of a population can be modeled to derive an estimate of the population that is active, but remains hidden (Böhning, 2010). Data is collected at one or more observation periods and each individual observed is recorded in the capture history that usually spans three or more capture periods. The logic behind the capture-recapture process is that previously encountered individuals will eventually be recaptured and from this distribution, estimates of the larger population can be derived. But for this logic to hold true, the population under study must meet three assumptions: 1) the population is closed, in which no entry or exit occurs; 2) the population is homogenous; and 3) the probability of capture is held constant throughout the observation period. If these assumptions are met, the number of observations, or captures, follow the Poisson distribution (van der Heijden et al., 2014) and the population estimate maximizes the likelihood of observed frequencies.

But estimating the size of criminal populations poses a series of challenges. First and foremost, the volatile nature of illicit activity dictates that offending is not constant or uniform across criminal populations, but that incidents are more likely to be infrequent and cluster together; however, Poisson models are specifically designed to model such rare events and their application in this context negates many potential issues that would

otherwise arise using models designed for non-count distributions.[18] Second, more serious offenses, including financial and market crimes, have smaller populations of offenders (Clarke & Weisburd, 1990). The transient nature of offending and offender populations triggers issues associated with the application of capture-recapture methods, primarily the necessity for large samples if observations are infrequent (van der Heijden et al., 2014). Such issues are thought to be addressed through the volume of data that was collected and extracted for analysis (See Table 3.1). Relatedly, the third obvious issue is that data may violate particular assumptions associated with capture-recapture that hinder efforts to produce estimates of offender prevalence (Tcherni et al., 2016). Capture-recapture assumptions are indeed commonly violated by offender populations, each of which must be adequately addressed through careful considerations to choice of estimation model, research design, and data analysis.

3.4. Zelterman's Truncated Poisson Estimator

Truncated Poisson models provide the necessary framework of methods required to estimate populations of hidden offenders. Zelterman's (1988) truncated Poisson estimator is one such model that has proven to be a rigorous method for estimating the size of criminal populations. Compared to other capture-recapture methods that more strictly adhere to the assumptions of the Poisson distribution (Böhning & van der Heijden, 2009), Zelterman's estimator relaxes these assumptions by remaining robust against model misspecification (Böhning, 2010). Zelterman (1988) argued that the assumptions for capture-recapture methods are likely not a true reflection of the Poisson distribution over the entire range of the count, but that it may better reflect a truncated Poisson distribution. The truncated Zelterman estimator, shown in Figure 3.1, considers only initial captures and recaptures to derive estimates:

[18] Poisson distributions represent count data and are used to model rare events or a small number of discrete independent observations, such as natural disasters, infections resulting from an epidemic, and deaths resulting from war or genocide, but are also applicable if non-observations or non-occurrences cannot be counted (Garson, 2013).

$$Z = \frac{N}{1 - e^{\left(-2 * \frac{n2}{n1}\right)}}$$

Figure 3.1 Zelterman's truncated Poisson estimator

where *Z* is the total population, *N* is the total number of individuals observed during the observation period, *n1* is the number of individuals observed once, and *n2* is the number of individuals observed twice in the observation period.

By discounting the observed activity from the minority of repeated offenders, Zelterman's model minimizes the effects of population heterogeneity first raised by Greene (1984).[19] The rationale for these omissions is that offenders who are not as visible are more meaningful for the estimation of hidden offenders. The advantages of Zelterman's estimator include its simple design, which makes it easy to fit to data, and that it works rather well with contaminated distributions (Böhning & van der Heijden, 2009).[20] Another distinct advantage is that it can produce estimates from just a single dataset, whereas other capture-recapture methods require at least two, but usually three or more complementary datasets, to derive estimates.

3.4.1. Controlling for an Open Population

Despite relaxing assumptions traditionally associated with capture-recapture methods, Zelterman's estimator still assumes that the population of interest is a closed population. In closed populations, the entire population, observed and hidden, remains static during the entire observation period. If this assumption is violated, the population will be overestimated (Böhning & van der Heijden, 2009), although adhering to the closed population assumption poses obvious challenges when measuring offender prevalence. Offender mobility, detection avoidance, or desistence from offending are relevant examples that illustrate why the assumption of a closed population is not simply met (Bouchard, 2007a; Rossmo & Routledge, 1990). But violating this assumption is not a

[19] Repeated offenders not only offend more frequently, but also commit more serious offences over longer durations of time (Piquero, Farrington, & Blumstein, 2003).

[20] Note that substantial contamination in the data nonetheless yields overestimates in population size (Böhning, 2010).

major limitation (see Kendall, 1999) if, for instance, the window periods between capture are relatively short, as few individuals are likely to enter or be displaced from a population in a relatively short span of time. Under these circumstances, population displacement is unlikely to be large enough to radically effect population estimates. Although the Internet is void of temporal and spatial factors that have traditionally limited offender mobility, such that displacement is commonplace in online marketplaces (Hutchings & Holt, 2015).

This being the case, adequate precautions must be taken to control for the open populations of stolen data markets. One way to control for violation of the closed population assumption is to ensure that capture periods span an appropriate length of time to scope the population under study. If capture periods are too long, the accuracy of any estimate would be in question. On the other hand, capture periods that are too short affect the probability of capture. Because population displacement was a real concern, shorter window periods between captures were thought to provide more assurance to the accuracy of the derived estimates and offset violation of the closed population assumption. Previous research indicated that average participation trajectory in hacker chat rooms did not exceed 2 weeks and that by one month the majority of population were displaced (Benjamin & Chen, 2014); therefore, 12 capture periods were implemented to scope populations on a monthly basis. That is, the capture process was continuous, recording which and how many market offenders were active in each of the 12 month-long capture periods. This timed measurement was easily implemented, given the fact that the user-generated text posted to online forums is timestamped.

By implementing short, successive capture periods, the effects of violating the closed population assumption were substantially reduced, if not negated. For the consolidated dataset, the capture process counted the number of times actors were identified across the 3 samples, each spanning a year in length. As the cross-sample distribution is a generalization of a much larger marketplace, the data theoretically models the entire hidden population that is active across all stolen data markets within a year timespan. Under such parameters, the closed population assumption is also not violated, given that the likelihood of severe departure from this assumption is minimized by the analysis of capture at an aggregate level, which theoretically encompasses all marketplaces active online at the time of data collection and analysis. The corresponding

capture frequencies for each of the samples are shown in Table 3.2. The small degree of pseudonym overlap across datasets lends credence to the use of capture-recapture methods and suggests participation across markets is rare.

3.5. Zelterman's Regression

Criminal populations are also not homogeneous. Even in relatively homogenous populations there is still heterogeneity present that is hidden due to the covert or transient features of the population, especially if data collection spans longer periods of time (Rossmo & Routledge, 1990). Greene (1984) first highlighted the importance of removing heterogeneity when deriving estimates of criminal populations, after reanalyzing data from a previous study (see Greene & Stollmack, 1981). To account for heterogeneity within a population of general offenders in Washington DC, models were fit to each of the 20 distinct age categories characterizing the samples. Compared to the results of Greene and Stollmack's (1981) study, the data did not fit the models well, and the parameter estimates and the heterogeneity were factors in the poor model fit. Thus, if ignored, heterogeneity causes error in the form of downward bias in population estimates (van der Heijden et al., 2013). The drawback of Zelterman's estimator in this context is that using truncated data provides more conservative estimates due to the inability to control for observed heterogeneity within the data, compared to other models, such as Chao's (1989) estimator, that do no remove heterogeneity—that is, take into account the full Poisson distribution of offender observations. In comparison, more complex capture-recapture models that enable the inclusion of covariates to account for heterogeneity lead to tighter estimates and provide additional context concerning the population under study (Rossmo & Routledge, 1990); however, these methods lack the flexibility of Zelterman's estimator model.

Table 3.2 Capture frequencies for N market actors active during 2015

Sample	*N* captures											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
Marketplace 1	389	18	0	0	0	0	0	0	0	0	0	0
Marketplace 2	466	45	12	13	2	2	4	2	1	3	1	1
Marketplace 3	349	28	8	2	2	1	1	2	0	0	0	0
Across Marketplaces	1,299	25	1	–	–	–	–	–	–	–	–	–

To address this issue, Böhning and van der Heijden (2009) extended Zelterman's estimator into a maximum likelihood framework that leverages the principles of logistic regression to measure the impact of covariates on population estimates. What Böhning and van der Heijden called 'Zelterman's regression' was published as a STATA macro and was used to calculate the regression parameters for the capture-recapture distributions derived from each sample. Unlike Zelterman's estimator model, confidence intervals are calculated along with the population estimates, larger ranges in which estimates may plausibly be confined, along with goodness of fit statistics to assess model fit to the data. While construction of a baseline model in which no covariates are included is the direct equivalent to Zelterman's estimator (Figure 3.1), the covariate adjusted model controls for heterogeneity within the data. Added covariates not only provide an added degree of context into the estimate, but also determines which characteristics within a population are associated with increased or decreased probability of capture. But the added covariates only control for observed heterogeneity that is ignored using Zelterman's estimator and cannot control for hidden heterogeneity within the data. Regardless, previous research indicates that Zelterman's regression is robust despite this limitation (Böhning & van der Heijden, 2009; Bouchard & Lussier, 2015; Bouchard et al., 2012). Estimates of the covariate adjusted model (G^2) are larger than those produced by Zelterman's estimator if the model is statistically significant.

Why are estimates larger? By factoring in characteristics of the known sample that are otherwise not considered through the estimator model, the covariate adjusted model increases the complexity associated with the calculation of estimates, thereby broadening the scope of the population. The added complexity increases uncertainty, but also precision, in the derived estimates. Though any uncertainties associated with the estimates produced by the covariate adjusted model are adequately addressed through model specification. Estimates will not greatly vary from those produced by Zelterman's estimator when the covariate adjusted model is not significant, as added covariates do not control for observed heterogeneity in the data or effect the probability of capture. Furthermore, nested models can be compared to eliminate unnecessary covariates and determine the most parsimonious model, but if the covariate adjusted model results in poor model fit, Zelterman's estimator is by default the preferred model.

3.5.1. Controlling for Heterogeneity within a Population

Grouping actors into subpopulations is one means to account for heterogeneity within otherwise homogenous populations and has the added benefit of producing estimates for each subgroup (Bouchard, 2007a, 2008; Bouchard & Tremblay, 2005; Bouchard et al., 2012; van der Heijden et al., 2013, 2015). Perhaps the most obvious difference in the composition of stolen data markets is role occupancy among actors. The marketplace is frequented by subpopulations of buyers, vendors, money launderers, and facilitators seeking to purchase and monetize funds and while their underlying behaviour collectively drive the economics and social organization of the marketplace, the motivations of these subpopulations differ. Through manual revision, the actors comprising the samples were coded as belonging to one or more of these subpopulations, each of which were denoted by dichotomous variables. In addition to these subpopulations, the risk aversive behaviour of market actors, including the usage of financial intermediary escrow services and participation in verified markets, were also denoted by dichotomous variables. Market activity resulting from these engagements was also thought to be particularly relevant for controlling for heterogeneity within the samples, as the increased transparency involved in these transactions distinguishes actors seeking increased security from others who are willing to incur greater risk. The rationale is that these mechanisms force actors to engage in open market activity, thereby serving to regulate behaviour in a marketplace that is otherwise characterized by shady actors engaging in risky transactions.

Capture frequencies by covariate are shown in Table 3.3. Besides the differences in observed frequencies of market actors within marketplaces, the covariate frequencies for the risk aversive behaviour indicate an interesting development: many actors use escrow services and participate in verified markets in marketplace 2, indicating that trust and confidence in the market is high, whereas frequency of actors who leverage these mechanisms of trust is lower in marketplace 3, and much lower still in marketplace 1, suggesting lower levels of trust and much higher degrees of uncertainty, which increases opportunities for unscrupulous actors to defraud others in the marketplace.

Table 3.3 Capture frequencies by covariate for N market actors active during 2015

Covariates	Marketplace 1				Marketplace 2				Marketplace 3				Across Marketplaces			
	f_1	f_2	f_{3+}	$\sum$	f_1	f_2	f_{3+}	$\sum$	f_1	f_2	f_{3+}	$\sum$	f_1	f_2	f_3	$\sum$
Subpopulation																
Buyers	204	11	0	**215**	182	13	18	**213**	90	12	4	**106**	516	13	0	**529**
Vendors	91	6	0	**97**	161	17	31	**209**	105	5	15	**125**	401	15	1	**417**
Moneylaunderers	68	2	0	**70**	96	25	9	**130**	161	18	2	**181**	370	9	0	**379**
Facilitators	62	1	0	**63**	77	8	20	**105**	56	4	0	**60**	216	11	0	**227**
Risk Aversive Behaviour																
Escrow services	9	2	0	**11**	83	21	13	**117**	73	12	2	**87**	212	3	0	**215**
Verified markets	5	1	0	**6**	114	11	13	**138**	70	5	3	**78**	216	6	0	**222**

Subpopulation

Buyers were coded as those actors posting threads looking to buy stolen funds in any of its many forms or those responding to vendor advertisements giving or seeking contact information, or generally showing interest in what was being offered by vendors. *Vendors* were coded as those actors providing stolen funds to the marketplace through thread advertisements or responding to prospective buyers indicating the products they offer or by providing contact information. *Money launderers* were fraudsters who indicated that they were in the possession of funds—through previous transactions with vendors, but mostly through unknown means—that needed to be transformed into an obscure form of currency. *Facilitators* were the peripheral actors who took advantage of these opportunities by offering money laundering services, including cash outs and drops, either at a fixed price or for a percentage of the funds needed to be laundered. Actors who did not generate any related content were coded as *unclassified*, with a total of 474 unclassified actors (26%) across the samples. These actors were not included in analysis

for two reasons: 1) they did not have a clear role in the market and, more importantly, 2) these actors were not persistent in the marketplace over longer periods of time, so they would merely inflate the population estimates.

Risk Aversive Behaviour

Escrow services are security mechanisms provided by market actors known as guarantors, who are positioned as financial intermediaries to mitigate the risky nature of the exchange in online marketplaces. Actors who used these services, despite the role(s) they occupied in the market, were coded accordingly. Market actors were coded as partaking in *verified markets* if they posted in threads reserved for verified advertisements by vendors and facilitators, were otherwise assigned verification status, or were actor engaging with a known verified actor.

3.6. Assessing Violation of a Fixed Capture Probability

Although capture-recapture methods assume a fixed probability of capture for the population under study over the entire capture period, there is heterogeneity associated with individual capture probabilities within any target population (Böhning, 2010). Aside from role occupancy and inclination to engage in risk aversive behaviour, market actors also differ with regards to their longevity in the marketplace, or *career length*, and by volume of user-generated content, or *frequency* of activity, both of which may influence their probability of capture. The longer one's career length, the higher the probability they will be captured and recaptured during the observation period. Relatedly, as the majority of forum activity is generated by relatively few participants (Abbasi, Li, Benjamin, Hu, & Chen, 2014; Holt, 2013a), there are individual differences with regards to frequency of overt behaviour that influences one's visibility among the larger population. This latter point in particular infers that the assumption of a fixed capture probability is indeed violated, because the median frequency for each of the of subpopulations, in addition to that of the larger population, is 1. Thus, there is a decreased probability that any one market actor will be reencountered following their initial post, as the majority of the known population contributes just a single post to their respective forum(s).

Table 3.4 Descriptive statistics of career length and frequency for N market actors active during 2015

Subpopulations	Marketplace 1[a]	Marketplace 2	Marketplace 3	Across Marketplaces
Buyers	215	213	106	529
Career length	–	10.72	4.81	5.03
Frequency	1.30	6.50	2.30	3.60
N captures	1	1	1	1
Vendors	97	209	125	417
Career length	–	12.28	6.36	7.38
Frequency	1.80	10.40	3.80	6.80
N captures	1	1	1	1
Money launderers	70	130	181	379
Career length	–	9.06	6.57	6.22
Frequency	1.10	3.20	2.60	2.70
N captures	1	1	1	1
Facilitators	63	105	60	227
Career length	–	13.57	4.29	7.13
Frequency	1.10	9.00	2.00	5.10
N captures	1	1	1	1

[a] Career length not able to be derived for this sample

Due to the violation of this assumption, it was then necessary to assess whether and to what degree the characteristics of market actors varied by subpopulation. Distinct differences in career length and frequency would indicate that subpopulations have stark underlying differences, which would further influence their probability of capture and, thus, the derived estimates. Descriptive statistics of career length and frequency for the market subpopulations across samples are shown in Table 3.4. Career length was coded cumulatively during each observed capture period as the number of weeks elapsed since one's registration to the website to the time of last observed activity. Frequency was also coded cumulatively during each observed capture period, denoting the number of posts one made to marketplace threads. Across samples, vendors and facilitators tended to have the longest careers, along with higher post frequency, indicating that they were more

active in the marketplace over longer periods of time; however, these differences are not so stark as to have a substantial impact on the probability of capture, especially given 12 months spanning observation period. Within samples, the mean scores vary, but once again there are not substantial differences in the characteristics across subpopulations. The median number of captures is also shown for each subpopulation. As previously indicated, the capture distributions are highly skewed, such that the majority market actors are only captured once within and across samples.

3.7. Analytic Strategy

Two related capture-recapture methods, Zelterman's estimator and Zelterman's regression, were used to scope the populations frequenting stolen data markets, as well as the composition of subpopulations comprising the markets during 2015. These complementary models were selected specifically for analysis, as both have been proven to be effective at providing precise estimates of the populations comprising illicit markets (Bouchard, 2007a; Bouchard & Tremblay, 2005; Bouchard et al., 2012) and criminal populations more generally. Analysis was facilitated by STATA 14, using a macro developed by Böhning and van der Heijden (2009), who extended Zelterman's estimator into a covariate adjusted model. The results provided by Zelterman's models are presented as a series of stepwise procedures that involved each of the data samples. Each sample was analyzed separately to provide estimates for each market, but were also consolidated into a single dataset from which the number of market offenders frequenting any marketplace could be generalized.

First, it was necessary to establish baseline estimates using Zelterman's estimator model. These estimates were produced solely through the capture-recapture distributions extracted from the 4 data samples. Estimates were then improved through the covariate adjusted model. With the addition of covariates, denoting various subpopulations of market actors and risk aversive behaviour, the estimates produced for each market increased quite substantially. The third procedure involved a thorough examination of the parameters produced using Zelterman's regression model. A closer look at the coefficients provided indication as to why some models produced better estimates and indicated which of the added covariates had a statistically significant and meaningful impact on the

probability of capture. The fourth component of analysis involved reporting the estimates by subpopulation, along with their corresponding capture percentages. Results for each subpopulation provided context regarding the composition of each marketplace and how they differed, the composition across marketplaces, and which market offenders were most visible. The final component of analysis involved assessing the plausibility of the estimates—an integral, but often overlooked, part of any capture-recapture analysis. As no prior studies have sought to estimate the size of stolen data markets, comparisons had to made to similar samples as reported in previous studies (Motoyama et al., 2011; Yip et al., 2012; Zhao et al., 2016).

Chapter 4.

Results

4.1. Establishing Baseline Estimates

How many market actors were active within and across marketplaces during a year span, but remained hidden? This question was addressed through a series of stepwise procedures, the first of which involved fitting Zelterman's estimator to the capture-recapture distributions to produce initial working estimates for each marketplace. The results shown in Table 4.1 indicate that a substantial number of market actors who were active did not partake in open market activity. Estimates indicate that only between 4 and 6% of all market actors were observed across the samples. Coupled with the observable sample sizes, the estimator model produced estimates ranging from 3,044 (CI = 1,680–3,622) to 4,471 (CI = 2,484–6,725) for the 3 marketplaces and an estimated 36,415 (CI = 21,343–48,837) actors active in any online marketplace. Estimates had relatively tight confidence intervals, but if the models truly provided good fit to the data samples, then the capture distribution produced by the model should resemble the observed capture-recapture distributions. Because Zelterman's estimator model is void of goodness of fit indicators that are produced for Zelterman's regression, Neyman's chi-square test (X^2), given in Figure 4.1, was calculated for each of the 4 models.

$$X^2 = \sum \frac{(n_j - \hat{u}_j)^2}{{}^{j-1}n_j}$$

Figure 4.1 Neyman's X^2

where *n* is the number of observations in a sample that are captured or recorded, *j* is the capture frequency of *n*, and $\hat{u}$ is the theoretical distribution of *n* with *j* captures. Results showed that Zelterman's models provide good fit to the capture distribution of each sample ($X^2 = 0.00$), as the observed capture distributions significantly differed from their theoretical distributions. Notice that the null estimate produced for marketplace 1 (Z = 4,064) is substantially larger than the null estimates provided by the models for marketplaces 2 (Z = 2,781) and 3 (Z = 2,651), despite similarities in sample size. Capture-recapture

distributions that have smaller numbers of repeated captures provide better fit to the estimator model (Bouchard & Lussier, 2015). Because the capture-recapture distribution for this sample contained no actors captured 3 or more times—parameters that are omitted once the model is fitted to the data—Zelterman's estimator yielded larger estimates than the more conservative estimates of marketplaces 2 and 3.

Table 4.1 Frequency distribution for *N* captures of market actors in stolen data markets during 2015 and estimated prevalence of the active population

	Marketplace 1	Marketplace 2	Marketplace 3	Across Marketplaces
Z estimate	4,471	3,333	3,044	36,415
CI (95%)	(2,484 – 6,725)	(2,032 – 3,529)	(1,680 – 3,622)	(21,343 – 48,837)
N captures				
0[a]	4,064	2,781	2,651	35,090
1	389	466	349	1,299
2	18	45	28	25
3+	0	41	16	1
Neyman's X^2	0.00	0.00	0.00	0.00

[a] *Z* estimate denoting the number of active, but hidden, market actors ($\widehat{N}$)

A caveat of Zelterman's estimator model is that estimates are based on a fixed probability of capture for each market actor comprising the respective samples—an assumption which is not likely to hold any real validity. It is more realistic that capture probabilities differ within and across subpopulations, among actors who prefer to partake in more open, and transparent exchanges to minimize risk and for those that occupy dual roles within the marketplace. In fact, research suggests that capture probability is likely to differ amongst market actors given that few market actors establish many more contacts than others, such that tie formation is exponentially distributed (Holt, 2013a, 2013b). For instance, vendors establish more contacts than buyers (Décary-Hétu & Laferrière, 2015; Zhao et al., 2016), indicating that vendors are more active and, thus, more visible in open market settings. Another drawback of Zelterman's estimator model is that it is ill equipped to estimate the size of subpopulations embedded in a given sample; however, the

estimator model is useful in establishing baseline estimates from which its extended, covariate adjusted model can be compared. By factoring covariates into the estimation model, observed heterogeneity can be modeled—the added covariates correct for fixed capture probability and may result in better model fit to the data and, thus, more reliable estimates.

4.2. Improving Estimates with a Covariate Adjusted Model

For each of the capture-recapture distributions, three nested covariate adjusted models were fitted to the data. The null models do not include covariates, so the estimates are equivalent to those produced by Zelterman's estimator model. The partial models included covariates controlling for role occupancy among market actors assuming the roles of buyers, vendors, money launderers, and/or facilitators. In addition to controlling for each of these subpopulations, the full model included covariates that indicated whether actors were risk aversive and used escrow services and/or frequented verified marketplaces, which increase transparency and serve to regulate behaviour in a marketplace otherwise characterized by ambiguous and risky transactions. The increased transparency associated with these transactions was thought to potentially increase the likelihood of capture amongst actors comprising the samples. The Zelterman regression models are depicted in Table 4.2. Results show that adding covariates to the models inflated the population estimates and expanded the confidence intervals. Generally, the larger the estimate, the wider the confidence interval (Böhning & van der Heijden, 2009).

Provided the data fit the models well, the observed capture-recapture distributions within and across marketplaces should follow the distributions generated by Zelterman's regression models. Model specification should ultimately hinge on which provides the best fit to the data. Akaike's Information Criterion (AIC) is used to determine the model of best fit. AIC scores hold no intrinsic meaning, but are nonetheless useful for comparing nested models. As AIC penalizes models including nonsignificant covariates, lower scores indicate better model specification (see Bozdogan, 1987). Models are also judged by their corresponding *p* value, which indicates whether or not they are statistically significant. Model (G^2) scores and *p* values are not provided for the null models produced by Zelterman's estimator, only for the covariate adjusted model. Population estimates for

each model are denoted by $\widehat{N}$ and are calculated along with a 95% confidence interval. Because smaller ranges indicate more reliable estimates, confidence intervals are another means to assess model fit. Save for the models generated by the capture distribution from marketplace 1, Table 4.2 shows that the covariate adjusted models all provide good model fit ($p \leq 0.001$). In such circumstances, AIC scores are the decisive factor in the choice of model.

Table 4.2 Estimates of N market actors in stolen data markets during 2015

	AIC	G^2	*P*	$\widehat{N}$	CI (95%)
Marketplace 1 (n = 407)					
Null Model	149.46	–	–	4,064	(2,484 – 6,725)
Partial Model	153.73	3.72	0.45	6,226	(765 – 11,687)
Full Model	154.56	6.90	0.33	6,380	(1,127 – 11,632)
Marketplace 2 (n = 552)					
Null Model	334.01	–	–	2,781	(2,032 – 3,529)
Partial Model	287.85	54.16	0.00	5,921	(2,254 – 9,588)
Full Model	288.51	57.50	0.00	6,214	(2,293 – 10,134)
Marketplace 3 (n = 393)					
Null Model	201.47	–	–	2,651	(1,680 – 3,622)
Partial Model	192.23	17.24	0.001	4,063	(1,556 – 6,571)
Full Model	183.71	29.76	0.00	5,482	(1,507 – 9,458)
Across Marketplaces (N = 1,325)					
Null Model	250.00	–	–	35,090	(21,343 – 48,837)
Partial Model	210.39	47.61	0.00	79,860	(28,276 – 131,443)
Full Model	213.50	48.50	0.00	84,219	(25,859 – 142,578)

4.3. Estimating the Regression Parameters of Capture-Recapture Distributions

The Zelterman regression parameters for each of the capture-recapture distributions are reported in sequential order, beginning with marketplace 1 and working toward the consolidated sample representing a generalization of the larger marketplace

for stolen data. A thorough examination of each model's parameters provides further indication as to why one nested model may be preferred over another, which covariates have a significant and meaningful impact on the probability of capture, and whether the likelihood of capture increases or decreases accordingly. Interpretation of the results and what they mean in the market context are discussed throughout.

Table 4.3 Zelterman regression parameters for marketplace 1 (n = 407)

Regression parameters	Null Model		Partial Model		Full Model	
	MLE-*Z*	SE	MLE-*Z*	SE	MLE-*Z*	SE
Intercept	-3.07***	0.24	-3.88***	0.97	-3.69***	0.97
Subpopulations						
Buyers			0.93	0.92	0.62	0.93
Vendors			1.09	0.95	0.83	0.96
Money launderers			0.06	0.93	-0.05	0.95
Facilitators			-0.54	1.18	-0.70	1.23
Risk Aversive Behaviour						
Escrow services					1.33	0.89
Verified markets					1.29	1.20

*$p < 0.05$ **$p < 0.01$ ***$p < 0.001$

The parameters of the Zelterman's regression for data gathered from marketplace 1 are listed in Table 4.3. As indicated in Table 4.2, none of the added covariates are statistically significant due to the large value of their standard errors (SE) (Böhning & van der Heijden, 2009). In other words, none of the added covariates were found to have a significant impact on the probability of capture. In the context of this specific marketplace, this would suggest that occupying any of the 4 specified market roles does not increase one's probability of capture. Furthermore, covariates controlling for increased visibility through risk aversive behavior, or transactions involving escrow or verified markets, also do not increase the likelihood of capture. As such, the null model estimate provided by Zelterman's estimator (Table 4.1), is the preferred model. According to this estimate, only 9% of the actors, who were active in the marketplace during any point in time during 2015, were actually visible and for every 1 actor captured, 11 remained hidden.

Zelterman's regression parameters for the capture-recapture distribution extracted from marketplace 2 are listed in Table 4.4. The partial model (G^2 = 54.16; p < 0.001) provided the best fit to the data, modeling the heterogeneity of the marketplace's known population by controlling for the role occupancy of each actor. Results of the partial model show that role occupancy, whether it be that of a buyer, vendor, money launderer, and/or facilitator had a significant impact on the probability of capture. These covariates remained significant in the full model (G^2 = 57.50; p < 0.001), though the added covariates controlling for whether actors used escrow services or engaged in verified transactions did not have any significant effect on the probability of capture. Correspondingly, the model fit was slightly worse and the range of the confidence intervals was also slightly larger.

Table 4.4 Zelterman regression parameters for marketplace 2 (n = 552)

Regression parameters	Null Model		Partial Model		Full Model	
	MLE-*Z*	SE	MLE-*Z*	SE	MLE-*Z*	SE
Intercept	-2.20***	0.15	-4.81***	0.55	-5.01***	0.56
Subpopulations						
Buyers			1.09*	0.46	1.25*	0.50
Vendors			2.16***	0.54	2.29***	0.55
Money launderers			3.29***	0.54	2.88***	0.58
Facilitators			1.35**	0.45	1.35**	0.48
Risk Aversive Behaviour						
Escrow services					0.82	0.45
Verified markets					-0.07	0.47

*p < 0.05 **p < 0.01 ***p < 0.001

Estimates derived by the partial model show that the observed population (n = 407) is active among a much larger population of market actors (Z = 5,921; CI = 2,254–9,588). From this estimate it was inferred that just 8.82% of the actors convening in this marketplace during 2015 were actually visible. From these figures it can also be deduced that for every 1 actor observed in the sample, nearly 15 remained undetected. Compared to the other models, this estimate is more than 2 times larger than the estimate produced by the null model (Z = 2,871; CI = 2,032–3,529) and only slightly more conservative than

that produced by the full model (Z = 6,214; CI = 2,293–10,134). The coefficients for the partial model show that money launderers had the highest probability of capture, followed by vendors. Facilitators and buyers by comparison are much more transient in this market. These findings suggest that within this specific marketplace, money launderers are the most visible. Perhaps this increased visibility is the cost of seeking out specialized labour to launder stolen funds. As vendors are also more likely to be detected, this suggests that they too are more active in the open marketplace and for longer periods of time, compared to buyers and facilitators. From a criminal achievement perspective, this makes sense as lucrative opportunities can only be realized through continued advertisement of one's products and prolonged survival in the market.

Table 4.5 Zelterman regression parameters for marketplace 3 (n = 393)

Regression parameters	Null Model		Partial Model		Full Model	
	MLE-Z	SE	MLE-Z	SE	MLE-Z	SE
Intercept	-2.52***	0.20	-4.70***	0.67	-5.42***	0.74
Subpopulations						
Buyers			2.09***	0.54	2.61***	0.60
Vendors			1.23	0.72	1.76*	0.74
Money launderers			2.27***	0.64	1.55*	0.72
Facilitators			-0.28	0.59	-0.28	0.61
Risk Aversive Behaviour						
Escrow services					1.86**	0.63
Verified markets					0.65**	0.22

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

The regression parameters for the capture-recapture distribution derived from marketplace 3 are shown in Table 4.5. The full model provided the best fit to the data (G^2 = 29.76; p = 0.00) and modeled the heterogeneity of the marketplace's known population by controlling for role occupancy and risk aversive behaviour. The estimate produced by the full model (Z = 5,482; CI = 1,507–9,458) was 2 times larger than that of the null model (Z = 2,651; CI = 2,032–3,529) and 25% larger than the estimate calculated by the partial model, controlling only for role occupancy (Z = 4,063; CI = 1,556–6,571). This estimate

suggests that a mere 8.53% of actors convening in this marketplace during 2015 were actually captured and that for every 1 actor observed in the sample, nearly 11 remained hidden. The covariates in the full model are all significant and had a positive impact on the probability of capture, except for the dichotomous measure denoting facilitators in the marketplace, suggesting that facilitators are the most hidden actors in this market. As indicated by the coefficients, buyers had the greatest probability of capture, though actors operating with increased transparency in the open market, especially those dealing through escrow services, were also more likely to be captured than those willing to engage in riskier exchanges with potentially unscrupulous actors. The significant and positive values of these two coefficients suggest strong enforcement of security measures in this market, resulting in high degrees of trust and stability in the marketplace.

Table 4.6 Zelterman's regression parameters across marketplaces (N = 1,325)

Regression parameters	Null Model		Partial Model		Full Model	
	MLE-Z	SE	MLE-Z	SE	MLE-Z	SE
Intercept	-3.95***	0.20	-9.27***	0.85	-9.24***	0.85
Subpopulations						
Buyers			1.86***	0.45	1.80***	0.46
Vendors			2.57***	0.48	2.55***	0.48
Money launderers			1.63**	0.49	1.62**	0.51
Facilitators			1.68***	0.44	1.72**	0.45
Risk Aversive Behaviour						
Escrow services					0.42	0.55
Verified markets					-0.42	0.68

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

The parameters for Zelterman's regression for the cross-sample capture distribution are presented in Table 4.6. Recall that the samples from each of the 3 marketplaces were consolidated to generate a capture-recapture distribution of market actors who frequented multiple markets within 2015 to generalize an estimate for the number of actors who were active across all marketplaces. The estimate provided by the null model indicated that the market was comprised of 35,090 actors (CI = 21,343–

48,837). This estimate suggests that the aggregate sample comprises less than 4% of all actors who were active in stolen data markets during 2015; however, with improved model specification enabled by fitting the covariate adjusted model to the data, the scale of the marketplace more than doubles.

The partial model provided the best fit to the data (G^2 = 47.61; p = 0.00), so it is used for the interpretation of the results. All 4 covariates controlling for role occupancy are statistically significant and had a positive impact on the probability of capture. The value of the coefficients suggest that vendors experience the greatest probability of capture across marketplaces, followed in order by buyers, facilitators, and money launderers. According to the estimate produced by this model (Z = 79,860; CI = 28,276–131,443), 1.66% of all actors frequenting stolen data markets are represented in the cross-sample distribution. Despite the capture-recapture distribution (N = 1,325) reflecting a highly skewed proportion of actors who frequent only a single marketplace (n = 1,299), as it is much rarer for actors to span 2 (n = 25) or even 3 (n = 1) marketplaces, the model produced a valid estimate due to advantages of Zelterman's models and capture-recapture methods more generally, which are well-equipped to model such rare events (Garson, 2013).

4.4. Estimating the Size of Subpopulations of Market Actors

After estimating the size of individual markets and the marketplace as a whole, the next question of interest concerned determining the composition of the various subpopulations of market actors within and across markets. Deriving estimates by subpopulation provides substantive context to the estimates and may present additional insight into the economics and organization of stolen data markets. Table 4.7 shows the population estimates for the various subpopulations under study for each marketplace. Results show that the size of subpopulations fluctuates from marketplace to marketplace. This finding suggests that the composition of markets do in fact differ, such that not all markets are comprised in the same way—something that undoubtedly affects the dynamics within and across markets. Notice that the size of each subpopulation also exceeds the estimate provided by the best fitting models (Z estimate), the reason being that market actors may, and sometimes do, occupy more than one role in their respective

marketplace(s). Though the number of actors occupying dual roles was not large (n = 203), representing roughly 15% of the market, which supports the idea that specialization in offending is more common amongst online offender populations in comparison to generalization in offending amongst conventional offender populations.

Table 4.7 Observed and estimated prevalence amongst subpopulations

	Observed	Estimated	Capture %
Marketplace 1 (n = 407)			
Z estimate	407	4,064	10.01%
Buyers	215	2,432	9.83%
Vendors	97	1,097	8.84%
Money launderers	70	792	8.83%
Facilitators	63	713	8.83%
Marketplace 2 (n = 552)			
Z estimate	552	5,921	9.32%
Buyers	213	3,710	5.74%
Vendors	209	1,224	17.08%
Money launderers	130	323	40.25%
Facilitators	105	886	11.85%
Marketplace 3 (n = 393)			
Z estimate	393	5,482	7.17%
Buyers	106	766	13.84%
Vendors	125	2,232	5.60%
Money launderers	181	1,902	9.52%
Facilitators	60	1,439	4.17%
Across Marketplaces (N = 1,325)			
Z estimate	1,325	79,860	1.66%
Buyers	529	32,829	1.61%
Vendors	417	12,455	3.35%
Money launderers	379	27,124	1.40%
Facilitators	227	9,750	2.33%

First examined are the estimates for the primary market actors, buyers and vendors. Estimates reveal that marketplaces 1 and 2 are denominated by buyers, who comprise nearly 60% of the marketplaces' populations. In comparison, vendors represent 27% and 22% of the marketplaces' populations, respectively. These findings dictate that buyers outnumber vendors more than 2 to 1 in marketplace 1 and nearly 3 to 1 in marketplace 2. Marketplace 3, in contrast, is heavily populated by vendors, who comprise 40% of the actors convening in that market. So much of the marketplace being comprised by actors on the supply side suggests two possibilities: 1) that this marketplace is grossly inefficient, as many vendors must compete for limited populations of prospective buyers or 2) that the estimates produced by the model resulted in downward bias in the population of buyers and/or upward bias in the population of vendors. Nonetheless, such skewed numbers of vendors would surely effect the economics of this market, in which competition for customers is either fierce or there is possibly a substantial number of scammers lurking in the market. Across marketplaces, buyers are estimated to comprise nearly 41% of all market actors active in 2015, compared to vendors who comprise nearly 16% of the market. These figures indicate that across marketplaces, there are nearly 3 buyers for every 1 vendor. Findings suggest that the demand side of the market does in fact exceed the supply side, which provides further evidence for the argument that stolen data markets are seller's markets (Holt et al., 2015, 2016). Also of interest was the overlap between buyers and vendors, or market actors who may potentially pose as resellers—actors who buy volumes of data, only to repackage their purchases for bulk resale at steeper prices. The overlap between buyers and vendors (n = 37) in the data was less than 3%, and less than 1% of actors that remain hidden are estimated to occupy this dual role in the larger marketplace.

Next, estimates are examined for the peripheral actors. Money launderers comprise 19% and 34% of the estimated population in marketplaces 1 and 3, in contrast to facilitators, who represent 17% and 26% of the populations frequenting these markets. Whereas the estimated numbers of money launderers only marginally outnumber facilitators in marketplaces 1 and 3, the estimated number of facilitators greatly outnumber money launderers in marketplace 2; however, the estimates spanning across all marketplaces reveal that a more substantial proportion of actors frequenting stolen data markets are involved in money laundering activities. Across all markets, money launderers

are estimated to outnumber facilitators by nearly 3 to 1 and comprise nearly 34% of the total estimate. Meanwhile, facilitators are estimated to forge 12% of this total population and although facilitators comprise the smallest subpopulation, the financial services that they provide are nonetheless essential for many actors to achieve the profit-motive and avoid detection (Morselli & Giguère, 2006). The higher estimates for money launderers make intuitive sense for a number of reasons. First, money launderers are likely to overlap with buyers, as these offenders must first acquire funds before they can be squandered. Assuming that many, if not most, of these individuals do not acquire the funds directly through data theft, their only other means of acquiring stolen financial data is through contact with vendors, although overlap in these roles occurred in just 0.22% of the aggregate data.[21] Combining the estimated figures of buyers and money launderers, the demand side of stolen data markets comprises 75% of the estimated market population.

Related to this larger inquiry is the proportion of the subpopulations who were actually captured. These proportions were derived by dividing the observed count of the subpopulation by the estimated count of the subpopulation and then multiplied by 100 to arrive at the percentage of each subpopulation that was captured during the 12-month observation period. Capture percentages are shown in the far-right column of Table 4.7. Across the different samples, results illustrate that just a relatively small segment of each subpopulation was visible. In marketplace 1, between 9% and 10% of all actors active in the market were visible, although recall that these estimates are provided by Zelterman's estimator, assumes a fixed capture probability for each offender in the sample. The covariate adjusted model corrected for the assumption of fixed capture probability and results show that capture percentage varied greatly across subpopulations. In the data gathered from marketplace 2, only 6% of buyers were captured, compared to 40% of money launderers. For marketplace 3, the derived capture percentages indicate just 4% of facilitators were captured, compared to nearly 14% of buyers. Across marketplaces, the capture percentage ranged from 1.40% for money launderers to over 3.35% for vendors; therefore, results indicate that only a small percentage of the market actors, who were actually active during 2015, were observed through data analysis.

[21] I do not believe that my reasoning is necessarily invalid, but rather that access to more complete data is required where the source of laundered funds is known.

4.5. Assessing the Capture-Recapture Estimation Models

Even though the estimates are derived by well-fitted models, there is no certainty as to whether they are actually accurate. The potential violation of capture-recapture assumptions, as well as uncertainty as to whether the sample size is reflective of the larger population, makes assessing the plausibility of estimates an important component of analysis (Bouchard, 2007a). But unlike previous work on illegal drug users and dealers (Bouchard & Tremblay, 2005), forced labour and human trafficking (van der Heijden et al., 2015), marijuana production and cultivation (Bouchard, 2007a, 2008), and sex offenders (Bouchard & Lussier, 2015), there are no satisfactory baseline estimates of the populations of online offenders involved in market crimes from which to compare the estimates produced in this study. Given these circumstances, one means to assess estimates is to compare the proportions of actors estimated to be active among the larger recorded cumulative population for each marketplace—that is, the total number of users who registered to the online market over the duration of its life course—and compare these figures to previous studies that have used similar data samples for analysis. The data used by Motoyama et al. (2011) and Yip et al. (2012) are especially relevant for such an assessment, as these researchers had access to datasets consisting of user-generated data from both public threads and private exchanges. In essence, this meant that these researchers were privy to all data that was generated by all offenders participating in the online marketplaces comprising their samples. With access to such rich data, deducing the total number of actors who frequented a single marketplace was a rather simple task.

But for proper comparisons to be made with the capture-recapture estimates produced in this study, the data from these studies had to first be manipulated to arrive at annual estimates of the number of actors, who were active in the marketplaces. To arrive at figures reflecting only actors with recorded activity, the number of actors comprising the samples were first multiplied by the percentage of actors who had registered to the marketplaces, but remained completely dormant. These figures were then divided by the number of months spanning each of the samples to arrive at a monthly average, which was multiplied by 12 to get an annual estimate of active users for each of the samples. These estimated figures are crude, but nevertheless provided annual proportions of actors

who were active among the larger recorded cumulative population from which the estimates produced by Zelterman's models could be compared.

In Yip et al.'s (2012) sample of 4 carding forums, the proportion of actors that were active in a calendar year ranged from 12.21% to 26.95%. For Motoyama et al.'s (2011) sample involving 6 such websites, the proportion of offenders active in the online marketplace over a 1-year period approached upwards to 30.77%, although, the two largest websites with regards to known registration (n = 38,377 and n = 33,986) are perhaps most reflective of the websites sampled for the current study.[22] For these two marketplaces specifically, an estimated 5.78% and 7.50% of the total number of actors registered to these marketplaces were actually active during a year's span. These figures, along with the lower bound estimate derived from Yip et al.'s data, are reflective of the figures shown in Table 4.8, which indicate that between just 5% and 12% of the recorded cumulative population was active at least some point throughout 2015. Similarity, these figures suggests that these estimates are indeed plausible.

It was next necessary to assess estimates produced using the cross-sample capture-recapture distribution, which represents the larger population active across all markets, but not observed in the sample used in this study. It was first pertinent to examine whether actor overlap across multiple marketplaces was similar to figures found in previous research to assess the applicability of the cross-sample capture-recapture distribution to produce population estimates for the marketplace. Motoyama et al. (2011) produced a matrix that compared actor overlap between any 2 of the 6 markets included in their sample. Results indicated that between 0.33% and 21.47% of actor usernames overlapped across markets. As this is quite a large range, it is useful to consider the effect of language on the overlap between markets. Under these parameters, overlap was much larger across German-speaking markets (mean = 11.96%) than across English-speaking

[22] These comparison samples are limited in their generalizability, though they do serve as a rough baseline from which to extrapolate figures. Variations in these numbers are to be expected given differences across studies with regards to volumes of collected data and the procedures used to sample data.

markets (mean = 3.80%).[23] Zhao et al. (2016) also analyzed actor overlap across pairs of 12 marketplaces and found that overlap was not common, involving less than 5% of all actors. For this study, the cross-sample capture distribution indicates that 2% of actors frequented multiple marketplaces, which is similar to the lower bound of the two estimates derived from Motoyama et al.'s data and Zhao et al.'s findings. Results show that the cross-sample distribution used to produce the estimate across any market is consistent with other findings that suggest overlap is rare.

Table 4.8 Prevalence of market actors active during 2015 and cumulative populations (as indexed by website statistics)

Sample	*Z*	Cumulative Population	% Active
Marketplace 1	4,471	89,088	5.02%
Marketplace 2	6,473	81,793	7.91%
Marketplace 3	5,875	50,386	11.66%
Across Marketplaces	79,860	1,316,251[a]	6.07%

[a] Cumulative population of the known number of registrants across the 38 forums indexed during data collection

The final assessment involved determining the cumulative populations of online marketplaces spanning the Internet. The initial survey prior to data collection found 38 such marketplaces, 32 of which disclosed the population for the market. Across these marketplaces, the cumulative population totalled 1,316,251 market actors. The null model estimate of 35,090 indicates that 1.55% of this cumulative population was active during 2015, whereas the partial estimate of 79,860 (CI = 28,276–131,443) suggests 6.07% of this same population was active during this time period, with a plausible range of 2.15% to 9.99%. In addition to being the better fitting model, given the size of the cumulative population spanning known marketplaces, its likely that the estimate provided by the

[23] Perhaps the reason for smaller overlap across websites catering to English-speaking fraudsters is that they are less insular than the German-speaking websites (Holt et al., 2016) and cater to larger more diverse populations, as English is a common spoken language between fraudsters operating online.

partial model is the more realistic figure. Although its hard to deduce a satisfactory answer, as it is not actually known how many websites comprise the marketplace for stolen data.

The closest comparisons can once again be made to Yip et al.'s (2012) and Motoyama et al.'s (2011) studies, in which at least 5.78% and 12.21% and upwards to 25.95% and 30.77% of all actors populating the marketplaces in their samples, respectively, were concluded to be active over a 12-month span. The estimate produced by the partial model is at the conservative end of this range; however, the comparison is somewhat inequitable due to differences in the scope of analysis, as well as the composition of the data between these studies and the current study. The estimated proportion derived from the cross-sample capture-recapture distribution (6.07%) fit between the ranges derived from the samples used in analysis—the 5% – 12% of the cumulative population frequenting all such marketplaces during 2015. Thus, the extrapolated figure may in fact be plausible.

Chapter 5.

Discussion

This study sought to address two simple research questions: 1) how many actors frequent stolen data markets in one calendar year? and 2) what is the composition of subpopulations comprising the markets? To address these research questions, Zelterman's estimator and its extended covariate adjusted model, were fitted to the capture distributions generated from each of the data samples. Opposed to other count models or capture-recapture methods used to estimate the size of hidden populations, Zelterman's models were preferred for a number of reasons. First, Zelterman's models correct for issues stemming from overdispersion that are associated with other adjusted count models like truncated Poisson regression models that cause downward bias in population estimates (Böhning & van der Heijden, 2009). Count models are also subject to a number of assumptions, which are often violated by data collected from criminal populations. Count models assume a closed population, homogeneity among the population, and that the probability of capture is constant amongst the population. These assumptions must be respected when producing estimates with many, if not most, capture-recapture models, although the increased flexibility provided by Zelterman's models relaxes these assumptions. For the reasons described, Zelterman's models have proven to be robust in efforts to estimate the size of criminal populations. Previous applications of these methods for the estimation of populations comprising illegal drug markets (Bouchard, 2007a, 2008; Bouchard & Tremblay, 2005; Bouchard et al., 2012) gave further assurance that they provided the right framework to address the research questions.

Analysis consisted of a series of stepwise procedures to ensure accuracy in the produced estimates. Separate analyses were undertaken for each market, after which the 3 samples were consolidated into a single dataset to produce an estimate of the number of actors, who were active in any marketplace. Zelterman's estimator first provided baseline estimates from which more complex models could be compared. Results indicated that the observed population of each marketplace was only a fraction of their true size; however, the estimator model was not informative other than providing figures

of sheer market size, given that it lacks the capacity to control for observed heterogeneity within the data. Estimates were improved through better model specification by extending Zelterman's estimator into its covariate adjusted model, referred to as Zelterman's regression, to model observed heterogeneity through the addition of covariates, which also added substantive context to the population estimates (Böhning & van der Heijden, 2009; van der Heijden et al., 2013). Next in the series of procedures involved a thorough examination of the parameters produced using Zelterman's regression model. Closer inspection of the coefficients indicated why some models produced better estimates than others by informing which of the added covariates had a significant and meaningful impact on probability of capture. The final component of analysis involved deriving estimates for each subpopulation, along with the percentage of actors who were actually observed through data analysis. The results for each subpopulation provided additional context regarding the composition of each marketplace and how they differed, the composition across marketplaces, and which market actors were the most visible. Interpretation of findings, what they mean with regards to the dynamics of stolen data markets, and how they fit within the current literature are discussed in detail throughout the following sections.

5.1. Size Matters

Commonly perpetuated within the relevant literature is the notion that online illicit marketplaces are growing in size and complexity (Ablon et al., 2014). Prior studies interested in the dynamics and social organization of stolen data markets (Afroz et al., 2013; Décary-Hétu & Laferrière, 2015; Décary-Hétu & Leppänen, 2013; Holt, 2013a, 2013b; Holt & Lampke, 2010; Holt et al., 2013, 2015, 2016; Hutchings & Holt, 2015; Motoyama et al., 2011; Soudijn & Zegers, 2012; Yip et al., 2012, 2013) have provided empirical evidence highlighting the complexities of these marketplaces, including innate social processes that serve to regulate behaviour, maximize performance, and ensure efficiency; however, no previous studies have applied the necessary methods to estimate the size of stolen data markets and their scope cannot be surmised through description or simple counts of actors who are visible in the marketplace.

Motoyama et al. (2011) and Zhao et al. (2016) possessed the necessary data samples and had the basic concept for capture-recapture (e.g., measuring overlap between markets). Zhao and his colleagues, for instance, counted a gross total of 10,187 actors across data collected from 12 marketplaces, with less than 5% of actor overlap between all pairs generated from the 12 data samples;[24] however, they did not extend the measured overlap into such a framework to estimate the size of the population. This study took this next step, adding to the literature on offender prevalence and providing additional support for the application of capture-recapture methods for estimating the size of illegal markets (Bouchard, 2007a, 2008; Bouchard & Tremblay, 2005; Bouchard et al., 2012), and criminal populations more generally (Collins & Wilson, 1990; Böhning et al., 2004; Brecht & Wickens, 1993; Bouchard & Lussier, 2015; Greene, 1984; Greene & Stollmack, 1981; Hay et al., 2008; Hser, 1993; Leyland et al., 1993; Mastro et al., 1994; Rhodes, 1993; Riccio & Finkelstein, 1985; Roberts & Brewer, 2006; Rossmo & Routledge, 1990; van der Heijden et al., 2003, 2013, 2014, 2015; Weaver & Collins, 2007; Wickens, 1993).

Analysis began by scoping the populations with a baseline model to provide estimates from which more complex models could be compared. The baseline estimates produced by Zelterman's estimator model showed that a substantial number of actors (roughly 89% – 91% across samples) who were active during the capture period, were not observed and, thus, included in the samples for data analysis. Overall, findings suggest that the observable size of the markets subject to analysis represent between 5% – 12% of the true extent of offender prevalence—that is, just 5% – 12% of the cumulative population recorded through the forum's statistics were actually active during 2015. Using the distribution measuring overlap across samples, the size of the marketplace estimated to exceed 36,000 offenders. Model specification was improved using the covariate adjusted Zelterman model, and with it, the estimates of marketplaces 2 and 3 more than doubled in size, whereas nearly 80,000 offenders were estimated to comprise the larger marketplace.

[24] This figures suggest that without accounting for the known overlap, at the very most 509 actors in their sample were active in multiple markets, although generating an overlapping distribution across all 12 samples would represent a much smaller proportion of actors in their dataset.

Do these figures suggest that stolen data markets are large or small? Once again, as no previous work has sought to estimate the number of actors frequenting stolen data markets, there are no comparisons from which to judge the estimates produced in this study. That said, Zhao et al.'s samples suggest that the populations of stolen data markets are likely to number in the tens of thousands (n = 10,187) and given that the 80,000 figure represents a 'global' estimate, the estimates produced in this study seem plausible at the very least. Furthermore, given that this is the first known study to produce population estimates of illicit online marketplaces, Ablon et al.'s (2014) assertion that markets are growing in size is premature. Considering the increases in the prevalence of incidents and the direct financial losses resulting from online fraud, its tempting to speculate whether the offender population frequenting stolen data markets is also growing. Though the current population is almost certainly larger than the population frequenting early markets, such as ShadowCrew and DarkMarket (Glenny, 2011), more recent trends are difficult to surmise. For instance, the upwards trend in prevalence statistics may not be the direct result of more offenders, but may plausibly be attributed to the onset of more opportunities for offenders to commit online frauds (Brennan & Dauvergne, 2011).

While the statement is likely to be true, according to the principles of economics, illicit markets follow cyclical processes in which the number of actors on the demand side may at times exceed those on the supply side, and vice versa. Empirical assessment of trends in market activity requires that estimates be produced through dynamic modeling efforts that measure population stability over time (see Homer, 1993); however, as the prevalence estimates produced in this study reflect the population size of stolen data markets only for a specified unit of time (2015), results only portray a 'snapshot' of the marketplace's population (Anglin, Caulkins, & Hser, 1993). Though the estimates produced in this study do not take into account the dynamic changes in markets that impact population stability, this study nonetheless provides a framework for future research to improve upon and apply more complex methods to arrive at more reliable estimates of offender populations active online.

5.2. Differences in Capture Probability Across Markets

A thorough examination of the regression parameters for each model indicated that occupying one of the 4 specified market roles generally led to an increased probability of capture, though probabilities differed within and across markets. Such findings suggest that heterogeneity does indeed exist from one market to the next. Across markets, occupying any of the 4 subpopulations also had a significant and meaningful impact on the probability of capture, though vendors were the most visible. In a marketplace characterized by a highly skewed distribution of activity (Holt, 2013a, 2013b) and a high degree of population displacement (Hutchings & Holt, 2015), recall that Motoyama et al. (2011) found that the most trusted vendors were those who generated a substantial volume of market activity over long periods of time. From a criminal achievement perspective, the increased visibility of vendors makes perfect sense—to realize success in a competitive marketplace, vendors must increase their visibility to the other actors through continuous advertisement of the products and services they are providing to the market.

One of the more interesting findings was that actors seeking to add security to otherwise risky transactions, through escrow services and pursuing contacts in verified marketplaces, only had a significant probability of capture in marketplace 3. The extent to which actors in marketplace 3 engaged in such behaviour suggests a marketplace where security mechanisms that counter against dishonest behaviour are strongly enforced, which fosters trust within the marketplace and maximizes market performance (Holt et al., 2015). Although results also show that these characteristics are rare, as across marketplaces these same covariates were not found to have a significant impact on the probability of capture. Contrary to the findings of Holt and Lampke (2010), who concluded that the majority of buyers participating in stolen data markets limit their contact to only verified vendors to mitigate against risks of dealing with unscrupulous actors, the findings of this study indicate that the majority of actors frequenting stolen data markets are willing to incur risk; therefore, though risks and uncertainty may very well threaten the profit motives of actors (Herley & Florêncio, 2010), results nonetheless indicate that embarking on a criminal career in stolen data markets is an inherently risky venture.

5.3. Composition of Markets by Subpopulation

Results also indicated that markets differed in composition, reflecting subtle heterogeneity that exists across otherwise homogenous markets. These findings are interesting for a number of reasons. First, differences in composition in the subpopulations of offenders reflects the unique dynamics and social organization of each marketplace. For instance, marketplaces 1 and 2 were heavily populated by buyers, thus appearing to be geared toward the distribution and sale of stolen financial data, which suggests that opportunities are perhaps more lucrative for vendors in these two markets (Bulakh & Gupta, 2015; Holt et al., 2016). In contrast, marketplace 3 is heavily populated by money launderers and facilitators, potentially indicating that this market specializes in laundering funds that are used to fund subsequent frauds. Such variability in the composition of markets support calls for future research to analyze *active* markets with different surface characteristics to gain a more thorough understanding as to how and why markets may vary (Holt & Smirnova, 2014).

Across markets, findings were as expected—actors on the demand side of the market outnumbered those on the supply side. Though outnumbered, the supply side of the market was more visible, the demand side was estimated to comprise approximately 75% of the marketplace. Buyers comprised 41% of the global marketplace, whereas money launderers were estimated to comprise nearly 34% of the market. On the supply side, vendors and facilitators represented 16% and 12% of the market size, respectively. The estimated proportion of buyers to vendors (less than a 3:1 ratio) across marketplaces contradicts the findings of Zhao et al. (2016), who found that the number of known buyers (n = 2,121) were much fewer in number than were vendors (n = 3,165) across their sample, consisting of 12 illicit online marketplaces.[25] Is this ratio large enough to ensure that demand and/or consumption exceeds supply in stolen data markets? The likely answer is yes. Because contact formation between buyers and vendors, for instance, is highly skewed, the majority of vendors have few contacts, whereas few vendors have many contacts, who in turn have the largest market shares (Afroz et al., 2013; Décary-Hétu & Laferrière, 2015; Yip et al., 2012); therefore, the marketplace is highly lucrative only for a

[25] The discrepancy between the number of buyers and vendors is likely to be at least partly attributed to the classification algorithm developed by Zhao et al. to annotate their data.

very small number of vendors—similar to the concentration of wealth in legitimate markets (Vitali, Glattfelder, & Battiston, 2011). Another possible explanation for the relatively small buyer to vendor ratio is that stolen data markets are niche markets and, as research on illegal markets indicates, demand to supply ratios may in fact be smaller in niche markets. For instance, Bouchard and his colleagues (2012) found that the prevalence and corresponding ratio of Canadian meth users and dealers was much smaller than ratios for more widely used drugs like marijuana or even for other synthetic drugs like ecstasy.

5.4. Limitations

Four limitations hinder the estimates produced in this study and should be considered in the interpretation of results. The first limitation concerns the length of the capture periods implemented in the research design. Capture periods were implemented on a monthly basis to counter against the transient nature of actors frequenting stolen data markets (Décary-Hétu, & Laferrière, 2015; Holt, 2013b; Hutchings & Holt, 2015) and, thus, the violation of the closed population assumption; however, the potential drawback of choosing shorter capture periods is that overlap across capture periods becomes smaller, especially with elapsed time, which can have an adverse effect on capture probability and inflate the confidence intervals of the population estimates (van der Heijden et al., 2013). Although the confidence intervals produced along with estimates were relatively tight, even for the estimates produced using the cross-sample distribution, indicating that the estimates are stable. The robustness of the models further suggests that any adverse effects inflicted on the estimates through the short capture periods are negligible compared to what otherwise would be the case if longer capture periods were implemented as part of the research design.

Second, though analysis was undertaken to assess the plausibility of the estimates, the lack of a standard from which to compare the findings makes it difficult to determine whether the models are biased (Anglin et al., 1993). In other words, there is no way of knowing with any certainty whether the estimates produced by Zelterman's models are actually on the mark. The closest comparisons were made to previous studies that benefitted from having access to all activity archived within forums, including open threads and private dialogue between actors (Motoyama et al., 2011; Yip et al., 2012). But the

comparisons between the estimates produced in this study and the data samples used by Motoyama et al. and Yip et al. are somewhat inequitable due to differences in composition across data samples and differences in the scope and sophistication of analysis. Due to the paucity of research on offender prevalence, this issue is not uncommon to other studies that have sought to estimate the number of offenders in a given population (Bouchard & Lussier, 2015). Despite the uncertainty surrounding the plausibility of the estimates, Zelterman's methods have proved robust in a number of contexts, including estimating the size of other illicit markets (Bouchcard, 2007a, 2008; Bouchard & Tremblay, 2005; Bouchard et al., 2012), which lends assurance to the estimates produced in this study.

Third, estimates are hindered by the fact that it was not possible to produce estimates for two key subpopulations: suppliers and rippers. Though it was originally intended for population estimates to be derived for both of these subpopulations, during coding, it became clear that inferences could not be made as to who these actors were in the data. With regards to suppliers, there was no guarantee that actors seeking to purchase malware or vulnerabilities to facilitate data theft were in turn going to supply the market, or whether they even purchased and deployed the malware at all.[26] If the number of suppliers were indeed able to be identified, it's at least interesting to ponder whether the proportion of suppliers to vendors in stolen data markets would be similar to that in illegal drug markets, in which there are many suppliers (Bouchard et al., 2012), or if the activity is monopolized by relatively few actors. Likewise, it was not possible to denote rippers, because with their removal from the market by administrators, comes their removal from the forum's roster and the removal of their user-generated content from the website. Even if they could be identified, it is impossible to deduce actors who are banned for flagrant fraudulent activity, but continually re-enter markets under a different pseudonym (this issue is discussed in more detail below). Nonetheless, counts of rippers would likely vary from market to market. For instance, in lemon markets, the number of estimated rippers would likely comprise a more significant proportion of all actors convening in those marketplaces (Herley & Florêncio, 2010; Holt et al., 2013; Yip et al., 2013), compared to other 'legitimate' markets. In this study, marketplace 1 would have the

[26] This is a good illustration of one of the limitations of forum data more generally: often archived in the data is *intent* to commit illicit activity, not record of *actual* crime.

greatest potential to assume the label of a lemon market, given the trivial number of actors that leveraged the increased security provided by escrow services (n = 11) or established contact with vendors in verified markets (n = 6).

Fourth, it was impossible to control for actors employing multiple pseudonyms. The inability to control for actors with multiple aliases introduces a certain, but unknown degree of contamination into the samples, which effects the quality of the population estimates (Böhning, 2010). But contamination imposed by an unknown number of individuals using multiple aliases is not an uncommon issue for data collected on criminal networks. Sparrow (1991) suggested a method for determining actors employing multiple aliases, though the necessary methods needed to make such determinations is outside the scope of the current study. Furthermore, Sparrow's strategy may not even be applicable in online networks, which remove barriers traditionally limiting offender mobility and experience mass population displacement (Hutchings & Holt, 2015). Another strategy worthy of note is the use of linguistic analysis to identify doppelgängers, although previous research has found that large depositories of user-generated data is needed to make proper inferences as to whether any two pseudonyms share the same identity (Afroz, Caliskan-Islam, Stolerman, Greenstadt, & McCoy, 2014). Due to the large volumes of data needed to undertake analysis, such a method is not likely to be useful in an online marketplace, as very few users would able to be classified. Furthermore, these techniques have methodological limitations of their own and due to these limitations, the confidence that one can have in such applications is suspect at best. If there were the ability to control for the number of actors using multiple pseudonyms, the estimates produced by Zelterman's models would surely be biased downward, but to what degree remains unclear.

5.5. Future Research

For new measurement strategies to properly scope the extent of online fraud, the many conceptual issues related to data collection and ensuing methodology must be addressed. Four directions for future research that address these issues are discussed in sequential order: 1) improving the quality of data collection; 2) the extension of capture-recapture methods; 3) the application of survival analysis methodologies to determine the

parameters that sustain survival in the criminal careers of online offenders; and 4) developing a framework for measuring consumption in stolen data markets.

5.5.1. Improving Quality of Data Collection

Estimates of offenders committing online frauds are even harder to deduce than are estimates regarding the prevalence of offending, as victims and police are often left in the dark about the identity and whereabouts of the offender(s)—omissions that are reflected in official statistics (Mazowita & Vézina, 2009; Tcherni et al., 2016). As evidenced in this study, data sourced from websites facilitating online fraud involving the sale, purchase, and use of stolen financial data is one means to access these hidden populations of offenders. But these websites provide just one such data source. Holt, Strumsky, Smirnova, and Kilger (2014) advocate a triangulation method to collect and synthesize different types of online data (e.g., blogs, chat rooms, forums, social media websites, etc.) that are frequented by online offenders. Such an approach to data collection provides a unique opportunity to saturate the recorded activity of online offenders and access diverse, yet complementary datasets that are otherwise usually unavailable to researchers.[27] This triangulation method is particularly suited to capture-recapture methods like the multiple-multiplier method, which requires multiple, complementary datasets to derive estimates on a single population. Furthermore, through a systematic approach to data collection, such as the use of custom software, diverse types of data can be manipulated and consolidated into a single dataset, so that models that are not as restrictive in their assumptions, such as Zelterman's, can easily be fitted to the data.

As little is empirically know about population stability in illicit networks, future research should also consider data samples spanning multiple (3 – 5) years, for the analysis of population trends. By modeling observed growth, cyclical patterns, and/or recession, researchers can begin to hypothesize why such trends occur in the marketplace

[27] Online data can also be further supplemented with conventional curated data, such as official statistics, as advocated by Williams et al. (2016), who consolidated statistics on crime occurring in the boroughs of London, England with corresponding geo coordinated Twitter data generated by users 'tweeting' about local incidents. The augmented data showed great promise for the increased precision in predictive modeling of crime and social disorder.

(Tremblay et al., 2009), which will ultimately lead to a more thorough understanding of the degree to which illicit network's experience population growth, stability, or displacement from one time period to the next.

5.5.2. Extension of Capture-Recapture Methods

Because all models are subject to error in model specification and estimates can differ simply by using one model over another, there is benefit to using multiple estimation models (Hser, 1993; Wickens, 1993). For instance, estimates produced through Chao's estimator (1989) can be directly compared to those produced by Zelterman's model to further assess the plausibility of the derived estimates. Though the two estimators are similar in many ways, Chao's model is generally argued to produce more conservative estimates and for this reason, is referred to as the lower bound estimator (see Böhning, 2010).[28] Chao's estimator was specifically developed to be used with scarce data sources in which frequency counts of capture are small—a common outcome in open populations. Unlike Zelterman's model that removes heterogeneity, Chao's estimator assumes the presence of heterogeneity, which is controlled for through the calculation of smaller estimates, especially if the sample size is large. Böhning, Lerdsuwansri, Vidal-Diez, Viwatwongkasem, and Arnold (2013) have also recently extended Chao's model into a zero-inflated Poisson regression framework. As is the case with Zelterman's regression, the covariate adjusted Chao model increases precision in the derived estimates by controlling for observed heterogeneity within data samples through added covariates.

In addition to considering alternative capture-recapture methods to estimate offender populations, future research should also consider measuring alternative definitions of prevalence that provide insight into the economics of stolen data markets, including illicit revenues generated by actors and the monetary value attached to commodities exchanged through the marketplace (Anglin et al., 1993; Tcherni et al., 2016). Such a framework has the potential to build upon the work of Holt et al. (2016) and Bulakh and Gupta (2015), by producing more precise estimates of revenues that can be

[28] Zelterman's estimator is not always larger than Chao's (Böhning, 2010; Böhning & van der Heijden, 2009). Bohning (2010) found that when few capture periods (e.g., 2) are implemented in the research design, population estimates produced by Zelterman's estimator are more conservative than those produced by Chao's.

accrued by actors on both the supply and demand side of the marketplace. But given the issues posed by variation in the breadth of description between vendor advertisements and, thus, the extant 'missingness' that ensues in the data, alternative data sources directly related to the marketplace, such as individual vendor 'shops', websites that catalogue the products, goods, and services, that vendors have in stock, may be better suited to derive such estimates.

Using this data source, capture-recapture models could also control for additional covariates that were not able to be controlled for in the estimation models produced in the current study, including the type of advertised data and the geographic origin of the data. Controlling for these factors has the potential to improve estimates specifically and crime measurement more broadly, in two key ways. First, specific type(s) of online fraud can be better differentiated and clearly distinguished, such as the prevalence of credit and debit card fraud. Second, crime and/or victimizations rates can be supplemented by deriving estimates of victimization from the volumes and geographic origins of advertised data in online markets. These measurements would have the added value of enabling direct comparison to the prevalence figures disclosed through official crime statistics, victimization surveys, and organizations such as the Canadian Banker's Association and the Canadian Anti-Fraud Centre.

5.5.3. Determining Survival in Online Criminal Careers

Structured data collected from stolen data markets also provides meaningful samples from which to examine the activity of market offenders from a criminal career perspective, as the archived nature of the data is necessary to analyze longitudinal patterns of offending (Blumstein & Cohen, 1987; Blumstein, Cohen, & Farrington, 1988; Piquero & Benson, 2004). Opposed to the macro level approach of the current study, the criminal career perspective is pertinent to understanding prevalence of offending among individual offenders, as well as their patterns of co-offending (Farrington, 1992). Gaining a more thorough understanding of criminal achievement calls for a framework that considers how offenders connect with one another and their subsequent opportunity structures for crime. Opportunities structures for crime granted through social capital are focal to criminal career achievement, yet few researchers have conceptualized

opportunities for offending (Piquero & Benson, 2004). Criminological theories assume that offender access to opportunities are more or less ubiquitous, and do not account for the prerequisite human and social capital necessary to leverage opportunities, which vary by individual and across crime types (Birkbeck & LaFree, 1993; Loughran et al., 2013).

As part of a larger research project, a forthcoming study will apply survival analysis methodologies to measure the criminal career trajectories of market actors using the data sample(s) used in this study. Social network analysis provides the necessary framework needed to measure the social capital that market actors accumulate through their ego networks to determine how these local level network structures impact the quality of criminal opportunities available through stolen data markets. Kaplan-Meier models will first provide the 'survival times' of actors, after which Cox regression models will consider different measures of social capital and their impact on 'survival' in online criminal careers.

5.5.4. Determining Consumption in Stolen Data Markets

Future research should also seek to develop a framework for measuring consumption in stolen data markets. Empirical research has shown that economic frameworks are indeed applicable to an online setting, as research by Nagurney (2015) found that the economics of stolen data markets are similar to those impacting the global drug market. As with the costs associated with global shipping and distribution that affect drug prices in domestic markets, Nagurney's findings indicate that the associated costs of data theft and its distribution through the supply chain in stolen data markets impacts the advertised prices set by vendors. Not only would the development of such a framework permit the application of more stringent economic theory and modeling to online illicit marketplaces, but economic models also complement prevalence estimates. As evidenced by Bouchard and his colleagues (2012), modeling the consumption rates of synthetic drugs provided an alternative measure of the size of the synthetic drug market in Canada, due to the likelihood that the number of users and dealers follow trends in consumption. Because the structured data posted online to vendor shops provides the appropriate longitudinal data sources denoting price of advertised commodities and corresponding auxiliary data, econometric modeling methods can be applied to estimate the scope of activity generated through stolen data markets.

Chapter 6.

Conclusion and Implications

Online property crimes involving the theft, sale, purchase, and fraudulent use of financial data are now considered more prevalent than traditional property crimes, such as burglary and theft (Anderson et al., 2013; Tcherni et al., 2016); however, since official statistics of online frauds began being collected, estimates derived from these statistics have been criticized due to the inconsistency in yearly reporting stemming from differences in definitions and recurrent modifications to data collection instruments. Such inconsistencies make official data sources uninformative, as yearly trends can neither be attributed to actual changes in crime trends or measurement error with any degree of confidence (Williams & Levi, 2012).[29] The current lack of reliable data has led to speculation in the perceived prevalence of online fraud across the academic, government, and financial sectors, and these perceived differences have undoubtedly affected the development of policy and proper unified responses.

In the absence of a crime prevention strategy, law enforcement interventions have taken a traditional approach to disrupting the stolen data market through the targeted removal of key players and/or the seizure of online domains (see Glenny, 2011). Though as part of a report produced by the Science and Technology Committee of the UK House of Lords (2007), Anderson argued in favour of enforcement strategies consisting of randomized attacks. By placing the least and less serious offenders at equal risk to be detected and targeted as the most serious offenders, the argument is that random attacks would cause a greater deterrence effect than what would otherwise be the case through just the removal of the most serious offenders. Although the small world characteristics of stolen data markets (Yip et al., 2012) suggests that a combination of targeted *and* random

[29] But as police data collection for cybercrimes becomes more comprehensive, crime reporting will improve and the proportion of frauds that occur online will be more accurately portrayed by official statistics, resulting in more adequate data for the measurement of online fraud (Williams & Levi, 2012; Tcherni et al., 2016).

attacks is the best strategy for their disruption (Kalm, 2013).[30] Such an intervention strategy incorporates both Anderson's argument advocating for randomized enforcement, while also taking into account the substantive social network literature that advocates for the targeted removal of key players as the most effective means to disrupt illicit networks (Bichler & Malm, 2015; Everton, 2012; Morselli, 2009).

While law enforcement interventions have proven to be successful at infiltrating these marketplaces and seizing their online domains, the estimated scope of the marketplace also suggests that disruption strategies consisting solely of traditional law enforcement interventions are likely to be futile in the long term (Thomas et al., 2015). Not only are these intervention strategies time intensive and costly with regards to scarce law enforcement resources, but research by Keegan, Ahmed, Williams, Srivastava, and Contractor (2010) suggests that like illegal drug markets, online fraud networks are resilient to such attacks (see Bouchard, 2007b). Other problems with these approaches are that relatively few actors are targeted and apprehended, displacing the majority of offenders in the short term, only to re-up their activities in other markets. These approaches have also pushed online marketplaces increasingly 'underground' through the adoption of enhanced screening measures, including elaborate vetting processes, restricting the marketplace to only those participants who are either invited or vouched for by known actors, and/or outright displacement to anonymity networks (e.g., Tor, I2P, etc.). As a result, these hard to reach populations have become even harder to reach.

Opposed to the traditional disruption strategies undertaken by law enforcement, *Sybil attacks* have been advocated as an alternative method for the disruption of stolen data markets. Sybil attacks are intended to disrupt market performance and efficiency by raising uncertainty and perceived risk among the larger population of market actors through a series of slanderous accusations (Décary-Hétu & Laferrière, 2015; Franklin et

[30] Small world networks are decentralized, but contain dense 'clusters' or subgroups (Watts & Strogatz, 1998). Granovetter (1973) found that clusters of actors are more likely to be bridged together by weak ties (e.g., associates) than strong ties (e.g., friends). Not only do actors tend to form more weak ties, they are also crucial to individual and group performance (Granovetter, 1973; Watts & Strogatz, 1998), as the social capital embedded within social networks is more accessible to a greater number of actors through weak, rather than strong, ties (Everton, 2012). For instance, Yip et al. (2012) found that the small world properties of stolen data markets enable buyers and vendors to connect with one another more easily than in a randomly constructed network of the same size, yielding serious implications for the facilitation of fraud.

al., 2007; Hutchings & Holt, 2016; Yip et al., 2013). Sybil attacks are initiated by deploying a number of decoys into the market, who engage in a series of fictitious transactions with one another, which are openly or publically disclosed to the larger marketplace. Under the façade of these fictitious exchanges, actors accrue positive feedback in an attempt to establish positive reputations and trust among the other actors comprising the marketplace. If and when actual market actors engage with these fictitious actors to purchase products and services, payment is received by the fictitious vendors, although they do not complete their end of the transfer. As a result, the fictitious vendors are denounced as rippers, which invites speculation as to the reliability of the mechanisms that foster trust and certainty in the marketplace. But are Sybil attacks really an effective strategy for the disruption of stolen data markets? Research by Décary-Hétu and Laferrière (2015) suggests that there is no long term impact on the level of activity generated in stolen data markets following the identification and removal of fraudulent actors, and that the short term impact is much too trivial to suggest that Sybil attacks are an effective intervention strategy. Nonetheless, pre- and post-intervention research designs are needed to properly assess whether these strategies truly have any merit (Hutchings & Holt, 2016; Thomas et al., 2015).

Because current law enforcement strategies are largely reactive in nature and the offender population frequenting stolen data markets is large, such interventions are likely to be ineffective in the long term, if continued to be relied upon as the primary solution to curbing the upward trend in online fraud. For such reasons, many law enforcement agencies have been resigned to place the onus squarely on the public to protect themselves against victimization, which Levi and Williams (2013) argue stems from the disparity in cooperation between federally commissioned ‘cyber security strategies’ and local policing efforts to combat cybercrime. To bridge this gap, scholars have argued for the development and implementation of community policing initiatives that leverage the large population of Internet users, as an effective strategy for cybercrime prevention (for thorough discussions see Bossler & Holt, 2013; Brenner, 2007; Jones, 2007). The current problem with these approaches is that because so few online community policing models are in operation, there is no precedent or evidence for how best to administer these programs (Bossler & Holt, 2013; Jones, 2007); however, police organizations do currently use features of online community policing strategies. Police have embraced open source

intelligence, or information that is publically and freely available on the Internet, to assist with evidence collection and have adopted various forms of social media, including Facebook® and Twitter®, to connect with and request assistance from the public. Online reporting portals have also provided the public with additional and widely accessible means to report cybercrimes to the relevant authorities (Bossler & Holt, 2013). For instance, the Canadian Anti-Fraud Centre has recently developed a 'Fraud Reporting System' that can be accessed online to report suspicious activity and known incidents.[31]

But rather than attempt to police the Internet, a better means to reduce opportunities for cybercrime is through widespread adoption of adequate cyber-security countermeasures. Legislation that mandates the reporting of data breaches to individuals whose identities and credentials have been compromised, such as PIPEDA, has shown to be effective at reducing the financial impact of victimization resulting from fraud subsequent to data theft. By considering state level responses and victimization rates across a sample of American states that had mandatory notification and those that had no such laws, Romanoksy, Telang, and Acquisti (2011) examined whether mandatory data breach reporting had any effect on reducing incidents of identity theft. Their results suggested that data breach notification had the desired effect: identity theft was detected and reported more frequently in states that had adopted mandatory notification compared to those that did not adopt such legislation. Though the 6% decrease in incidents stemming from notification is arguably a relatively modest outcome, this study nonetheless suggests that cyber security policy is an effective step towards crime prevention and that similar studies are needed to form evidence based policy.

Though mandatory notification laws have improved institutional accountability and cyber security more generally, further standards must be developed to improve the manner in which financial and corresponding personal data is stored by organizations across the public and private sectors. Some have advocated for policymakers to develop adequate legislation or refine existing policy that mandates organizations that collect and store data to develop cyber security safeguards that adhere to a minimum set of standards. At the very least, minimal regulations should mandate that any organization

[31] http://www.antifraudcentre-centreantifraude.ca/reportincident-signalerincident/index-eng.htm

storing data must implement procedures for data encryption (Hutchings & Holt, 2016). Furthermore, the sheer size of the stolen data market dictates that policy that seeks to enforce crime controls through minimum sets of standards for cyber-security and data protection must ensure that standards raise security to such a degree that it has a substantial impact on the economics shaping the marketplace (Jones, 2007), particularly the supply and durability of products, pricing structures, and gross revenues (Thomas et al., 2015). As Nagurney (2015) argues, to be effective, such security measures must considerably increase the acquisition cost of data, which in turn increases transaction costs and advertised prices, thereby reducing demand for illegally acquired financial data by 'pricing out' a significant proportion of prospective buyers. Unfortunately, barring changes to public policy, such lofty standards are unlikely to be initiated from within organizations, as they do not have any current economic incentive to develop and implement such standards.

Given the size of the money launderer and facilitator populations, which together encompass roughly 46% of the stolen data market, further development of anti-money laundering policies are particularly relevant for crime control. Though the potential for the Internet to facilitate the money laundering process has long been, and continues to be, recognized (Financial Action Task Force, 2010a, 2013; Financial Transactions and Analysis Centre of Canada, 2010, 2015; Solicitor General of Canada, 1998), loopholes continue to exist in policy that circumvent anti-money laundering efforts (Financial Action Task Force, 2004, 2010b, 2012).[32] Due to the trend toward reducing barriers associated with online transactions to further the growth of Internet commerce (Filipkowski, 2008), mandatory due diligence practices, such as 'Know Your Customer' policies, and suspicious activity reporting requirements continue to be weakly enforced through online transactions. For example, money transfer services, such as MoneyGram® and Western Union®, only implement anti-money laundering procedures for transactions exceeding CAD$1000 (MoneyGram, 2008; Western Union, 2012). But even these measures are

[32] International standards for anti-money laundering policies are developed by the Financial Action Task Force. In Canada, the *Proceeds of Crime (Money Laundering) and Terrorist Financing Act* defines the statutes regulating financial activity and anti-money laundering efforts. Countries that do not comply with minimum standards, which are often jurisdictions referred to as 'tax heavens', are 'blacklisted' by the G7, which has serious implications for their economies due to the combined political and economic power of the G7 member countries.

weakly enforced, as the sender is only required to produce one valid form of customer identification (e.g., driver's licences); therefore, anti-money laundering controls must be considerably strengthened and consistently implemented to reduce opportunities for laundering money online (Hutchings & Holt, 2016).

But yet again, lost in the discussion is the underlying size of offender populations. The aim of this study was to estimate the size of stolen data markets, in addition to the number of offenders estimated to have frequented any online marketplace during 2015. Estimating the size of offender populations are pertinent to developing adequate policy and for informing the scope of interventions from law enforcement (Anglin et al., 1993). Previous studies have advocated for real-time, active monitoring of the populations frequenting these environments as a means of crime control (Décary-Hétu & Aldridge, 2015). Though at nearly 80,000 offenders, the estimated size of the global marketplace suggests that offender populations are much larger than what can be observed solely through monitoring open market activity. Rather, researchers should consider the size of the offender populations participating within and across markets as a key metric in pre- and post-intervention designs as a means to consider and/or compare the effectiveness of policies and intervention strategies.

References

Abbasi, A., Li, W., Benjamin, V. A., Hu, S., & Chen, H. (2014, September). Descriptive Analytics: Examining Expert Hackers in Web Forums. In *Joint Intelligence and Security Informatics Conference* (*JISIC)* (pp. 56-63). IEEE.

Ablon, L., Libicki, M. C., & Golay, A. A. (2014). *Markets for Cybercrime Tools and Stolen Data: Hackers' Bazaar*. Rand Corporation. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_reports/RR600/RR610/RAND_RR610.pdf

Afroz, S., Garg, V., McCoy, D., & Greenstadt, R. (2013, September). Honor among thieves: A common's analysis of cybercrime economies. In *eCrime Researchers Summit (eCRS), 2013* (pp. 1-11). IEEE.

Afroz, S., Islam, A. C., Stolerman, A., Greenstadt, R., & McCoy, D. (2014, May). Doppelgänger finder: Taking stylometry to the underground. In *2014 IEEE Symposium on Security and Privacy* (pp. 212-226). IEEE.

Aldridge, J., & Decary-Hetu, D. (2014). *Not an "eBay for Drugs": The Cryptomarket "Silk Road" as a Paradigm Shifting Criminal Innovation* (SSRN Scholarly Paper No. ID 2436643). Rochester, NY: Social Science Research Network.

Allen, M. (2016). Police-reported crime statistics in Canada, 2015. *Juristat*, *36*(1), 1-55. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2016001/article/14642-eng.pdf

Anderson, R., Barton, C., Böhme, R., Clayton, R., Van Eeten, M. J., Levi, M., & Savage, S. (2013). Measuring the cost of cybercrime. In R. Böhme (Ed.), *The Economics of Information Security and Privacy* (pp. 265-300). Springer Berlin Heidelberg.

Andresen, M. A., & Felson, M. (2012). Co-Offending and the Diversification of Crime Types. *International Journal of Offender Therapy and Comparative Criminology*, *56*(5), 811-829.

Anglin, M. D., Caulkins, J. P., & Hser, Y. (1993). Prevalence Estimation: Policy Needs, Current Status, and Future Potential. *Journal of Drug Issues*, *23*(2), 345-360.

Benjamin, V. A., & Chen, H. (2014, September). Time-to-Event Modeling for Predicting Hacker IRC Community Participant Trajectory. In *Joint Conference in Intelligence Security Informatics (JISIC)* (pp. 25-32). IEEE.

Bichler, G., & Malm, A. E. (Eds.). (2015). *Disrupting Criminal Networks: Network Analysis in Crime Prevention*. Boulder, CO: FirstForumPress.

Birkbeck, C., & LaFree, G. (1993). The situational analysis of crime and deviance. *Annual Review of Sociology*, *19*(3), 113-137.

Blumstein, A., Cohen, J., Roth, J., & Visher, C. A. (1986). Criminal careers and "career criminals": Report of the National Academy of Sciences Panel on Research on Criminal Careers.Blumstein, A., & Cohen, J. (1987). Characterizing criminal careers. *Science*, *237*(4818), 985-991.

Blumstein, A., & Cohen, J. (1987). Characterizing criminal careers. *Science*, *237*(4818), 985-991.

Blumstein, A., Cohen, J., & Farrington, D. P. (1988). Criminal career research: Its value for criminology. *Criminology*, *26*(1), 1-35.

Böhning, D., Suppawattanabodee, B., Kusolvisitkul, W., & Viwatwongkasem, C. (2004). Estimating the number of drug users in Bangkok 2001: A capture–recapture approach using repeated entries in one list. *European Journal of Epidemiology*, *19*(12), 1075-1083.

Böhning, D., & van der Heijden, P. G. M. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *The Annals of Applied Statistics*, *3*(2), 595-610.

Böhning, D. (2010). Some General Comparative Points on Chao's and Zelterman's Estimators of Population Size. *Scandinavian Journal of Statistics*, *37*(2), 221-236.

Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C., & Arnold, M. (2013). A generalization of Chao's estimator for covariate information. *Biometrics*, *69*(4), 1033-1042.

Bossler, A. M., & Holt, T. J. (2013). Assessing officer perceptions and support for online community policing. *Security Journal*, *26*(4), 349-366.

Bouchard, M., & Tremblay, P. (2005). Risks of arrest across drug markets: A capture-recapture analysis of "hidden" dealer and user populations. *Journal of Drug Issues*, *35*(4), 733-754.

Bouchard, M. (2007a). A capture-recapture model to estimate the size of criminal populations and the risks of detection in a marijuana cultivation industry. *Journal of Quantitative Criminology*, *23*(3), 221-241.

Bouchard, M. (2007b). On the resilience of illegal drug markets. *Global Crime*, *8*(4), 325-344.

Bouchard, M. (2008). Towards a realistic method to estimate cannabis production in industrialized countries. *Contemporary Drug Problems*, *35*(2-3), 291-320.

Bouchard, M., Morselli, C., Gallupe, O., Easton, S., Descormiers, K., Turcotte, M., & Boivin, R. (2012). *Estimating the Size of the Canadian Illicit Meth and MDMA Markets: A Multi-Method Approach*. Ottawa, ON: Pubic Safety Canada.

Bouchard, M., Joffres, K., & Frank, R. (2014). Preliminary analytical considerations in designing a terrorism and extremism online network extractor. In *Computational Models of Complex Systems* (pp. 171-184). Springer International Publishing.

Bouchard, M., & Lussier, P. (2015). Estimating the Size of the Sexual Aggressor Population. In A. Blokland and P. Lussier (Eds.), In A. Blokland & P. Lussier (Eds.), *Sex Offenders: A Criminal Career Approach* (pp. 351-371). Hoboken, NJ: Wiley.

Boyce, J., Cotter, A., & Perreault, S. (2014). Police-reported crime statistics in Canada, 2013. *Juristat*, *34*(1), 1-39. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2014001/article/14040-eng.pdf

Boyce, J. (2015). Police-reported crime statistics in Canada, 2014. *Juristat*, *35*(1), 1-40. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2015001/article/14211-eng.pdf

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345-370.

Brecht, M. L., & Wickens, T. D. (1993). Application of multiple-capture methods for estimating drug use prevalence. *Journal of Drug Issues*, *23*(2), 229-250.

Brennan, S., & Dauvergne, M. (2011). Police-reported crime statistics in Canada, 2010. *Juristat*, *31*(1), 1-39. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2011001/article/11523-eng.pdf

Brennan, S. (2012). Police-reported crime statistics in Canada, 2011. *Juristat*, 1-39. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2012001/article/11692-eng.pdf

Brenner, S. W. (2007). Cybercrime: re-thinking crime control strategies. In Y. Jewkes (Ed.), *Crime Online*. London, UK: Willian.

Broadhurst, R., Grabosky, P. Alazab, M., & Chon, S. (2014). Organizations and Cyber Crime: An Analysis of the Nature of Groups engaged in Cyber Crime. *International Journal of Cyber Criminology*, *8*(1), 1-20.

Bulakh, V., & Gupta, M. (2015). Characterizing Credit Card Black Markets on the Web. *Proceedings of the 24th International Conference on World Wide Web Conference* (pp. 1435-1440). ACM.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223-242.

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 1-15. doi: 10.1140/epjds/s13688-016-0072-6

Burt, R. S. (2000). The network structure of social capital. *Research in Organizational Behavior*, *22*, 345-423.

Canadian Anti-Fraud Centre. (2014). *Annual Statistical Report 2013: Mass Marketing Fraud and ID Theft Activities*. Canadian Anti-Fraud Centre Criminal Intelligence Analytical Unit. Retrieved from https://www.antifraudcentre-centreantifraude.ca/english/documents/Annual%202013%20CAFC/pdf

Canadian Banker's Association. (2015). *Credit Card Fraud and Interac Debit Card Fraud Statistics—Canadian Issued Cards*. Retrieved from http://www.cba.ca/contents/files/statistics/stat_creditcardfraud_en.pdf

Carrington, P. J., Brenna, S., Matarazzo, A., & Radulescu, M. (2013). Co-offending in Canada, 2011. *Juristat, 33*(1), 1-33. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2013001/article/11856-eng.pdf

Chao, A. (1989). Estimating Population Size for Sparse Data in Capture-Recapture Experiments. *Biometrics*, *45*(2), 427-438.

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.

Clarke, R. V., & Weisburd, D. (1990). On the distribution of deviance. In D. M. Gottfredson & R. V. Clarke (Eds.), *Policy and theory in criminal justice* (pp. 10-27). Avebury: Aldershot.

Collins, M. F., & Wilson, R. M. (1990). Automobile theft: Estimating the size of the criminal population. *Journal of Quantitative Criminology*, *6*(4), 395-409.

Darroch, J. N. (1958). The multiple-recapture census: I. Estimation of a closed population. *Biometrika*, *45*(3/4), 343-359.

Darroch, J. N. (1961). The two-sample capture-recapture census when tagging and sampling are stratified. *Biometrika*, *48*(3/4), 241-260.

Dauvergne, M. (2008). Crime Statistics in Canada, 2007. *Juristat*, *28*(7), 1-17. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/85-002-x2008007-eng.pdf

Dauvergne, M., & Turner, J. (2010). Police-reported crime statistics in Canada, 2009. *Juristat*, *30*(2), 1-37. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2010002/article/11292-eng.pdf

Décary-Hétu, D., & Dupont, B. (2012). The social network of hackers. *Global Crime*, *13*(3), 160-175.

Décary-Hétu, D., & Dupont, B. (2013). Reputation in a dark network of online criminals. *Global Crime*, *14*(2-3), 175-196.

Décary-Hétu, D., & Leppänen, A. (2013). Criminals and signals: An assessment of criminal performance in the carding underworld. *Security Journal*, *29*(3), 442-460. doi: 10.1057/sj.2013.39.

Décary-Hétu, D., & Aldridge, J. (2015). Shifting through the Net: Monitoring of Online Offenders by Researchers. *The European Review of Organised Crime*, *2*(2), 122-141.

Décary-Hétu, D., & Laferrière, D. (2015). Discrediting Vendors in Online Criminal Markets. In G. Bichler & A. E. Malm (Eds.), *Disrupting Criminal Networks: Network Analysis in Crime Prevention* (pp. 129-151). Boulder, CO: FirstForumPress.

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

Dupont, B. (2012). The criminal ecology of payment systems: How 'identity theft' evolved from plastic counterfeiting to the 'crime of the century'. S. Leman-Langlois (Ed.), Technocrime, Policing and Surveillance (pp. 13-27). New York., NY: Routledge.

Dupont, B., Côté, A. M., Savine, C., & Décary-Hétu, D. (2016). The ecology of trust among hackers. *Global Crime* (ahead of print), 1-23.

Eck, J. E. (1995). A general model of the geography of illicit retail marketplaces. *Crime and Place*, *4*, 67-93.

Eck, J. E., & Weisburd, D. (1995). *Crime place in crime theory*. In J. E. Eck. And D. Weisburd (Eds.), Crime and place, crime prevention studies (pp. 1-33). *Monsey, NY: Criminal Justice Press.*

Everton, S. F. (2012). Disrupting Dark Networks. *Structural Analysis in the Social Sciences* (Vol. 34). New York, NY: Cambridge University Press.

Farrington, D. P. (1992). Criminal career research in the United Kingdom. *British Journal of Criminology*, *32*(4), 521-536.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, *59*(3), 591-603.

Filipkowski, W. (2008). Cyber Laundering: An Analysis of Typology and Techniques. *International Journal of Criminal Justices Sciences*, *3*(1), 15-27.

Financial Action Task Force. (2004). *FATF Forty Recommendations*. Retrieved from http://www.fatf-gafi.org/media/fatf/documents/FATF%20Standards%20-%2040%20Recommendations%20rc.pdf

Financial Action Task Force. (2010a). *About the FATF*. Retrieved from http://www.fatf-gafi.org/about/

Financial Action Task Force. (2010b). *Money Laundering Using New Payment Methods*. Retrieved from http://www.fatf-gafi.org/media/fatf/documents/reports/ML%20using%20New%20Payment%20 Methods.pdf

Financial Action Task Force. (2012). *FATF Forty Recommendations*. Retrieved from http://www.fatf-gafi.org/media/fatf/documents/recommendations/pdfs/FATF_Recommendations.pdf

Financial Action Task Force. (2013). *Guidance for a Risk-Based Approach: Prepaid Cards, Mobile Payments and Internet-based Payment Services*. Retrieved from http://www.fatf-gafi.org/media/fatf/documents/recommendations/Guidance-RBA-NPPS.pdf

Financial Transactions and Reports Analysis Centre. (2010, July). *Money Laundering and Terrorist Financing (ML/TF) Typologies and Trends for Canadian Money Services Businesses (MSBs)*. Typologies and Trends Report. Retrieved from http://www.fintrac-canafe.gc.ca/publications/ typologies /2010-07-eng.pdf

Financial Transactions and Reports Analysis Centre. (2015). *Mass Marketing Fraud: Money Laundering Methods*. Typologies and Trends Report. Retrieved from http://www.fintrac.gc.ca/publications/ typologies /2015-02-eng.pdf

Florêncio, D., & Herley, C. (2013). Sex, lies and cyber-crime surveys. In B. Schneier, (Ed.), *Economics of Information Security and Privacy III* (pp. 35-53). Springer New York: New York, NY.

Frank, R., Westlake, B., & Bouchard, M. (2010). The Structure and Content of Online Child Exploitation Networks. *Proceedings of the 10th ACM SIGKDD Workshop on Intelligence and Security Informatics* (pp. 3-11). ACM.

Franklin, J., Perring, A., Paxson, V., & Savage, S. (2007). An inquiry into the nature and causes of the wealth of internet miscreants. *Proceedings of the Conference on Computer and Communications Security* (pp. 375-388). ACM.

Garson, D. G. (2013). *Generalized Linear Models/Generalized Estimating Equations*. Statistical Associates: Blue Book Series.

Glenny, M. (2011). *Darkmarket: Cyberthieves, cybercops and you*. New York, NY: Random House.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *73*(6), 1360-1380.

Greene, M. A., & Stollmack, S. (1981). Estimating the number of criminals. In J. A. Fox (Ed.), *Models of Quantitative Criminology* (pp. 1-24). New York, NY: Academic Press.

Greene, M. A. (1984). Estimating the size of a criminal population using an open population approach. In *Proceedings of American Statistical Association, Survey Research Methods Section* (pp. 8-13).

Hay, G., Gannon, M., MacDougall, J., Eastwood, C., Williams, K., & Millar, T. (2008). Capture–recapture and anchored prevalence estimation of injecting drug users in England: national and regional estimates. *Statistical Methods in Medical Research*, *18*(4), 323-339.

Herely, C., & Florêncio, D. (2010). Nobody sells gold for the price of silver: Dishonesty, uncertainty and the underground economy. In T. Moore, D. J. Pym, and C. Ioannidis (Eds.), *Economics of Information Security and Privacy* (pp. 33-53). New York, NY: Springer.

Holt, T. J. (2010). Exploring Strategies for Qualitative Criminological and Criminal Justice Inquiry Using On-Line Data. *Journal of Criminal Justice Education*, *21*(4), 466-487.

Holt, T. J., & Lampke, E. (2010). Exploring stolen data markets online: products and market forces. *Criminal Justice Studies*, *23*(1), 33-50.

Holt, T. J. (2011). Examining the Language of Carders. In B. H. Schell and T. J. Holt (Eds.), *Corporate Hacking and Technology-Driven Crime: Social Dynamics and Implications* (pp. 127-143). New York, NY: Hersey.

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*(1), 891-903.

Holt, T. J. (2013a). Examining the forces shaping cybercrime markets online. *Social Science Computer Review*, *31*(2), 165-177.

Holt, T. J. (2013b). Exploring the social organisation and structure of stolen data markets. *Global Crime 14*(2-3), 155-174.

Holt, T. J., Chua, Y. T., & Smirnova, O. (2013, September). An exploration of the factors affecting the advertised price for stolen data. In *eCrime Researchers Summit (eCRS), 2013* (pp. 1-10). IEEE.

Holt, T. J., & Smirnova, O. (2014). Examining the Structure, Organization, and Processes of the International Market for Stolen Data. *Washington DC: National Criminal Justice Reference Service*.

Holt, T. J., Smirnova, O., Strumsky, D., & Kilger, M. (2014). Advancing Research on Hackers Through Social Network Data. In C. D. Marcum and G. E. Higgins (Eds.), *Social Networking as a Criminal Enterprise* (pp. 145-167). Boca Raton, FL: CRC Press.

Holt, T. J. (2015). Qualitative criminology in online spaces. In H. Copes and J. M. Miller (Eds.), *The Routledge Handbook of Qualitative Criminology*. New York, NY: Routledge.

Holt, T. J., Smirnova, O., Chua, Y. T., & Copes, H. (2015). Examining the risk reduction strategies of actors in online criminal markets. *Global Crime*, *16*(2), 81-103. doi: 10.1080/17440572.2015.1013211

Holt, T. J., Smirnova, O., & Chua, Y. T. (2016). Exploring and Estimating the Revenues and Profits of Participants in Stolen Data Markets. *Deviant Behavior*, *37*(4), 353-367.

Homer, J. B. (1993). A system dynamics model for cocaine prevalence estimation and trend projection. *Journal of Drug Issues*, *23*(2), 251-279.

Housley, W., Procter, R., Edwards, A., Burnap, P., Williams, M., Sloan, L., & Greenhill, A. (2014). Big and broad social data and the sociological imagination: A collaborative response. *Big Data & Society*, *1*(2), 1-15.

Hutchings, A. (2014). Crime from the keyboard: organised cybercrime, co-offending, initiation and knowledge transmission. *Crime, Law and Social Change*, *62*(1), 1-20.

Hutchings, A., & Holt, T. J. (2015). A Crime Script Analysis of the Online Stolen Data Market. *British Journal of Criminology*, *55*(1), 596-614.

Hutchings, A., & Holt, T. J. (2016). The online stolen data market: disruption and intervention approaches. *Global Crime* (ahead of print). doi: 10.1080/17440572.2016.1197123

Hser, Y. I. (1993). Population estimation of illicit drug users in Los Angeles County. *Journal of Drug Issues*, *23*(2), 323-334.

IBM. (2014). *X-Force Threat Intelligence Quarterly 1Q 2014*. IBM Security Systems. Retrieved from http://spyders.ca/wp-content/uploads/2014/04/IBM-X-Force-Threat-Intelligence-Quarterly-Q1-2014.pdf

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, *52*(1/2), 225-247.

Jones, B. R. (2007). Virtual Neighbourhood Watch: Open Source Software and Community Policing against Cybercrime. *Journal of Criminal Law and Criminology*, *97*(2), 601-630.

Kalm, K. (2013, June). Illicit network structures in cyberspace. In *2013 5th International Conference on Cyber Conflict (CyCon)* (pp. 1-13). IEEE.

Keegan, B., Ahmed, M. A., Williams, D., Srivastava, J., & Contractor, N. (2010, August). Dark gold: Statistical properties of clandestine networks in massively multiplayer online games. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on* (pp. 201-208). IEEE.

Kendall, L. W. (1999). Robustness of closed capture-recapture methods to violations of the closure assumption. *Ecology*, *80*(8), 2517-2525.

Kilger, M., Arkin, O., & Stutzman, J. (2004). Profiling. *The Honeynet Project (2nd Ed.), Know your enemy.* Addison Wesley Professional.

Kong, R. (2006). *A Feasibility Report on Improving the Measurement of Fraud in Canada, 2005* (No. 85-569-XIE). Retrieved from http://www.statcan.gc.ca/pub/85-569-x/85-569-x2006001-eng.pdf

Leukfeldt, E. R. (2014). Cybercrime and social ties. *Trends in organized crime*, *17*(4), 231-249.

Leukfeldt, E. R. (2015). Organised Cybercrime and Social Opportunity Structures: A Proposal for Future Research Directions. *The European Review of Organised Crime*, *2*(2), 91-103.

Leukfeldt, E. R., Kleemans, E. R., & Stol, W. P. (2016). Cybercriminal Networks, Social Ties and Online Forums: Social Ties versus Digital Ties within Phishing and Malware Networks. *British Journal of Criminology*, *56*(2), 1-19. doi:10.1093/bjc/azw009

Levi, M., & Williams, M. (2013). Multi-agency partnerships in cybercrime reduction: Mapping the UK information assurance network cooperation space. *Information Management & Computer Security*, *21*(5), 420-443.

Leyland, A., Barnard, M., & McKeganey, N. (1993). The use of capture-recapture methodology to estimate and describe covert populations: An application to female street-working prostitution in Glasgow. *Bulletin de Méthodologie Sociologique*, *38*(1), 52-73.

Loughran, T. A., Nguyen, H., Piquero, A. R., & Fagan, J. (2013). The Returns to Criminal Capital. *American Sociological Review*, *78*(6), 925-948.

Lusthaus, J. (2012). Trust in the World of Cybercrime. *Global Crime*, *13*(2), 71-94.

Mastro, T. D., Kitayaporn, D., Weniger, B. G., Vanichseni, S., Laosunthorn, V., Uneklabh, T., & Limpakarnjanarat, K. (1994). Estimating the number of HIV-infected injection drug users in Bangkok: a capture-recapture method. *American Journal of Public Health*, *84*(7), 1094-1099.

Mazowita, B. & Vézina, M. (2014) Police-reported cybercrime in Canada, 2012. *Juristat*, *34*(1), 1-24. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2014001/article/14093-eng.pdf

Mcguire, M. (2012). *Organised Crime in the Digital Age*. London, UK: John Grieve Centre for Policing and Security.

MoneyGram. (2008). *Canada Agent Anti-Money Laundering and Terrorist Financing Prevention Compliance Regime*. Retrieved from http://corporate.moneygram.com/Documents/Corp%20site%20docs/Compliance/Canada%20Agents/014165.pdf

Moore, T., Clayton, R., & Anderson, R. (2009). The economics of online crime. *The Journal of Economic Perspectives*, *23*(3), 3-20.

Morselli, C. (2003). Career opportunities and network-based privileges in the Cosa Nostra. *Crime, Law and Social Change*, *39*(4), 383-418.

Morselli, C., & Giguère, C. (2006). Legitimate strengths in criminal networks. *Crime, Law and Social Change*, *45*(3), 185-200.

Morselli, C., Giguère, C., & Petit, K. (2007). The efficiency/security trade-off in criminal networks. *Social Networks*, *29*(1), 143-153.

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

Morselli, C., Grund, T. U., & Boivin, R. (2015). Network Stability Issues in a Co-Offending Population. In G. Bichler & A. E. Malm (Eds.), *Disrupting Criminal Networks: Network Analysis in Crime Prevention* (pp. 47-65). Boulder, CO: FirstForumPress.

Motoyama, M., McCoy, D., Levchenko, K., Savage, S., & Voelker, G., M. (2011). An Analysis of Underground Forums. *Proceedings of the 2011 ACM SIGCOMM Conference* (pp. 7180-7186). ACM.

Nagurney, A. (2015). A multiproduct network economic model of cybercrime in financial services. *Service Science*, *7*(1), 70-81.

Newman, R. M., & Waters, T. F. (1989). Differences in brown trout (Salmo trutta) production among contiguous sections of an entire stream. *Canadian Journal of Fisheries and Aquatic Sciences*, *46*(2), 203-213.

Nonnecke, B., & Preece, J. (2001, December). *Proceedings of the 7th Americas Conference on Information Systems* (pp. 1521-1530).

Peretti, K. (2008). Data Breaches: What the Underground World of Carding Reveals. *Santa Clara High Technology Law Journal*, *25*(2), 375-413.

Perreault, S. (2011). Self-reported Internet victimization in Canada, 2009. *Juristat*, *36*(1), 1-31. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2011001/article/11530-eng.pdf

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

Piquero, N. L., & Benson, M. L. (2004). White-collar crime and criminal careers specifying a trajectory of punctuated situational offending. *Journal of Contemporary Criminal Justice*, *20*(2), 148-165.

Pollock, K. H., Nichols, J. D., Brownie, C., & Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs*, *107*(1), 3-97.

Ponemon Institute. (2015). *2015 Megatrends in Cybersecurity*. Retrieved from http://www.raytheon.com/news/rtnwcm/groups/gallery/documents/content/rtn_233811.pdf

Preece, J., Nonnecke, B., & Andrews, D. (2004). The top five reasons for lurking: improving community experiences for everyone. *Computers in human behavior*, *20*(2), 201-223.

Rafaeli, S., Ravid, G., & Soroka, V. (2004, January). De-lurking in virtual communities: a social communication network approach to measuring the effects of social and cultural capital. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004* (pp. 10-pp). IEEE.

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

Reuter, P. (2013). Are estimates of the volume of money laundering either feasible or useful? In B. Unger and D. van der Linde (Eds.), *Research Handbook on Money Laundering* (pp. 224-231). Cheltenham, UK: Edward Elgar Publishing.

Rhodes, W. (1993). Synthetic estimation applied to the prevalence of drug use. *Journal of Drug Issues*, *23*(2), 297-321.

Riccio, L. J., & Finkelstein, R. (1985). Using police arrest and clearance data to estimate the number of burglars operating in a suburban county. *Journal of Criminal Justice*, *13*(1), 65-73.

Richet, J. L. (2013). *Laundering Money Online: A Review of Cybercriminals' Methods*. United Nations Office on Drugs and Crime. Retrieved from http://arxiv.org/ftp/arxiv/papers/1310/1310.2368.pdf

Roberts, J. M., & Brewer, D. D. (2006). Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capture–recapture method. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 745-756.

Romanosky, S., Telang, R., & Acquisti, A. (2011). Do Data Breach Disclosure Laws Reduce Identity Theft? *Journal of Policy Analysis and Management*, *30*(2), 256-286.

Rossmo, D. K., & Routledge, R. (1990). Estimating the Size of Criminal Populations. *Journal of Quantitative Criminology*, *6*(3), 293-314.

Science and Technology Committee. (2007). *Personal Internet Security* (Vol. 2). House of Lords. Retrieved from http://www.publications.parliament.uk/pa/ld200607/ldselect/ldsctech/165/165i.pdf

Seber, G. A. (1965). A note on the multiple-recapture census. *Biometrika*, *52*(1/2), 249-259.

Silver, W. (2007). Crime statistics in Canada, 2006. *Juristat*, *27*(5), 1-15. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/85-002-x2007005-eng.pdf

Solicitor General of Canada. (1998). *Electronic Money Laundering: An Environmental Scan*. Department of Justice Canada. Retrieved from http://justice.gc.ca/eng/pi/rs/rep-rap/1998/wd98_9dt98_9/wd98_9.pdf

Soudijn, M., R., J., & Zegers, B. C. H. T. (2012). Cybercrime and virtual offender converge settings. *Trends in Organized Crime*, *15*(2), 111-129.

Sparrow, M. K. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks*, *13*(3), 251-274.

Taylor-Butts, A., & Perreault, S. (2009). *Fraud Against Businesses in Canada: Results from a National Survey* (No. 85-571-X). Retrieved from http://www.statcan.gc.ca/pub/85-571-x/85-571-x2009001-eng.pdf

Tcherni, M., Davies, A., Lopes, G., & Lizotte, A. (2016). The Dark Figure of Online Property Crime: Is Cyberspace Hiding a Crime Wave? *Justice Quarterly*, *33*(5), 890-911. doi: 10.1080/07418825.2014.994658

Thomas, K., Huang, D., Wang, D., Bursztein, E., Grier, C., Holt, T. J., Kruegel, C., McCoy, D., Savage, S., & Vigna, G. (2015). Framing Dependencies Introduced by Underground Commoditization. *Proceedings of the Workshop on Economics of Information Security (WEIS)* (pp. 1-24).

Tremblay, P., Bouchard, M., & Petit, S. (2009). The size and influence of a criminal organization: A criminal achievement perspective. *Global Crime*, *10*(1-2), 24-40.

Trend Micro. (2015). *Report on Cybersecurity and Critical Infrastructure in the Americas*. Organization of American States. Retrieved from http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/reports/critical-infrastructures-west-hemisphere.pdf

United Nations. (2015). *Cybersecurity: A global issue demanding a global approach*. United Nations Department of Economics and Social Affairs. Retrieved from http://www.un.org/en/development/desa/news/ecosoc/cybersecurity-demands-global-approach.html

van der Heijden, P. G., Cruyff, M., & van Houwelingen, H. C. (2003). Estimating the size of a criminal population from police records using the truncated Poisson regression model. *Statistica Neerlandica*, *57*(3), 289-304.

van der Heijden, P. G. M., Cruts, G., & Cruyff, M. (2013). Methods for population size estimation of problem drug users using a single registration. *International Journal of Drug Policy*, *24*(2), 614-618. doi: 10.1016/j.drugpo.2013.04.002

van der Heijden, P. G., Cruyff, M., & Böhning, D. (2014). Capture Recapture to Estimate Criminal Populations. In *Encyclopedia of Criminology and Criminal Justice* (pp. 267-276). New York, NY: Springer New York.

van der Heijden, P. G., de Vries, L., Böhning, D., & Cruyff, M. (2015). Estimating the size of hard-to-reach populations using capture-recapture methodology, with a discussion of the International Labour Organization's global estimate of forced labour. K. Kangaspunta (Ed.), *United Nations Office on Drugs and Crime Forum on Crime and Society*, Vol. 8. New York, NY: United Nations.

Vitali, S., Glattfelder, J. B., & Battiston, S. (2011). The network of global corporate control. *PloS one*, *6*(10), 1-6.

Wallace, M. (2009). Police-reported crime statistics in Canada, 2008. *Juristat*, *29*(3), 1-37. Retrieved from http://www.statcan.gc.ca/pub/85-002-x/2009003/article/10902-eng.pdf

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440-442.

Weaver, R., & Collins, M. P. (2007, October). Fishing for phishes: Applying capture-recapture methods to estimate phishing populations. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit* (pp. 14-25). ACM.

Wechinger, F. (2011). The Dark Net: Self-Regulation Dynamics of Illegal Online Marketplaces for Identities and Related Services. *Intelligence and Security Informatics Conference* (pp. 209-213). IEEE.

Westlake, B. G., Bouchard, M., & Frank, R. (2011). Finding the key players in online child exploitation networks. *Policy & Internet*, *3*(2), 1-32.

Westlake, B. G., & Bouchard, M. (2016). Liking and hyperlinking: Community detection in online child sexual exploitation networks. *Social Science Research* (ahead of print), 1-14. doi: 10.1016/j.ssrresearch.2016.04.010

Western Union. (2012, May). *Agent Anti-Money Laundering Compliance Manual*. Western Union Financial Services (Canada), Inc. Retrieved from https://weborder.harlingdirect.com/documents/products/product_3021_details.pdf

Wickens, T. D. (1993). Quantitative methods for estimating the size of a drug-using population. *Journal of Drug Issues*, *23*(2), 185-216.

Williams, M., & Levi, M. (2012). Perceptions of the eCrime controllers: Modelling the influence of cooperation and data source factors. *Security Journal*, *28*(3), 252-271. doi: 10.1057/sj.2012.47

Williams, M., & Burnap, P. (2016). Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. *British Journal of Criminology*, *56(2),* 211-238.

Williams, M., Burnap, P., & Sloan, L. (2016). Crime Sensing with Big Data: The Affordances and Limitations of using Open Source Communications to Estimate Crime Patterns. *British Journal of Criminology* (ahead of print), 1-21. doi: 10.1093/bjc/azw031

Wolfgang, M. E., Figlio, R. M., & Sellin, T. (1972). Delinquency in a birth cohort. Chicago, IL: University of Chicago Press.

World Bank. (2016). *Internet users (per 100 people)*. World Bank Open Data. Retrieved from http://data.worldbank.org/indicator/IT.NET.USER.P2?locations=CA

Yip, M., Shadbolt, N., & Webber, C. (2012, June). Structural analysis of online criminal social networks. In *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on* (pp. 60-65). IEEE.

Yip, M., Webber, C., & Shadbolt, N. (2013). Trust among cybercriminals? Carding forums, uncertainty and implications for policing. *Policing and Society*, *23*(4), 516-539.

Zelterman, D. (1988). Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of Statistical Planning and Inference*, *18*(2), 225-237.

Zhang, S., & Chin, K. L. (2002). Enter the Dragon: Inside Chinese Human Smuggling Organizations. *Criminology*, *40*(4), 737-768.

Zhang, S., & Chin, K. L. (2003). The declining significance of triad societies in transnational illegal activities: A structural deficiency perspective. *British Journal of Criminology*, *43*(3), 469-488.

Zhao, Z., Sankaran, M., Ahn, G. J., Holt, T. J., Jing, Y., & Hu, H. (2016). Mules, Seals, and Attacking Tools: Analyzing 12 Online Marketplaces. *IEEE Security & Privacy*, *14*(3), 32-43.