Sifting through the Net: Monitoring of Online Offenders by Researchers

Research note

Sifting through the Net: Monitoring of Online Offenders by Researchers

David Décary-Hétu and Judith Aldridge*

Abstract: Criminologists have traditionally used official records, interviews, surveys, and observation to gather data on offenders. Over the past two decades, more and more illegal activities have been conducted on or facilitated by the Internet. This shift towards the virtual is important for criminologists as traces of offenders' activities can be accessed and monitored, given the right tools and techniques. This paper will discuss three techniques that can be used by criminologists looking to gather data on offenders who operate online: 1) *mirroring*, which takes a static image of an online resource like websites or forums; 2) *monitoring*, which involves an on-going observation of static and dynamic resources like websites and forums but also online marketplaces and chat rooms and; 3) *leaks,* which involve downloading of data placed online by offenders or left by them unwittingly. This paper will focus on how these tools can be developed by social scientists, drawing in part on our experience developing a tool to monitor online drug "cryptomarkets" like Silk Road and its successors. Special attention will be given to the challenges that researchers may face when developing their own custom tool, as well as the ethical considerations that arise from the automatic collection of data online.

Keywords: Cybercrime – Internet research methods – Crawler – Cryptomarkets avid e cary- e tu
is Assistant Professor at the School of Criminology, University of Montreal, and Researcher at the International Centre of Comparative Criminology (ICCC). Email: david.decary-hetu@umontreal.ca

Judith Aldridge is Senior Lecturer at the School of Law, University of Manchester.

The European Review of Organised Crime 2(2), 2015. 122-141

ISSN: 2312-1653

Introduction

Connected services and devices are more and more a part of our daily lives. We spend most of our days connected to the Internet in one way or another (Oliveira, 2014), so much so that it is now difficult to differentiate between time spent online and offline. This is also of course true for offenders. Offenders converge in online settings to communicate and collaborate within criminal networks using a crime-as-a-service model, thereby increasing their efficiency while limiting their risks. In this paper, we show how researchers can take advantage of this shift to online convergence settings through developing new techniques to monitor and better understand offenders. We divide the tools available to researchers into three categories: mirroring, active monitoring, and leaks. We then describe the experience of developing an online monitoring tool based on our own experience of building one. We conclude with a discussion of the future challenges that researchers will face when using online traces left by offenders for research purposes as well as the ethical considerations that need to be addressed by researchers.

The Rise of the Network Society

The term *network society* comes from Castells (1996) who described the development and adoption of information technologies that made time and space constraints virtually disappear through instantaneous communications. These changes led to the globalisation of social interactions and business relationships, shifting interactions from a bureaucratic and hierarchical structure to a horizontal and networked one. This created much more fluid communications over multiple coexisting networks (Wellman, 2002), enabling actors to join in or leave networks as their ability to communicate and participate developed. The network society and the ubiquity of the Internet means that we can now be a part of many networks at the same time (Boase and Wellman, 2006). Relationships inside these networks are sparsely-knit and ephemeral in nature, often connecting individuals who may not share many common traits. Living in a networked society means an increased social network for people both professionally and personally. A body of literature has adopted this network framework to understand offenders (Morselli, 2009, Sparrow, 1991, Krebs, 2002). Offenders are analysed as entrepreneurs who collaborate with their peers on a project-by-project basis, the “crime-as-a-service” model (Manky, 2013). This can be seen in the case of online financial fraud where fraudsters will network with hackers to develop viruses that can take over computers and steal credit card information (e.g. Holt et al., 2008). Once the virus is written, the fraudsters and malware writers split up and may never work together again. Similar relationships exist between fake or stolen prescription vendors and spam specialists (e.g. Krebs, 2014). Spammers are hired to deliver ads to potential customers who are directed to websites owned by the prescription vendors. Once the spam campaign is completed, prescription vendors can decide to hire a different spammer or use another method altogether to reach their potential customers.

This networked social organisation has created more than ever a need for convergence settings where offenders can meet, network, and advertise their goods and services (Motoyama et al., 2011). Online settings for this activity include discussion forums, chat rooms, and newsgroups. Online convergence settings offer both synchronous and asynchronous methods of communications that can be open or private. Online convergence settings represent an opportunity for criminologists who can take advantage of the vast quantity of online traces to better understand offenders (Chen, 2011). Indeed, many settings have now been active for over a decade and have stored hundreds of thousands if not millions of public messages and member profiles. These messages and profiles provide an image of offenders that includes information additional to their illegal activities: as offenders spend time networking with others, they often discuss personal and philosophical topics, providing an expanded understanding of offenders and their characteristics (Holt, 2010). Messages are typically posted online under handles (fake names) which do not change over time and across settings, given the time and energy invested in creating these online personas. This enables researchers to study the evolution of online communities and criminal organisations through time.

Although there is a lack of empirical evidence that some “traditional” criminal organisations have moved parts of their activities online, some hypothesise that the growing profits to be made online will inevitably draw these groups to the Internet. Emerging research has documented these moves in relation to activities including smuggling and online gambling and in various locations including China and some former U.S.S.R. countries (see, for instance, Broadhurst et al., 2014; Bhattacharjee, 2011; Kshetri, 2013; Lavorgna, 2015; Lavorgna and Sergi, 2014).

The Internet as a Source of Data in Academic Research

The Locard principle stipulates that all criminal activities leave traces (Horswell and Fowler, 2004). As we move further and further into an always connected, networked world, offenders are increasingly likely to interact with each other in online convergence settings—and to leave traces of their interactions online. Online traces have been collected and used by researchers and criminologists for over two decades. Such traces can be collected either manually or automatically.

Manual collection of online traces (see Mann and Sutton, 1998; Durkin and Bryant, 1999; Williams and Copes, 2005 for examples) can be used even by researchers with limited technological skills. It involves copying and pasting online content from discussion forums or newsgroups in text files stored on researchers’ computers. While time consuming, this process has been used to gather relatively large datasets (e.g., the first page of 285 discussion threads in the case of Williams and Copes, 2005) and has provided deep insight into the characteristics and operations of offenders.

Automatic collection, on the other hand, is the use of software to automatically collect the content located on a web server or in online chat rooms or newsgroups. Automatic collection provides

advantages over manual collection as it allows for the collection locally of all the data posted in these convergence settings rather than only a subset, meaning that researchers can have access to very large and powerful datasets that are by definition representative since no sampling is involved. This provides researchers with a permanent and secure copy of the data. Williams and Copes (2005) describe how the forum they were monitoring lost all of its data due to some unknown computer problem. This means that a vast trove of information was forever lost and that researchers who have manually collected only partial data will be unable to go back to the source, for example if they need to validate or collect additional data. It is also often easier to search for content locally rather than online. Indeed, many online platforms do not have adequate search engines to find the content researchers want. An online drug market may not allow visitors to select listings from a specific country or for a specific type of drug, for example. Data that has been collected automatically can be indexed and searched very easily using the built-in tools of modern operating systems. Finally, automatic collection empowers researchers to collect very large datasets that can be used to study wide-ranging phenomena and large communities. Décary-Hétu et al. (2012) demonstrated that it was possible to explain the social structure of the hacker community that specialises in the illegal distribution of copyrighted content such as movies, software and books. This data was collected automatically through the *mirroring* (more on this technique below) of a website where hackers posted the name of the products they had illegally distributed over previous years. These features of automatic data collection do not make manual collection obsolete. On the contrary, manual collection is still useful for researchers with less developed technological skills and/or researchers who know exactly what they need and are able to find it using online platform search engines. In many cases however, the automatic collection of online data may be preferable for the reasons presented above.

There are three types of automatic online data collection: the mirroring of traces, the active monitoring of traces, and the exploitation of leaks. *Mirroring*, also known as web crawling, is the indexing and copying of web pages (Olston and Najork, 2010). This is the technique that Google uses to index the Internet. Crawlers—custom software built to mirror websites—start by downloading a single web page and by indexing all the hyperlinks it contains. Crawlers then visit the linked web pages one by one, searching for more content to download and more links to follow. This process has the advantage of capturing all the traces left online on a web site like a discussion forum and requires very little manual work. As the crawler downloads the raw HTML code, moreover, it is possible to search that code for traces that may be hidden in comments or invisible text. Such traces may include the name of the programmer or clues about his or her location. To mirror websites, HTTrack is commonly used, as it is free and fairly user-friendly (Marill et al., 2004). The software must be used in conjunction with another class of software known as web scrapers. Crawlers like HTTrack will only crawl web sites and download web pages, and are unable to extract key information from web pages; that is the web scraper's role. Web scrapers can be taught what content is important on a web page (e.g., name of person posting a message, content of message, date the message was posted) and how to store that information in a database or spreadsheet. This can be challenging if the web pages collected do not have a common layout and/or structure. The scraper must be able to recognise the desired content to extract and doing so requires a level

of similarity across the web pages. Web crawlers are very easy to detect for system administrators as they tend to follow a discernible pattern of downloading rapidly one page after the other. Most webmasters do not care or take action against web crawlers but it has been our experience that some may go as far as blacklisting the IP address of the web crawler. Some webmasters have no choice but to take defensive measures like these given that crawlers can put a burden on web servers if they are not carefully set up. By rapidly downloading a large number of web pages, they can add to the load of a server, sometimes making it unable to deliver web pages to legitimate visitors (Thelwall and Stuart, 2006). Most crawlers have a setting which limits the speed at which a website is crawled in order to prevent this from happening.

There are a number of examples of researchers employing mirroring techniques to understand criminal networks. Christin (2013) used the HTTrack software to index all of the listings, vendor profiles and feedback from the original *Silk Road* marketplace, an online, anonymous illicit drug market. Chen's work (2012) provides an even more detailed account of how mirroring can be used to gather data on terrorists and other types of offenders. Finally, Décary-Hétu et al. (2014) created their own custom tool to download a list of all the pirated copyrighted content that was distributed online between 2003 and 2009. Their study demonstrates the importance of peer reputation as well as the scale on which mirroring can be used.

A second method of gathering online traces is the *active monitoring* of the kinds of traces that emerge from the synchronous and more ephemeral communications that occur in online chatrooms and social networks like Instagram, Facebook, and Twitter (Fallmann et al., 2010). Content on these platforms is often short lived and needs to be collected as soon as it is posted, before it is taken down or replaced by newer content. Active monitoring crawlers must be able to monitor server communications, analyse their content, and extract the required information from them. These crawlers must be able to deal with large simultaneous influxes of data where many individuals share content at the same time. Offenders who network through synchronous communications are typically wary of being monitored. They often protect the convergence settings where they meet with passwords or remove unknown or inactive participants that only 'listen' and never participate in the communications. Active monitoring crawlers must therefore be able to connect and reconnect automatically and to change their online pseudonyms dynamically in order to keep monitoring offenders effectively. While active monitoring provides untainted traces of offenders, some participants may be aware that their communications are monitored. They could modify their behaviour accordingly, even if it may prove difficult for them to keep their guards up for extended periods of time. Chat room and social network communications generate rich qualitative data that provides in-depth understanding of convergence settings. This comes at a cost, as setting up an active monitoring crawler can be difficult. Gaining access to the most private convergence settings also takes time and requires researchers to interact with offenders to gain their trust, thereby carrying with it additional implications as regards the ethics of conducting this kind of research.

There are a number of examples of researchers actively monitoring traces to understand criminal

networks. Décary-Hétu et al. (2014) used active monitoring to gather data on offenders who talked about hacking in Internet Relay Chat (IRC) chat rooms. Their work led to a methodology for building activity logs to detect offenders who use multiple online identities. Franklin et al. (2007) used a similar technique to find that many offenders who sold stolen credit card numbers in IRC chat rooms were actually scammers trying to steal from naïve buyers. Stone-Gross et al. (2009) monitored IRC chat rooms where hackers who had taken control of over 180,000 computers met, enabling these researchers to understand how the hackers communicated with the infected computers, and how this could be prevented.

The third and final type of trace that can be gathered is known as a *leak*. Criminal markets are by nature competitive with participants fighting for market share (Reuter, 1983). Given the impossibility of establishing public credibility or advertising, offenders must use their reputation to find and attract new partners (Décary-Hétu, 2013). As reputation is one of the most prized assets of offenders (Anderson, 1999), it is often reputation itself that is the target of offenders who want to harm their competitors. One way to attack reputation is to release an offender's private information online. This practice, known as *doxing (Coleman, 2014)*, provides researchers with information about the identity of offenders. Such information is often posted to text-sharing websites like Pastebin where the poster of the information can remain anonymous. Leaks can also target whole convergence settings. Many backups of discussion forums have been leaked online (see Motoyama et al., 2011), disclosing publicly all of the public and private messages of the forum participants as well as administrative information connected to participants (IP addresses, promotions, and rankings). Leaks are extremely useful traces as they provide information that would not normally be available publicly. They are however of unknown origin and it is often not possible to verify if these traces have been tampered with in any way. Leaks tend to be taken down rapidly and so must be downloaded as soon as they are published.

Motoyama et al. (2011) collected leaks from six different forums and measured their participants' social network as well as forum dynamics and regulation. They found that offenders who had a higher status, a good reputation and a large social network were more successful in selling illicit goods online. They also found that many participants were banned from forums, mainly for trying to create multiple accounts to scam others. The work of Afroz et al. (2013) is based on a similar dataset of traces and defines the characteristics of a successful convergence setting. These include a growing number of participants over time, effective official regulation by forum administrators and easy communication tools.

Collecting traces of illegal activity online provides researchers with new and innovative datasets that are free from the bias of official criminal justice derived data. Data derived from criminal justice agencies usually either relates to small geographical areas, or is a sample of a larger population. Moreover, data derived from these sources always involves offenders who have come into contact with the criminal justice system, and not with those who have not—excluding therefore the more "successful" offenders (Oosthoek, 1978). Digital traces collected automatically provide data on *all* offenders operating in those settings: those that have been in contact with the criminal justice

system as well as those that have not. This offers a more representative picture of offending communities (see Décary-Hétu et al., 2014). As valuable at it is however, this methodology requires skills that only minority of researchers possess. The next section will seek to help researchers interested by this methodology by presenting the steps and challenges associated with the development of an automatic and online data collection tool.

Creating a Custom Web Crawler

Using the Internet as a source of data for academic research poses many challenges, even for researchers who are knowledgeable of new digital technologies. In this section, we will draw on our own experience of developing the DATACRYPTO software, a tool for monitoring the sale of illicit goods and services on online hidden marketplaces, to explain how a custom web crawler could be developed. The intention is to provide researchers who may be considering the collection of online traces to study criminality with a sense of how one might go about doing so, drawing in part on our experiences and the challenges we faced. DATACRYPTO is a crawler that mirrors and scrapes web sites and builds databases of product listings, vendor information and buyer feedback.

As a reference, the development of the DATACRYPTO tool was initially estimated to take about 4 months' time and cost around $13,000 US .

The first step when developing a crawler is to define the specifications of the tool (i.e., what data should be collected and how it should be presented to the researchers). Past research can provide an indication as to the type of data that would be useful to collect. In the case of online marketplaces, product listings and vendor profiles are often crawled and scraped (Christin, 2013; Aldridge and Décary-Hétu, 2014; Dolliver, 2015) though other platforms besides cryptomarkets may include friends' lists and images that may also be of interest. The most convenient format to store the data is the commaseparated value (csv) file type as it can be read by most software. No matter the format, data should be collected in a uniform fashion across multiple online data sources, allowing queries to be run against multiple sources simultaneously. This would allow, for example, a researcher to quickly build a dataset of all vendors selling (say) cannabis during a period of time no matter what market they were active on. This kind of feature allows researchers to describe and explain phenomena in general and not just the dynamics of a specific market, website, or online community. An additional important specification is the characterisation of the connection protocols that the crawler can use. Crawlers usually connect to online data sources through regular (" TTP") connections. Some sources however are only accessible through secured connections (" TTPS") that demand more work on the programmer's end. Online data sources sometimes protect the identity of their participants by hosting their infrastructure on the Tor network (Dingledine et al., 2004). This network is used to hide the true location of servers and to protect the identity of participants by routing Internet traffic through multiple anonymous relays. Other online data sources have also adopted an alternative to Tor, the Invisible Internet Protocol (I2P) (see Zantout and Haraty, 2011 for more details) which works in a similar fashion. These two technologies reinforce the need to evaluate the protocols that are used by the online data sources

and to include them in the crawler's specifications.

The second step when developing a crawler is to define the level of automation that is needed. Christin (2013) and Dolliver (2015) both used the HTTrack software tool, which needs to be configured and launched manually and which does not extract the information from the web pages it collects. Custom crawlers can be much more feature-packed. They can be set to launch at regular intervals and store the login information so that they are able to login to websites in order to collect data that is not openly accessible. The root page of a website can also be stored, providing the crawler with a fixed entry point from which to index a website. Most websites provide crawlers with *robots.txt* files that list the sections of the website that should be indexed and those that should not. Custom crawlers can be provided with similar files to reduce the time they spend indexing websites, focusing only on the sections that researchers require.

Once the specifications and automation are defined, researchers should seek a reliable developer by putting out a call for bids. These calls mostly draw the attention of software firms who can charge upwards of $100 USD per hour for their time. Given the limited budget of most research projects, a good alternative is to work with independent freelance developers who typically bill an hourly wage of $50 USD or less. Many websites offer freelancers with an opportunity to showcase their work and the feedback they have received from past clients (see Freelancer.com for example). The quality of the work evidenced on these sites varies considerably, and communication can sometimes be difficult when the researcher and the developer do not share the same first language; this is not uncommon. Researchers should also take advantage of the milestone feature of such sites which allows those hiring to set deadlines for the different parts of a projects, releasing payment at intervals once milestones have been reached.

It is important to manage freelancers on a day-to-day basis. Developers often have questions about the specification of projects, and researchers need to keep on top of the freelancers' progress. Selecting a freelancer that is in the same time zone as the researchers, therefore, can be critical to ensuring effective communication. Planning phone meetings and chats can be tricky when there is a large time difference between the researcher and the developer, with one or the other often needing communication to occur outside of typical 9-5 working hours, and therefore involving intrusion into late evenings or early mornings. This is often overlooked, but as we found, a crucial issue for researchers to keep in mind.

Problems with freelancers can have very important consequences for research projects. Should a freelancer quit or be fired, as was the case with the development of DATACRYPTO, other freelancers will seldom accept to build on the work begun by another developer, because it is usually more time consuming to pick up a partlycompleted development job than to develop from scratch. Because of this, all of the money invested in a freelancer who does not finish a project will likely need to be invested again in a new programmer. This makes working with freelancers a sometimes delicate undertaking, as a crawler that is not 100% complete will have to be discarded—whether it was 20 percent or 99 percent completed.

Challenges to Consider

Online convergence settings offer valuable opportunities for researchers to collect traces on offenders. The last section presented a non-technical explanation of the work involved in building a web crawler. This section will present the three main challenges that may impede the development of a web crawler.

The first challenge that researchers will face is the need to develop an understanding of the convergence settings they want to monitor, both from operational and technical points of view. Preliminary research should focus on how convergence settings operate and who their participants are. In the case of cryptomarkets that sell illegal goods and services, for example, some marketplaces keep all of the buyers' feedbacks while others only show the last *n* feedbacks, severely curtailing the ability for researchers to analyse the vendor-buyer relationships. Not understanding the feedback retention policy could seriously affect the quality of analyses with these aims, potentially leading to underestimates or overestimates. Researchers must also seek to understand the technologies used in the convergence settings they wish to monitor. This will allow them to provide freelancers with more detailed instructions and to closely monitor the work they do. There is no need for researchers to learn how to code; there is a need however to understand how coding works and what tools are at their disposal. This is very transparent in the case of CAPTCHAs. Many websites require that their users solve a CAPTCHA (a series of distorted characters that are difficult to read by computers) to log in. Freelancers may offer to build complex software that can read CAPTCHAs but it is much easier to connect to commercial services such as DEATH BY CAPTCHA. This service solves CAPTCHAs by enlisting the help of workers who are paid a few pennies each time they solve a CAPTCHA. Their answer to the CAPTCHA can be returned automatically to a crawler which uses it to log in to a website. Understanding how a convergence setting works ensures the quality of the data and the analyses while understanding technology ensures that the data collected is gathered as efficiently as possible.

While it is time-consuming for researchers to gain this understanding, it is even more challenging to maintain this knowledge over time. Indeed, new convergence settings using new technologies appear every month. Researchers need to adapt to this constant innovation by those who carry out illegal activities online. We are seeing already that convergence settings are becoming more and more difficult to locate using search engines like Google. Tens of financial fraud forums can be found via Google but the more private and elite convergence settings are increasingly accessible by invitation only in an attempt to evade the gaze of law enforcement agencies. Russian hacking forums, for example, operate effectively by allowing entry only to other Russians who can prove their location by providing answers to questions that those from other cultural/national settings could not manage successfully, thereby keeping the more powerful Western law enforcement agencies (as well as many researchers) away. Convergence settings also take advantage of legitimate technology, like the anti-robot service of CloudFlare, to protect themselves from

crawlers. The CloudFlare service makes sure that humans are connecting to a convergence setting using a series of tests that are difficult for robots to bypass. This evolution means that the current tools must be updated to evade these countermeasures (we discuss more on the ethical implications of this below). It also means that human intelligence will become more and more important as researchers will not be able to rely on automatic tools to identify convergence settings and harvest their data. Researchers will need to read the messages that offenders exchange in order to identify the latest trends and the more covert convergence settings.

The final challenge that researchers face is the management of big data. Automatic tools can collect millions of traces each day. Aggregating, sorting, and manipulating such large datasets (e.g., forums with hundreds of thousands of messages) are cumbersome activities that require creative thinking to manage effectively. One solution is to break these datasets into smaller subsets that can be stored and manipulated more easily. Another solution is to invest in hardware and software that can handle databases with hundreds of millions of entries. Researchers should work in teams where the responsibility of managing the data collected is distributed to those most suited to do it.

Ethical Considerations

The collection of Internet-derived content has raised many ethical issues, which include the public vs private nature of content published online, the informed consent of research participants, the harms to others and self, deception by researchers, and the potential for tactical displacement. These issues will be addressed in turn below.

The first question that the collection and use of Internet-derived content poses is whether that content is public or private in nature. Some, like Kitchin (2002), argue that content posted online in open settings like discussion forums is in the public domain, akin to documents in public archives, and so not subject to the same ethical considerations that apply to research with human subjects. The use of such content by researchers poses very little risk for those who post that content, according to this view, and should therefore be considered publicly available, not therefore requiring the consent of their authors. This stance is bolstered by the fact that authors typically use nicknames or pseudonyms and often refuse to divulge personal information. This argument assumes that authors will be aware that their content may be read by others beyond the intended recipients precisely because of its public nature. A more nuanced approach to the public versus private question focuses on the expectation that the authors have when they post content online (Binik et al., 1999; Sveningsson Elm, 2008). Working out what should be considered “public” and “private” on the Internet is not always straightforward. There exists a range of locations in which potential content may reside: in completely open locations on the “clearnet”, in locations that require solving CAPTCHAs to access, in locations requiring a registered account to view, and in locations requiring an invitation by members of a virtual community or its moderators. This

approach skirts the complexity of making the “public/private” determination by instead attempting to ascertain the expectations of the communities of individuals who form the subject of a research in order to determine whether obtaining individual consent is required. Researchers, however, may not be equipped to determine the expectations for privacy of individuals participating in these forums, and not all individuals will share the same expectation of privacy (Barratt, 2011). The question is then to determine whether researchers should set the bar according to the most open or the most private individual. Rosenberg (2010) suggests that researchers may turn to the norms of the virtual community itself for guidance. With the example of scraping cryptomarkets that bring together vendors selling illegal goods and services, many of these markets explicitly espouse “crypto-anarchist” (see May, 1994) and radical libertarian principles, leading us and others (e.g., Christin, 2013) to conclude these particular communities would have viewed their content as public.

Standard ethical practice dictates that those who participate in research should be given the opportunity to provide their consent after being fully informed of the purpose of the research, how the data they contribute will be used, alongside the implications of their participation (Berg et al., 2004). When Mann and Sutton (1998) collected digital trace data about hackers, they argued that it was not necessary to obtain informed consent as they were merely lurking and not interacting with this community of hackers, a view that is supported by some others (Garton et al., 1997, Finn and Lavitt, 1994, Reid, 1996). This lower threshold for consent draws on the assumption that individuals should have a lower expectation of privacy when posting content openly on the Internet (Brownlow and O'Dell, 2002). Some have argued for the appropriateness of spending some time in online settings collecting traces in order to become familiar with the culture of a particular community before making contact to obtain consent (e.g. Barratt, 2012). Others from the virtual ethnography tradition have taken a strong stand against “lurking” as a “one way process” wherein all power resides with researchers who “appropriate” data without the kind of dialogue that gives ethnography its meaning (Bell, 2006 p. 198).

Requiring informed consent of all possible individuals (see for example King, 1996) would be extremely challenging to achieve in most cases of online automated data collection, in which the goal is to obtain full population data rather than sampling. Many participants in these locations post only one or a few messages, and may not be contactable by researchers. The sheer number of forum participants is also likely to be an insurmountable hurdle given that some online settings can have thousands if not hundreds of thousands of participants; consensus to make full crawls of these settings is an impossibility. A compromise might be the requirement for researchers to obtain consent from the moderators or administrators of these convergence settings (Sixsmith and Murray, 2001), allowing for a dialogue to ensue whereby administrators can then express their concerns or require certain concessions from the researchers. Other researchers recommend the solution of requesting a waiver from research ethics committees to sidestep asking for individual informed consent (Hine, 2008).

When researchers collect digital traces of illegal activities openly, one result is that the community may change its behavior (see Garcia et al., 2009), thereby compromising data validity. The same

can be said also of covert crawling which follows discernible patterns that may be picked up by system administrators. The adaptation or tactical displacement that results from research activity (Clarke and Eck, 2014) could push offenders to move from open convergence settings to ones that are semi-private or private. These more exclusive settings require some level of authentication such as a referral by an active member or a one-time payment. Offenders may also move to region-specific convergence settings such as those hosting Russian-language online hacker communities. Individuals can become members in these communities only if they are able to prove their Russian origin by providing answers to questions that only those having resided in Russia could (e.g., by finishing off a popular saying). Another tactical displacement could be that offenders become less inclined to discuss publicly the location of their online convergence settings, making it more difficult for researchers to identify and collect content from them. This can be accomplished by administrators when they block search engine crawlers and ask members not to post the URL for the setting on public forums and chat rooms. A third tactical displacement could be the adoption of countermeasures. These measures can limit the number of pages one account can download per session, require that a CAPTCHA be solved before accessing certain pages or use scripting to discriminate between requests that come from a robot and those that come from a web browser. This technology has been developed and implemented by companies like CloudFlare, and is currently being used in some convergence settings. Offenders can also take more proactive countermeasures and “hack back” researchers if they have the required skills (olt, 2010). Researchers’ computers hosting the crawlers should therefore be connected to protected Internet lines that have no ties with the researchers’ institutions. The computers running the crawlers should not contain identifying information about either researchers or the subjects of their research. Researchers can reduce the odds of “hack back” by making sure that their computers are up-to-date and have anti-virus software installed. They could also use application white-listing software that vastly limit the ability of viruses to install and run on these computers (Holt et al., 2014). It has been suggested that police should target offenders in online convergence settings. Offenders are therefore likely to be on the lookout for signs of monitoring, and therefore prone to tactical displacement.

Some might assume that no harm can come from researchers collecting digital traces of offenders operating online. Our experience of crawling drug cryptomarkets suggests that this may not be a safe assumption. First, although vendors who list drugs for sale on these marketplaces use pseudonyms for their usernames, after law enforcement actions on these sites (e.g., Department of Justice, 2014) where arrests have taken place, the real names of those arrested alongside their cryptomarket pseudonyms have been made publicly available. As well, participants on cryptomarkets and other forums where researchers may be gathering digital traces of criminal activity have engaged in “doxing” (hacking with the intent to expose identity) in response to conflict that sometimes arises amongst participants, again potentially making a matter of public record the link between pseudonyms and real-world identities. For this reason, where crawls include pseudonyms, the data we collect may be only temporarily anonymous. It is therefore important that dissemination of research does not include reference to pseudonyms, and that when datasets are publicly archived or shared for research purposes, that this kind of potentially identifying

information is removed. Avoiding verbatim quotes from archived content is an additional strategy to avoid this problem (Sixsmith and Murray, 2001); others (e.g. Davey et al., 2012) recommend the use of quotations that are not easily referenced through Google.

The actions of convergence setting administrators to prevent crawling must be taken seriously by researchers, but whether their actions are motivated, on the one hand, by fears about law enforcement activities or crawls by competitors, or on the other hand, concerns about research *per se* is unclear, as the following anecdote illustrates. An alleged former employee of one particular cryptomarket has written an account of being privy to a story involving the marketplace administrator having discovered that the authors of this article were crawling his marketplace, supposedly with the intent of discovering the servers and using the information for the benefit of law enforcement (though we can confirm that this was not the case). The administrator was sufficiently concerned, according to the account, to have threatened the life of these researchers (though we can also confirm that this did not happen). This story suggests that collecting digital traces of criminal activity may be threatening for operators if they perceive their activities may be revealed. It also reveals that harm to research subjects may not be the only consideration for researchers. The literature on danger to researchers themselves has generally been confined to those who carry out “field” research, typically using ethnographic methodologies (e.g., Lee-Treweek and Linkogle, 2000) but our experiences suggests this should also be a consideration for researchers seeking to collect digital traces of illegal activities.

Researchers place themselves further into ethically ambiguous situations where they engage in deception, for example by participating in discussion forums or chat rooms by posing as criminal operators. Although in general research ethics guidelines increasingly disallow research that involves deception, criminology has a long and valued history of covert research that involves some level of deception, and some have asserted that avoiding this kind of research has the inevitable consequence that some investigation is simply “handed over to journalists, undercover police, or security personnel”, arguing instead that covert research should be “celebrated and supported” (Calvey, 2013: 546). It will undoubtedly be much more difficult for researchers these days, in light of the requirement for increasingly stringent ethical review, to get approval for research that involves this kind of deception when collecting digital traces of illegal activities, but researchers who elect to go this route will need to be robust in obtaining suitable ethical approval.

Conclusion

The automated collection of Internet-derived data is an exciting development in contemporary research methods. It makes available to researchers new topics for research and provides access to data that had previously been fragmented, limited or unreliable. A good example of this is the indexing of drug prices on cryptomarkets. Law enforcement agencies, alongside researchers, have been indexing street prices of drugs for decades but their data collection methodologies have been

very fragmented and prone to sampling problems. With cryptomarkets, it is now possible to get an accurate picture of the price of a range of illicit drugs in different quantities, internationally. This is just one example of the versatility and power of Internet-derived trace data. This paper highlighted the steps and challenges associated with the development of research tools that can crawl online convergence settings. Of course, crawlers are not necessarily the best tool to use when only a sample of information is required; but even in these cases, the use of automated tools can provide detailed context and opportunities for reliability and validity checks. This paper also highlighted the complex ethical questions that center around the use of Internet-derived data concerning illegal activities. These ethical issues are by no means resolved, and researchers examining illegal activities using digital trace data must take exceptional care when considering issues like consent and anonymity insofar as the Internet has, to an extent, made it a more complicated undertaking to implement standard ethical protocols designed for research in the offline world.

References

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

Aldridge J and Décary- étu (2014) Not an “Ebay for rugs”: The Cryptomarket “Silk Road” as a Paradigm Shifting Criminal Innovation. *Available at SSRN: http://ssrn.com/abstract=2436643 or http://dx.doi.org/10.2139/ssrn.2436643*.

Anderson E (1999) *Code of the street: decency, violence, and the moral life of the inner city,*

New York, W.W Norton.

Barratt MJ (2011) Discussing illicit drugs in public internet forums: Visibility, stigma, and pseudonymity. *Proceedings of the 5th International Conference on Communities and Technologies*, ACM, 159-168.

Barratt MJ (2012) The efficacy of interviewing young drug users through online chat. *Drug and Alcohol Review* 31*:* 566-572.

Bell D (2006) *An introduction to cybercultures,* London, Routledge.

Berg BL and Lune H (2004.)*Qualitative research methods for the social sciences*, Pearson Boston, MA.

Bhattacharjee Y (2011) *Why Does A Remote Town In Romania Have So Many Cybercriminals?* [Online]. Wired. Available:

http://connection.ebscohost.com/c/articles/58844948/why-does-remote-townromania-have-so-many-cybercriminals [Accessed 1 April 2015].

Binik Y M, Mah K, and Kiesler S (1999) Ethical issues in conducting sex research on the Internet. *Journal of Sex Research* 36*:* 82-90.

Boase J and Wellman B (2006) Personal relationships: On and off the Internet. *The Cambridge Handbook of Personal Relationships***,** pp. 709-723.

Broadhurst R, Grabosky P, Alazab M, and Chon S (2013) Organizations and Cyber crime: An Analysis of

Brownlow C and O'Dell L (2002) Ethical issues for qualitative research in on-line communities. *Disability & Society* 17**:** 685-694.

Calvey D (2013) Covert Ethnography in Criminology: A Submerged yet Creative Tradition. *Current Issues in Criminal Justice* 25**:** 541.

Castells M (1996) The rise of the networked society. Cambridge, MA, and Oxford: Blackwell Publishers.

Chen H (2011) *Dark web: Exploring and data mining the dark side of the web*, Springer Science & Business Media.

Christin N (2013) Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. *Proceedings of the 22nd International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee: 213-224.

Clarke R and Eck JE (2014) *Become a Problem-Solving Crime Analyst,* Abingdon, Routledge.

Coleman G (2014) *Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous*, Verso Books.

Davey Z, Schifano F, Corazza O, and Deluca P (2012) E-Psychonauts: Conducting research in online drug forum communities. *Journal of Mental Health* 21: 386-394.

Décary-Hétu D (2013) *Le capital virtuel: entre competition, survie et réputation.* PhD, University of Montreal.

Décary-Hétu D, Dupont B, and Fortin F (2014) Policing the Hackers by Hacking Them: Studying Online Deviants in IRC Chat Rooms. *In* Masys AJ (ed) *Networks and Network Analysis for Defence and Security.* Switzerland: Springer.

Décary-Hétu D, Morselli C, and Leman-Langlois S (2012) Welcome to the Scene: A Study of Social Organization and Recognition among Warez Hackers. *Journal of Research in Crime and Delinquency* 49**:** 359-382.

Department of Justice (2014) *Dozens of Online "Dark Markets" Seized Pursant to the Forfeiture Complaint Filed in Manhattan Federal Court in Conjumption with the Arrest of hte Operator of Silk Road 2.0* [Online]. Department of Justice. Available:

http://www.justice.gov/usao/nys/pressreleases/November14/DarkMarketTake down.php [Accessed 14 March 2015].

Dolliver DS (2015) Evaluating Drug Trafficking on the Tor Network: Silk Road 2, the Sequel. *International Journal of Drug Policy*. Online first:

http://www.sciencedirect.com/science/article/pii/S0955395915000110.

Durkin KF and Bryant CD (1999) Propagandizing pederasty: a thematic analysis of the on-line exculpatory accounts of unrepentant pedophiles. *Deviant Behavior* 20**:** 103-127.

Fallmann H, Wondracek G, and Platzer C (2010) Covertly probing underground economy marketplaces. In Kreibich C and Jahnke M (eds. *Detection of Intrusions and Malware, and Vulnerability Assessment.* Berlin: Springer.

Finn J and Lavitt M (1994) Computer-based self-help groups for sexual abuse survivors. *Social Work with Groups* 17**:** 21-46.

Franklin J, Perrig A, Paxson V, and Savage, S. (2007) An inquiry into the nature and causes of the wealth of internet miscreants. *ACM Conference on Computer and Communications Security*, USA.

Garcia AC, Standlee AI, Bechkoff J, and Cui Y (2009) Ethnographic approaches to the internet and computer-mediated communication. *Journal of Contemporary Ethnography* 38**:** 52-84.

Garton L, Haythornthwaite C. and Wellman B (1997) Studying online social networks.

Journal of ComputerMediated Communication 3.

Hine C (2008) Virtual Ethnography: Modes, Varieties, Affordances. In Fielding N, Lee RM, and Blank G (eds) *The SAGE Handbook of Online Research Methods.* London: Sage.

Holt T, Soles J, and Leslie L (2008) Characterizing malware writers and computer attackers in their own words. *The 3rd International Conference on Information Warfare and Security*, USA: 189.

Holt TJ (2010) Exploring Strategies for Qualitative Criminological and Criminal Justice

Inquiry Using OnLine Data. *Journal of Criminal Justice Education* 21: 466-487. Holt TJ, Smirnova O, Strumsky D, and Kilger M (2014) Advancing Research on Hackers Through Social Network Data. In Marcum CD and Higgins GE (eds) *Social Networking as a Criminal Entreprise.* Boca Raton: Taylor Francis.

Horswell J and Fowler C (2004) Associative evidence–the Locard exchange principle. In Horswell J (ed) *The Practice Of Crime Scene Investigation.* Boca Raton: CRC Press.

King SA (1996) Researching Internet communities: Proposed ethical guidelines for the reporting of results. *The Information Society* 12**:** 119-128.

Kitchin HA (2002) The Tri-Council on cyberspace: Insights, oversights, and extrapolations. In Van Den Hoonaard, WC (ed) *Walking the tightrope: Ethical issues for qualitative researchers.* Toronto: University of Toronto Press.

Krebs B (2014) *Spam Nation: The Inside Story of Organized Cybercrime-from Global Epidemic to Your Front Door,* Naperville, Sourcebooks, Inc.

Krebs VE (2002) Mapping networks of terrorist cells. *Connections* 24**:** 43-52.

Kshetri N (2013) *Cybercrime and cybersecurity in the global south,* New York, Palgrave Macmillan.

Lavorgna A (2015) Organised crime goes online: Realities and challenges. *Journal of Money Laundering Control* 18(2): 153-168.

Lavorgna A and Sergi A (2014) Types of organized crime in Italy. The multifaceted spectrum of Italian criminal associations and their different attitudes in the financial crisis an in the use of Internet technologies. *International Journal of Law, Crime and Justice* 42(1): 16-32

Lee-Treweek G and Linkogle S (eds) (2000) *Danger in the Field: Risk and Ethics in Social Research,* London: Routledge.

Manky D (2013) Cybercrime as a service: a very modern business. *Computer Fraud & Security* 6: 9-13.

Mann D and Sutton M (1998) NETCRIME: More Change in the Organization of Thieving. *British Journal of Criminology* 38**:** 201-229.

Marill JL, Boyko A, Ashenfelder M, and Graham L (2004) Tools and techniques for harvesting the World Wide Web. Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries, USA.

May T (1994) *Crypto Anarchy and Virtual Communities* [Online]. Available:

http://groups.csail.mit.edu/mac/classes/6.805/articles/crypto/cypherpunks/ma y-virtual-comm.html [Accessed 17 March 2014].

Morselli C (2009) *Inside Criminal Networks,* New York, Springer Science+ Business Media.

Motoyama M, Mccoy D, Levchenko K, Savage S., and Voelker GM (2011) An analysis of underground forums. *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, Germany.

Oliveira M (2014) *Canadians spending drastically more time online, comScore study shows* [Online]. Toronto: The Star. Available:

http://www.thestar.com/business/tech_news/2014/11/12/canadians_spending _drastically_more_time_online_comscore_study_shows.html [Accessed 1 April 2015 2015].

Olston C and Najork M (2010) Web crawling. *Foundations and Trends in Information Retrieval* 4**:** 175-246.

Oosthoek A (1978) *The Utilization of Official Crime Data*, Solicitor General Canada, Research Division.

Reid E (1996) Informed consent in the study of on-line communities: a reflection on the effects of computer-mediated social research. *The Information Society* 12**:** 169-174.

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

Rosenberg A (2010) Virtual world research ethics and the private/public distinction. *International Journal of Internet Research Ethics* 3: 23-36.

Sixsmith J and Murray CD (2001) Ethical issues in the documentary data analysis of Internet posts and archives. *Qualitative Health Research* 11**:** 423-432.

Sparrow MK (1991) The application of network analysis to criminal intelligence: An assessment of the prospects. *Social networks* 13**:** 251-274.

Stone-Gross B, Cova M, Cavallaro L, Gilbert B, Szydlowski M, Kemmerer R, Kruegel C, and Vigna G (2009) Your botnet is my botnet: analysis of a botnet takeover. *Proceedings of the 16th ACM conference on Computer and communications security*, USA.

Sveningsson Elm M (2008) How do various notions of privacy influence decisions in qualitative

internet research? In Markham A and Baym N.(eds.) *Internet Inquiry: Conversations About Method.* London: Sage.

the Nature of Groups engaged in Cyber Crime. *International Journal of Cyber Criminology* 8(1): 1-20.

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

Wellman B (2002) Designing the Internet for a networked society. *Communications of the ACM* 45**:** 91-96.

Williams JP and Copes (2005) " ow Edge Are You?" Constructing Authentic Identities and Subcultural Boundaries in a Straightedge Internet Forum. *Symbolic Interaction* 28**:** 67-89.

Zantout B and Haraty R (2011) I2P data communication system. *The Tenth International Conference on Networks*, Netherlands.