22

The Ethics of Web Crawling and Web Scraping in Cybercrime Research: Navigating Issues of Consent, Privacy, and Other Potential Harms Associated with Automated Data Collection

Russell Brewer, Bryce Westlake, Tahlia Hart, and Omar Arauza

Introduction

Over the past three decades, the internet has become an increasingly attractive location for conducting social science research (Askitas & Zimmermann, 2015; Hooley et al., 2012). Two driving forces are the

R. Brewer
School of Social Sciences, University of Adelaide, Adelaide, SA, Australia
e-mail: russell.brewer@adelaide.edu.au

B. Westlake (✉) · O. Arauza
Department of Justice Studies, San Jose State University, San Jose, CA, USA
e-mail: bryce.westlake@sjsu.edu

O. Arauza
e-mail: omar.arauza@sjsu.edu

T. Hart
College of Business Government and Law, Flinders University, Adelaide, SA, Australia
e-mail: tahlia.hart@flinders.edu.au

A. Lavorgna and T. J. Holt (eds.), *Researching Cybercrimes*,
https://doi.org/10.1007/978-3-030-74837-1_22

abundance of quality data available (personal information, communications, videos, images, and other data) and the ease at which such data can be accessed. As the volume of data available online increases, researchers have turned to automated data collection tools. These include web crawlers (a process also known as mirroring), which systematically browses (i.e., crawls) and indexes various web pages (Olston & Najork, 2010) and web scrapers, which access and download large volumes of data from websites based on user-defined criteria (Thomas & Mathur, 2019)—see, for instance, Chapters 3, 8, 10, and 11. In recent years, the number of studies that have used software integrating both web crawlers and web scrapers (automated collection software, hereafter) has increased, as is the degree of sophistication and creative means by which these technologies have been deployed (see Chapters 8 and 10 for an overview of these developments). The rapid rise in their use has meant that guidelines for their ethical operation have been slow to develop and adapt.

The deployment of automated software by researchers (in criminology and beyond) has given rise to debates over ethical concerns surrounding informed consent, privacy, and other risks and potential harms. These concerns arise because of the automated nature of the data collection process, including decisions made by programmers and researchers, as well as inconsistent approaches taken by institutional human research ethics committees. While some scholars have made progress toward identifying and addressing said ethical dilemmas in psychiatry (Sidhu & Srinivasraghavan, 2016; Silva et al., 2017), psychology (Harlow & Oswald, 2016; Landers et al., 2016), and social work (Bent-Goodley, 2007; McAuliffe, 2005; Millstein, 2000), criminology has been slow to identify, acknowledge, and respond to these issues, as well as tackle more discipline-specific concerns. Some early pioneering criminological work (e.g., Décary-Hétu & Aldridge, 2015; Martin & Christin, 2016; Pastrana et al., 2018) has identified and acknowledged some of the ethical dilemmas facing specific key online research environments (such as cryptomarkets and web forums), but have not fully considered other criminological contexts. While this work has been instrumental in setting the scene, we suggest that taking a holistic view of the criminological domains within which automated collection software operates can

provide a fuller understanding of the suite of ethical challenges. Identifying and addressing said challenges can serve to guide future applied research endeavors.

In this chapter, we aim to raise awareness among criminological researchers about the ethical challenges associated with automated collection software, which will be accomplished in two parts. First, we detail the extent and contexts within which automated software have been deployed within the field of criminology, which are useful in drawing out the unique contexts and ethical challenges facing the discipline. Notably, we demonstrate that the data collected by researchers often do not involve human subjects, or when they do, tend to involve experiences of criminality and/or victimization that ultimately require specific and due consideration. Second, we chronicle and critically engage with the ethical challenges confronting criminological researchers utilizing said software. In doing so, we argue that such data collection practices need not be unethical, provided special care is taken by the researcher to acknowledge and explicitly address the complexities surrounding consent, privacy, and a myriad of other potential harms (to subjects, websites, and researchers). We conclude by drawing together the key points emerging from the discussion to offer practical recommendations that we anticipate will provide researchers a path forward when navigating this burgeoning, yet challenging, terrain.

The Use of Automated Collection Software in Criminology

Criminologists have used automated software to collect data from myriad sources, emanating from both the surface and deep/dark web. This has included personal websites and blogs, social media, video streaming platforms, web forums, chat rooms, online marketplaces (both licit and illicit), and peer-to-peer networks. The data collected can be broadly classified into four types—media files, goods and services bought and sold online, digital communications regarding the commission of crimes, and experiences of victimization—and have been used to study a vast array of criminological phenomena.

First, the internet has transformed the way that *media files* are distributed and consumed. In some instances, the media is being distributed illegally (e.g., copyright infringement) or contains graphic content (e.g., child sexual abuse material [CSAM]). This has led criminologists to use automated software to investigate topics such as the impact of piracy on book sales (Hardy et al., 2014), the distribution of pirated (copyrighted) content (Décary-Hétu et al., 2014), the validity of anti-piracy tools on YouTube (Jacques et al., 2018), the automated identification of fake news videos (García-Retuerta et al., 2019), and the analysis of CSAM (Fournier et al., 2014; Kusz & Bouchard, 2020; Shavitt & Zilberman, 2013; Westlake et al., 2012, 2017).

Second, the global reach of the internet has facilitated an explosion of digital marketplaces—which has yielded unprecedented information about *goods and services (licit and illicit) that are being bought and sold online*. Researchers have leveraged automated software to find and collect vendor and transaction-based data on the sale of legal and illegal items through Darknet cryptomarkets and on the surface web. This has, for example, included credit cards (Bulakh & Gupta, 2015), drugs and precursor chemicals (Broadhurst et al., 2020; Cunliffe et al., 2017; Demant, Munksgaard, & Houborg, 2018; Demant, Munksgaard, Décary-Hétu, et al., 2018; Frank & Mikhaylov, 2020; Hayes et al., 2018; Paquet-Clouston et al., 2018), protected wildlife (Hansen et al., 2012; Xu et al., 2020), malware (Broadhurst et al., 2018), and other forms of contraband (Barrera et al., 2019; Broadhurst et al., 2020; Décary-Hétu & Quessy-Doré, 2017).

Third, the internet is often used to discuss the *commission of crimes or to incite others to engage in crime*. Criminologists have collected user-based data, including communications between users, to better understand the role of cyberspace in facilitating crime. This has included examining web forums for illicit, radical, sentiment (Mei & Frank, 2015; Scrivens et al., 2019) and violent agendas (Bouchard et al., 2014), collecting social media posts to study religious bigotry (Gata & Bayhaqy, 2020; Ozalp et al., 2020) predict real-world threats and security requirements (Subramaniyaswamy et al., 2017), and better understand the social mores of offending (Lyu et al., 2020). Other social media for social mores of movie piracy (Lyu et al., 2020). Automated collection

software has also been used to explore the sharing of information on how to commit cyberattacks (Crosignani et al., 2020; Décary-Hétu & Dupont, 2013; Macdonald et al., 2015; Pastrana et al., 2018). Finally, software has been used to monitor malicious websites on the dark web (Pannu et al., 2018) and gather intelligence on organized crime's human trafficking recruitment (McAlister, 2015).

Fourth, the internet provides an opportunity for people to discuss their witnessing, or opinion, of crime and share their *experiences of victimization*. This information is important for understanding previous crime (Keyvanpour et al., 2011), informing the public, and preventing future crime. To explore this, criminologists have used automated software to collect, primarily, text-based descriptions on social media. From this, they have explored the discourse around media from cellular phones, dash cams, and law enforcement body cams (Pitman et al., 2019), sharing of potential scams or threats (Gorro et al., 2017), and experiences of, and responses to, crime, to predict future online bullying, harassment, and scams (Abbass et al., 2020).

Considered together, the preceding discussion illustrates the diversity of criminological studies leveraging automated collection software, both in terms of the data sources used, and the social phenomena explored. Critical appraisal of these studies, according to data type, reveals myriad ethical challenges, for which researchers must consider before data collection should commence. These are explained in further detail below.

Navigating the Ethical Minefield of Automated Data Collection in Criminology

Researchers within criminology, and indeed across other disciplines, are prone to using automated software for data collection purposes without due consideration, given there appears to be no clear ethical guidelines or regulations governing their use (Capriello & Rossi, 2013; Martin & Christin, 2016; Thelwall & Stuart, 2006). As a result, scholars have continually called for the development of consistent ethical guidelines

(see further, Alim, 2013; Chiauzzi & Wicks, 2019; Gold & Latonero, 2018; Thelwall & Stuart, 2006). Without such guidance, researchers are expected to apply pre-existing institutional ethical (and legal) frameworks, which often fail to consider both technological advancements (Gold & Latonero, 2018; Thelwall & Stuart, 2006) and unique criminological contexts (Décary-Hétu & Aldridge, 2015). Further, responses by ethics committees may be influenced by their individual members' expertise and training rather than uniformed adherence to guidelines, whether they are directly relevant or not (McCann, 2016). These issues may raise concerns when proposing to use specialized technological tools in unique or novel settings and can result in the imposition of unnecessary or inappropriate restrictions that make the research unfeasible (Martin & Christin, 2016). Criminological researchers should not be deterred from using automated collection software for research, but do need to be cautious when approaching this method of data collection and also be informed about, and mitigate against, any potential risks or harms. These can overlap, but also diverge, depending on the types of data being collected. The following discussion engages with the principal issues arising from the criminological literature canvassed above and navigates the researcher through the ethical process, with particular emphasis on such emergent issues as consent, privacy, and potential harms that may arise.

The issue of informed consent (see also Chapters 16–20) is debated among researchers leveraging these technologies and is an issue arising from the software eliminating the need for researcher-subject interaction (see further, Décary-Hétu & Aldridge, 2015; Martin & Christin, 2016; Tsatsou, 2014). In offline research contexts, researchers are typically expected to obtain consent from human subjects to collect and analyze their data. However, since automated collection software extracts data that have previously been published online, and at a large scale, this process becomes problematic—regardless of the types of criminological data sourced by researchers. Without the fundamental interaction, human subjects associated with scraped data would not be able to consent. This bears out in practice as Alim (2014) found only 47% ($n = 64$) of surveyed multi-disciplinary researchers acquired consent for

scraped user profile data. A closer examination of criminological articles reviewed in this chapter revealed that very few researchers explicitly addressed the issue of informed consent, or even flagged other ethical considerations associated with the data collection process. This may be due to ambiguity around whether data being collected by the automated software has been derived from a human subject. For example, some scholars have debated whether data automatically extracted from online sources (e.g., prices, reputation data extracted from digital marketplaces) should meet accepted definitions of human subjects (see Alim, 2013; Gold & Latonero, 2018; Solberg, 2010 for further treatment of these arguments). Moreover, the issue of "ownership" over data appearing online is complex, with Martin and Christin (2016) arguing that obtaining informed consent from one group to participate in the research (e.g., webmasters), does not extend to other parties who may also be entitled (e.g., a user posting to a forum about their experiences being victimized). Accordingly, the apparent lack of engagement by criminological researchers observed here, which are consistent with trends reported by Pastrana et al. (2018), may point to tacit acknowledgment of arguments in the field that it is appropriate to waive informed consent under certain conditions (Martin & Christin, 2016). These circumstances are complex and interwoven and are elaborated upon below.

Informed consent can be waived in instances where the anticipated benefits of the research outweigh any potential risks associated with the research (these risks are canvassed in detail below). Criminological studies, in particular, can produce considerable public benefit by providing crucial information that enhances understandings of the motivations driving certain criminal behaviors, such as the commission of hacking (Décary-Hétu & Dupont, 2013) and the inciting of extremist sentiment (Scrivens et al., 2017). Elsewhere, such studies have been used to identify key trends in the distribution of illicit drugs in digital marketplaces (Martin et al., 2018a, 2018b; Moeller et al., 2020) and CSAM (Joffres et al., 2011; Westlake & Frank, 2016; Westlake et al., 2011). To investigate such areas, there is often a need for stealthy crawling to avoid interfering with the natural behavior of subjects (see further, Soska & Christin, 2015). Whist studies involving limited disclosure or deception

are often discouraged by institutional ethics committees, criminology has a long history of covert research, which have produced measurable benefits to public policy (Calvey, 2013; Décary-Hétu & Aldridge, 2015). As such, a criminological researcher looking to embark down such a path should be able to clearly articulate these benefits, while also being able to mitigate potential risks, particularly as they might relate to different data types.

Researchers seeking to waive consent need to ensure that their research activities will present a negligible or low risk of harm. This includes risk to the research subject, if one can reasonably be determined (e.g., users of a web forum, sellers/buyers on an e-commerce platform, those depicted within media files), as well as others who might be adversely affected by the research, such as website administrators and the website itself. Accordingly, researchers are typically required to consider and protect subject privacy, particularly when it pertains to the collection and storage of data, as well as in the reporting of results. However, ascertaining precisely what information appearing online should reasonably be considered in the "private" versus "public" domain is not necessarily straightforward and has attracted considerable scholarly debate (see further, Alim, 2014; Décary-Hétu & Aldridge, 2015; Solberg, 2010; Wilson et al., 2012). That is, there are various "public" fora online where a user has posted information online that is freely available for broader public consumption and therefore does not necessarily attract an expectation of privacy (e.g., Twitter). There are also domains that involve clear and identifiable "private" exchanges between individuals (e.g., direct messages). When it comes to collecting data online however, the separation between these two domains can quickly become blurred. For example, some digital marketplaces and web forums are not entirely "public," insofar as they require a user to first register as a member before access is granted—although registration may otherwise be free and open to anyone, without need for the researcher to compromise the website (Christin, 2013). Elsewhere, scholars have also drawn distinctions between online communities that have large memberships versus those that are only visible to a few members, arguing that the latter may assert a higher expectation of privacy (Martin & Christin,

2016). Accordingly, a researcher looking to employ automated collection software must carefully consider these contexts in order to draw conclusions about privacy in the online setting they wish to research. In guiding such decisions, some scholars have argued that assumptions about privacy should reflect and coincide with the norms of the community under study (Décary-Hétu & Aldridge, 2015; Martin & Christin, 2016; Pastrana et al., 2018; Turk et al., 2020).

Taking such a considered approach to privacy—before, but also during and after data collection has occurred—is vital to minimize any potential harms to subjects. The consolidation of a significant quantum of personal data has the potential to uncover associations or reveal subjects through collection of various information across different platforms. This can include data obtained through forums or social media, such as a list of contacts (or friends), the correspondence between parties, photographs, videos, "tags" and other metadata (Alim, 2013; Gold & Latonero, 2018). In circumstances where criminological researchers collect data pertaining to individuals who are desirable to law enforcement (e.g., users discussing the commission of a crime on web forums, those selling illicit items on cryptomarkets, and people sharing illicit media files), pressure could be applied to the researcher to provide such data. This could facilitate the arrest or prosecution of individuals (Israel, 2004; Martin & Christin, 2016) as well as other harms through instances of internet vigilantism (Chang, 2018). Elsewhere, the collection of personal data also presents a risk for victimization of new crimes—where any such information made publicly available could be used for spamming, phishing, and/or identity theft purposes (Giles et al., 2010; Menczer, 2011; Thelwall & Stuart, 2006). Finally, where scraped data could be analyzed, published, and subsequently read by and cause trauma to the subject, there is potential for re-victimization (Gueta et al., 2020). For example, in the examination of commonly experienced crimes and victimization on social media(Gorro et al., 2017), reproduction of profiles, photos, posts, and stories may be easily encountered by the victim or other individuals known to the victim. While the likelihood of such activities occurring is low given the expansive and global nature of web-based activities, this risk is nevertheless real and requires that the researcher approach with caution.

Given the potentially sensitive nature of personal data being scraped (regardless of data type), criminological researchers seeking to employ such methods must take several steps to minimize the potential for harm against subjects (e.g., those buying/selling items on digital marketplaces, posting comments or other media online, and even those persons contained within media files). This should be accomplished through a process of anonymizing individual outputs, avoiding analysis of identifiable information (Magdy et al., 2017; Xu et al., 2020), and securing the storage and transmission of any sensitive data (Tsatsou, 2014). This includes not only a subject's "user name" (where applicable), but also any other personal data (e.g., verbatim quotes, extracted biometric data) that might infer information back to a particular subject or even the source website (Décary-Hétu & Aldridge, 2015; Fussell, 2019). Practically, when researchers may not be able to strip all personal identifiers from data without compromising its useability (Israel, 2004), data should be reported on an aggregated level (Alim, 2014; Bouwman et al., 2013). This may prove particularly problematic for data sharing among researchers and in many cases will prohibit such practices. Furthermore, researchers must exercise care beyond the data collection process, and be attentive to data security, particularly as it pertains to data storage and processing procedures (Chiauzzi & Wicks, 2019; Magdy et al., 2017; Xu et al. 2020). To mitigate any potential harm, researchers are advised to maximize confidentiality and implement robust security safeguards, which include both strict access controls and data encryption (Alim, 2014; Bouwman et al., 2013; Gold & Latonero, 2018; Tsatsou, 2014). Finally, researchers need also be aware of any reporting requirements (e.g., being a mandatory reporter in a particular jurisdiction) and be mindful of those requirements prior to, during, and after data collection.

In addition to the potential harm against subjects, researchers seeking to collect data through the use of automated collection software must also be aware of the potential financial and technological harms to website administrators and the platforms themselves, and incorporate measures to mitigate against them. For example, to avoid overloading a server and preventing legitimate traffic from accessing a website (i.e., mimicking a DDoS attack, see further, Thewal & Stuart, 2006),

or abusing the TOR network (see further, Christin, 2013), automated collection software should distribute their requests to servers in a measured way (Menczer, 2011). This could be accomplished by mimicking a single human accessing a website one page at a time. This will also limit potential "spider traps" for researchers (i.e., the web scraper becomes trapped in an infinite loop), which duplicate data and waste bandwidth (Menczer, 2011; Thelwall & Stuart, 2006). To combat against such risks, some websites actively employ tactics to set limits on the ways that automated collection software can function on a website, such as using CAPTCHA services (Pastrana et al., 2018) or articulating unenforceable advisory protocols that specify parameters around what information contained can (and cannot) be collected via automated collection software, through "Robots Exclusion Protocols" (robots.txt). This necessitates that the researcher(s) develop and implement internal protocols which dictate such aspects relating to automated collection software downloading, downloading priorities, server request rates, re-visiting, CAPTCHA bypass, and scraper politeness (Capriello & Rossi, 2013; Menczer, 2011). However, we agree with other scholars (e.g., Hand, 2018; Pastrana et al., 2018) who suggest that there may be situations where, after review, it might be justifiable to ignore such protocols—particularly when the benefits of the research outweigh the potential harms. As such, carefully considering the context within which a researcher encounters such protocols is fundamentally important in determining a path forward.

Criminological researchers employing automated collection software also need to be aware of, and mitigate against, unique risks to themselves (see also Chapters 23 and 24). Given the domain of study, data being collected could be both illegal and cause the researcher trauma. For example, the collection and analysis of media files containing graphic content, such as child sexual abuse (Latapy et al., 2013; Westlake & Bouchard, 2016a, 2016b), could cause psychological harm and open researchers up to criminal charges for accessing and possession. Elsewhere, the collection and analysis of textual depictions of heinous crimes or serious victimization could be distressing to the researcher(s) (Pitman et al., 2019; Xin & Cai, 2018). Prior to engaging in research

of this nature, researchers need to mitigate against potential psychological harm by developing a study protocol. This would likely include taking care to separate personal electronic devices from data collection and analysis devices, requiring counseling for research team members, and determining how and when data will be analyzed within the department (e.g., office) to minimize accidental exposure to colleagues and students. To mitigate against potential dismissal or arrest, researchers should consult with ethics committees, departmental supervisors, and law enforcement about their research plan. In addition, there are situations where researchers may want to be discreet in their deployment of web scrapers. For example, researchers would be advised to not announce their intention to scrape data to the cryptomarkets, as doing so may impact the integrity of the data (e.g., changing buying and selling habits and biasing results), but also potentially put the researcher at risk, through possible reprisal (personal abuse, threats, physical or cyber-attacks from site users) (Décary-Hétu & Aldridge, 2015; Holt et al., 2014; Martin & Christin, 2016). Similar risks are also present for data collected from other sources—including from websites where subjects correspond about the commission of crimes or about their personal victimization. Accordingly, researchers should be mindful of the context and circumstances before implementing such practices that disclose their information. While such practices as publishing a user-agent HTTP header to inform website administrators of the scraper's nature (i.e., by providing the scraper's name, version, and links to further information about the research project) have merit in some circumstances (see further, Menczer, 2011), the disclosure of such information could potentially put the research at risk and should be carefully considered.

It is also important to flag that researchers who engage in automated collection may, in certain situations, be subject to potential litigation, particularly in cases where the robots.txt protocol is ignored/misinterpreted, or the website's terms of service (TOS) forbid the harvesting of data (Alim, 2014; Giles et al., 2010; Gold & Latonero, 2018; Sun et al., 2010). Some criminologists have weighed in on this debate and suggested that TOS, particularly those appearing on illicit websites (e.g., criminal marketplaces), are not legally enforceable (see further, Martin & Christin, 2016). Elsewhere, scholars have debated

whether the automated extraction of data that may be subject to copyright could present further risk of litigation (O'Reilly, 2007; Stokes, 2019). Given the multijurisdictional nature of legal proceedings, it is outside the scope of this chapter to provide researchers with resolute guidelines. In addition, it is difficult to provide specific advice to follow as legal aspects of data automatically collected are unclear, inconsistent, and difficult to interpret (Gold & Latonero, 2018; Landers et al., 2016). As a result, researchers should seek legal advice specific to their jurisdiction, as well as the context surrounding the website(s) and research endeavor(s) prior to data collection. However, from an ethical standpoint, scholars have argued that it can be permissible to breach TOS for research purposes, providing that the benefits of the research outweigh any potential harms (Freelon, 2018; Martin & Christin, 2016).

Beyond acknowledging and addressing any risks inherent in the research enterprise, a waiver of consent typically requires that attempts to do so would be impractical. Indeed, scholars have noted that obtaining informed consent using automated collection software is not only impractical, but often impossible (Tsatsou, 2014). Automated collection software typically seeks to obtain data for a full population (e.g., capturing all available data on a digital marketplace or web forum) as opposed to a more targeted sampling process. As such, the software extracts data for all users—whose true identities may be masked by avatars and pseudonyms, and who may be active on the website or long inactive. As such, researchers will typically not be in a position to obtain or collect reliable contact information for subjects under study before (or even after) the research is undertaken (Décary-Hétu & Aldridge, 2015).

This section has demonstrated that the use of automated collection software in criminological contexts is potentially rife with ethical challenges. Researchers need to thoroughly explore the ramifications of informed consent, particularly as it pertains to the type of data being collected and analyzed. Doing so requires a robust understanding of how subject privacy could be impacted by the research, and what protections will need to be implemented. Likewise, the investigation of criminal activity means that researchers need to fully understand and mitigate against the risks and potential harms, even if done unwittingly, that the

research could pose to subjects, websites, and themselves. If due consideration is afforded in the ways we have outlined above, we argue that it is possible to use automated collection software in ethical and conscientious manners for criminological study.

Conclusion

Criminologists have successfully deployed automated collection software to identify and extract various types of data across numerous sources, to better understand phenomena such as terrorism, CSAM, illicit drug distribution, and hacking. Such data collection strategies enable innovative studies that afford global recruitment possibilities (Tsatsou, 2014), and can cluster data at an efficient speed and low cost (Tavani, 1999). At the same time, they can overcome deficiencies commonly associated with more traditional research methods, including low survey responses (Gök et al., 2015) and the need for researcher involvement and training (Gök et al., 2015; Landers et al., 2016). This chapter has shown that despite a proliferation in the use of such technologies, criminology has been slow to identify, acknowledge and respond to the unique ethical challenges confronting their use, and tackle discipline-specific concerns. This chapter elucidated and critically engaged with these ethical challenges, and in doing so, argued that such data collection practices need not be unethical, providing that special care is taken to explicitly address and justify matters pertaining to consent, and mitigation against risks and potential harms (to subjects, websites, and researchers).

While the use of automated collection software presents numerous ethical challenges that the researcher must consider, we close by stressing that it is not our intention to discourage criminologists from employing such data collection techniques. Rather, our aim in this chapter is to encourage researchers to acknowledge and engage with these tools in an ethical way and thus open the door to novel and fruitful means of better understanding various crime problems. It is our hope that the discussion and recommendations presented offer a useful path forward and will enhance consistency in, and understanding of, ethical practices.

References

Abbass, Z., Ali, Z., Ali, M., Akbar, B., & Saleem, A. (2020). A framework to predict social crime through Twitter tweets by using machine learning. *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, 363–368.

Alim, S. (2013). Automated data extraction from online social network profiles: Unique ethical challenges for researchers. *International Journal of Virtual Communities and Social Networking (IJVCSN), 5*(4), 24–42.

Alim, S. (2014). An initial exploration of ethical research practices regarding automated data extraction from online social media user profiles. *First Monday, 19*(7).

Askitas, N., & Zimmermann, K. F. (2015). The Internet as a data source for advancement in social sciences. *International Journal of Manpower, 36*(1), 2–12.

Barrera, V., Malm, A., Décary-Hétu, D., & Munksgaard, R. (2019). Size and scope of the tobacco trade on the darkweb. *Global Crime, 20*(1), 26–44.

Bent-Goodley, T. B. (2007). Teaching social work students to resolve ethical dilemmas in domestic violence. *Journal of Teaching in Social Work, 27*(1–2), 73–88.

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

Bouwman, H., de Reuver, M., Heerschap, N., & Verkasalo, H. (2013). Opportunities and problems with automated data collection via smartphones. *Mobile Media & Communication, 1*(1), 63–68.

Bulakh, V., & Gupta, M. (2015). Characterizing credit card black markets on the web. *Proceedings of the 24th International Conference on World Wide Web,* 1435–1440.

Broadhurst, R., Ball, M., & Jiang, C. (2020). Availability of COVID-19 related products on Tor darknet markets. *Statistical Bulletin, no. 24*. Canberra: Australian Institute of Criminology.

Broadhurst, R., Ball, M., & Trivedi, H. (2020). Fentanyl availability on darknet markets. *Trends & issues in crime and criminal justice, no. 590*. Canberra: Australian Institute of Criminology.

Broadhurst, R., Lord, D., Maxim, D., Woodford-Smith, H., Johnston, C., Chung, H.W., et al. (2018). Malware trends on Darknet crypto-markets: Research review. *ANU Cybercrime Observatory*. Canberra.

Calvey, D. (2013). Covert ethnography in criminology: A submerged yet creative tradition. *Current Issues in Criminal Justice, 25*(1), 541–550.

Capriello, A., & Rossi, P. (2013). Spidering scripts for opinion monitoring. In H. Rahman & I. Ramos (Eds.), *Ethical data mining applications for socio-economic development*. IGI Global.

Chang, L. Y. C. (2018). Internet vigilantism co-production of security and compliance in the digital age. In Brewer R. (Ed.), *criminal justice and regulation revisited: Essays in honour of Peter Grabosky*. Routledge.

Chiauzzi, E., & Wicks, P. (2019). Digital trespass: Ethical and terms-of-use violations by researchers accessing data from an online patient community. *Journal of Medical Internet Research, 21*(2).

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.

Crosignani, M., Macchiavelli, M., & Silva, A. F. (2020). Pirates without borders: The propagation of cyberattacks through firms' supply chains. *SSRN Electronic Journal*.

Cunliffe, J., Martin, J., Décary-Hétu, D., & Aldridge, J. (2017). An island apart? Risks and prices in the Australian cryptomarket drug trade. *The International Journal of Drug Policy, 50*, 64–73.

Décary-Hétu, D., & Aldridge, J. (2015). Sifting through the net: Monitoring of online offenders by researchers. *European Review of Organised Crime, 2*(2), 122–141.

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

Décary-Hétu, D., & Quessy-Doré, O. (2017). Are repeat buyers in cryptomarkets loyal customers? Repeat business between dyads of cryptomarket vendors and users. *American Behavioral Scientist, 61*(11), 1341–1357.

Décary-Hétu, D., Dupont, B., & Fortin, F. (2014). Policing the hackers by hacking them: Studying online deviants in irc chat rooms. In A. J. Masys (Ed.), *Networks and network analysis for defence and security*. Springer.

Demant, J., Munksgaard, R., & Houborg, E. (2018). Personal use, social supply or redistribution? Cryptomarket demand on Silk Road 2 and Agora. *Trends in Organized Crime, 21*(1), 42–61.

Demant, J., Munksgaard, R., Décary-Hétu, D., & Aldridge, J. (2018). Going local on a global platform: A critical analysis of the transformative potential of cryptomarkets for organized illicit drug crime. *International Criminal Justice Review, 28*(3), 255–274.

Fournier, R., Cholez, T., Latapy, M., Chrisment, I., Magnien, C., Festor, O., & Daniloff, I. (2014). Comparing pedophile activity in different P2P systems. *Social Sciences, 3*(3), 314–325.

Frank, R., & Mikhaylov, A. (2020). Beyond the 'Silk Road': Assessing illicit drug marketplaces on the public web. In M. A. Tayebi., U. Glässer, & D. B. Skillicorn (Eds.), *Open source intelligence and cyber crime*. Springer.

Freelon, D. (2018). Computational research in the post-API Age. *Political Communication, 35*(4), 665–668.

Fussell, S. (2019). You no longer own your face. *The Atlantic*. Available at: https://www.theatlantic.com/technology/archive/2019/06/universities-record-students-campuses-research/592537/.

García-Retuerta, D., Bartolomé, Á., Chamoso, P., & Corchado, J. M. (2019). Counter-terrorism video analysis using hash-based algorithms. *Algorithms, 12*(5).

Gata, W., & Bayhaqy, A. (2020). Analysis sentiment about islamophobia when Christchurch attack on social media. *Telkomnika, 18*(4), 1819–1827.

Giles, C., Sun, Y., & Councill, I. (2010). Measuring the web crawler ethics. *Proceedings of the 19th International Conference on World Wide Web*, 1101–1102.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics, 102*(1), 653–671.

Gold, Z., & Latonero, M. (2018). Robots welcome? Ethical and legal consideration for web crawling and scraping. *Washington Journal for Law, Technology & Arts, 13*(3), 275–312.

Gorro, K. D., Sabellano, M. J. G., Maderazo, C. V., Ceniza, A. M., & Gorro, K. (2017). Exploring Facebook for sharing crime experiences using selenium and support vector machine. *Proceedings of the 2017 International Conference on Information Technology, 2*18–222.

Gueta, K., Eytan, S., & Yakimov, P. (2020). Between healing and revictimization: The experience of public self-disclosure of sexual assault and its perceived effect on recovery. *Psychology of Violence, 10*(6), 626–637.

Hand, D. J. (2018). Aspects of data ethics in a changing world: Where are we now? *Big Data, 6*(3), 176–190.

Hansen, A. L. S., Li, A., Joly, D., Mekaru, S., & Brownstein, J. S. (2012). Digital surveillance: A novel approach to monitoring the illegal wildlife trade. *PLoS ONE, 7*(12), e51156.

Hardy, W., Krawczyk, M., & Tyrowicz, J. (2014). Internet piracy and book sales: A field experiment. *Faculty of Economic Sciences, University of Warsaw Working Papers, 23*(140), 1–22.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods, 21*(4), 447–457.

Hayes, D. R., Cappa, F., & Cardon, J. (2018). A framework for more effective dark web marketplace investigations. *Information (basel), 9*(8), 186–204.

Holt T. J., Smirnova, O., Strumsky, D., & Kilger, M. (2014). Advancing research on hackers through social network data. In C. D. Marcum & G. E. Higgins (Eds.), *Social networking as a criminal enterprise*. Taylor Francis.

Hooley, T., Marriott, J., & Wellens, J. (2012). *What is online research? Using the Internet for social science research.* Bloomsbury Academic.

Israel, M. (2004). Strictly confidential? Integrity and the disclosure of criminological and socio-legal research. *British Journal of Criminology, 44*(5), 715–740.

Jacques, S., Garstka, K., Hviid, M., & Street, J. (2018). An empirical study of the use of automated anti-piracy systems and their consequences for cultural diversity. *SCRIPT-Ed, 15*(2), 277–312.

Joffres, K., Bouchard, M., Frank, R., & Westlake, B. G. (2011). Strategies to disrupt online child pornography networks. *2011 European Intelligence and Security Informatics Conference,* 163–170. IEEE.

Keyvanpour, M. R., Javideh, M., & Ebrahimi, M. R. (2011). Detecting and investigating crime by means of data mining: A general crime matching framework. *Procedia Computer Science, 3*, 872–880.

Kusz, J., & Bouchard, M. (2020). Nymphet or lolita? A gender analysis of online child pornography websites. *Deviant Behavior, 41*(6), 805–813.

Landers, R., Brusso, R., Cavanaugh, K., & Collmus, A. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods, 21*(4), 475–492.

Latapy, M., Magnien, C., & Fournier, R. (2013). Quantifying paedophile activity in a large P2P system. *Information Processing & Management, 49*(1), 248–263.

Lyu, Y., Xie, J., & Xie, B. (2020). The attitudes of Chinese online users towards movie piracy: A content analysis. In A. Sundqvist, G. Berget, J. Nolin, & K. Skjerdingstad (Eds.), *Sustainable digital communities* (pp. 169–185). Springer.

Macdonald, M., Frank, R., Mei, J., & Monk, B. (2015). Identifying digital threats in a hacker web forum. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 926–933.

Magdy, W., Elkhatib, Y., Tyson, G., Joglekar, S., Sastry, N. (2017). Fake it till you make it: Fishing for catfishes. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 497–504.

Martin, J., & Christin, N. (2016). Ethics in cryptomarket research. *International Journal of Drug Policy, 35*, 84–91.

Martin, J., Cunliffe, J., Décary-Hétu, D., & Aldridge, J. (2018a). Effect of restricting the legal supply of prescription opioids on buying through online illicit marketplaces: Interrupted time series analysis. *British Medical Journal, 361*, 1–7.

Martin, J., Cunliffe, J. D., Décary-Hétu, D., & Aldridge, J. (2018b). The international darknet drugs trade-a regional analysis of cryptomarkets. *Australasian Policing, 10*(3), 25–29.

McAlister, R. (2015). Webscraping as an investigation tool to identify potential human trafficking operations in Romania. *Proceedings of the ACM Web Science Conference,* 1–2.

McAuliffe, D. (2005). I'm still standing: Impacts and consequences of ethical dilemmas for social workers in direct practice. *Journal of Social Work Values and Ethics, 2*(1), 1–10.

McCann, M. (2016). The smartphones study: An analysis of disciplinary differences in research ethics committee responses to phone app-based automated data collection. *European Journal of Public Health*, *26* (suppl. 1).

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

Menczer, F. (2011). Web crawling. In B. Liu (Ed.), *Web data mining: Exploring hyperlinks, contents, and usage data, 311 Data-Centric Systems and Applications* (pp. 311–362). Springer.

Millstein, K. (2000). Confidentiality in direct social-work practice: Inevitable challenges and ethical dilemmas. *Families in Society, 81*(3), 270–282.

Moeller, K., Munksgaard, R., & Demant, J. (2020). Illicit drug prices and quantity discounts: A comparison between a cryptomarket, social media, and police data. *The International Journal of Drug Policy* (online first).

Olston, C., & Najork, M. (2010). Web crawling. *Foundations and Trends in Information Retrieval, 4*(3), 175–246.

O'Reilly, S. (2007). Nominative fair use and Internet aggregators: Copyright and trademark challenges posed by bots, web crawlers and screen-scraping technologies. *Loyola Consumer Law Review, 19*(3), 273–288.

Ozalp, S., Williams, M. L., Burnap, P., Liu, H., & Mostafa, M. (2020). Antisemitism on Twitter: Collective efficacy and the role of community organisations in challenging online hate speech. *Social Media + Society, 6*(2), 1–20.

Pannu, M., Kay, I., & Harris, D. (2018). Using dark web crawler to uncover suspicious and malicious websites. *International Conference on Applied Human Factors and Ergonomics* (pp. 108–115). Springer.

Paquet-Clouston, M., Décary-Hétu, D., & Morselli, C. (2018). Assessing market competition and vendors' size and scope on AlphaBay. *International Journal of Drug Policy, 54*, 87–98.

Pastrana, S., Thomas, D. R., Hutchings, A., & Clayton, R. (2018). Crimebb: Enabling cybercrime research on underground forums at scale. *Proceedings of the 2018 World Wide Web Conference*, 1845–1854.

Pitman, B., Ralph, A. M., Camacho, J., & Monk-Turner, E. (2019). Social media users' interpretations of the Sandra Bland arrest video. *Race and Justice, 9*(4), 479–497.

Scrivens, R., Davies, G., & Frank, R. (2017). Searching for signs of extremism on the web: An introduction to Sentiment-based Identification of Radical Authors. *Behavioral Sciences of Terrorism and Political Aggression, 10*(1), 39–59.

Scrivens, R., Gaudette, T., Davies, G., & Frank, R. (2019). Searching for extremist content online using the dark crawler and sentiment analysis. In M. Deflem & D. M. D Silva (Eds.), *Methods of criminology and criminal justice research (Sociology of Crime, Law and Deviance).* Emerald Publishing Limited.

Shavitt, Y., & Zilberman, N. (2013). On the presence of child sex abuse in BitTorrent networks. *IEEE Internet Computing, 17*(3), 60–66.

Sidhu, N., & Srinivasraghavan, J. (2016). Ethics and medical practice: Why psychiatry is unique. *Indian Journal of Psychiatry, 58*(6), 199–202.

Silva, E., Till, A., & Adshead, G. (2017). Ethical dilemmas in psychiatry: When teams disagree. *Bjpsych Advances, 23*(4), 231–239.

Solberg, L. B. (2010). Data mining on Facebook: A free space for researchers or an IRB nightmare? *University of Illinois Journal of Law, Technology & Policy, 2*, 311–343.

Soska, K., & Christin, N. (2015). Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. *In USENIX Security Symposium (USENIX Security),* 33–48.

Stokes, S. (2019). *Digital copyright: Law and practice*. Hart Publishing.

Sun, Y., Councill, I. G., & Giles, C. L. (2010). The ethicality of web crawlers. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 1,* 668–675.

Subramaniyaswamy, V., Logesh, R., Abejith, M., Umasankar, S., & Umamakeswari, A. (2017). Sentiment analysis of tweets for estimating criticality and security of events. *Journal of Organizational and End User Computing, 29*(4), 51–71.

Tavani, H. T. (1999). Informational privacy, data mining, and the Internet. *Ethics and Information Technology, 1*(2), 137–145.

Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology, 57*(13), 1771–1779.

Thomas, D. M., & Mathur, S. (2019). Data analysis by web scraping using python. *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, 450–454.

Tsatsou, P. (2014). Research and the Internet: Fast-growing Internet research. In P. Tsatsou (Ed.), *Internet studies: Past, present and future directions*. Ashgate Publishing Ltd.

Turk, K., Pastrana, S., & Collier, B. (2020). A tight scrape: Methodological approaches to cybercrime research data collection in adversarial environments. *Workshop on Actors in Cybercrime Operations*, 428–437.

Westlake, B. G., & Bouchard, M. (2016a). Criminal careers in cyberspace: Examining website failure within child exploitation networks. *Justice Quarterly, 33*(7), 1154–1181.

Westlake, B. G., & Bouchard, M. (2016b). Liking and hyperlinking: Examining reciprocity and diversity in online child exploitation network communities. *Social Science Research, 59*, 23–36.

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

Westlake, B. G., Bouchard, M., & Frank, R. (2012). Comparing methods for detecting child exploitation content online. *European Intelligence and Security Informatics Conference,* 156–163.

Westlake, B. G., Bouchard, M., & Frank, R. (2017). Assessing the validity of automated webcrawlers as data collection tools to investigate online child sexual exploitation. *Sexual Abuse, 29*(7), 685–708.

Westlake, B. G., & Frank, R. (2016). Seeing the forest through the trees: Identifying key players in online child sexual exploitation distribution networks. In T. Holt (Ed.), *Cybercrime through an interdisciplinary lens.* New York: Routledge.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science, 7*(3), 203–220.

Xin, Y., & Cai, T. (2018). Child trafficking in China: Evidence from sentencing documents. *International Journal of Population Studies, 4*(2), 1–10.

Xu, Q., Cai, M., & Mackey, T. K. (2020). The illegal wildlife digital market: An analysis of Chinese wildlife marketing and sale on Facebook. *Environmental Conservation, 47*(3), 206–212.