USING DIGITAL OPEN SOURCE AND CROWDSOURCED DATA IN STUDIES OF DEVIANCE AND CRIME

R.V. Gundur, Flinders University
Mark Berry, University of Bournemouth
Dean Taodang, Flinders University

Gundur, R.V., Mark Berry, and Dean Taodang. In Press. "Using digital open source and crowdsourced data in studies of deviance and crime." In *Researching Cybercrimes: Methodologies, Ethics, and Critical Approaches*, edited by Anita Lavorgna and Thomas Holt. London: Palgrave.

Abstract

As the internet has become cheaper, faster, and more widely used, the amount of data generated by people has increased exponentially. Much of the data is provided by users' activities, often mundane tasks, like making purchases, engaging in exercise routines, and consuming streaming content. In some cases, these tasks are leveraged by criminal actors; in others, these tasks include criminal activities. Using data to explore patterns of offending and victimization is critical to understanding crime trends in the digital age. This chapter explores how researchers have used open source data collection techniques and have solicited data using crowdsourcing to develop viable data sets to explore social scientific enquiries. These techniques illustrate that it is increasingly possible to gather and solicit information and analytical help to explore deviance, victimization, social taboos, and behaviors that are often kept out of public view but nonetheless impact society.

Key words: open source, data collection, social science, OSINT, crowdsourcing, digital methods

Data and the Social World

In the twenty-first century, people around the world, even in developing economies and rural communities, have become increasingly connected to the internet (James, 2021). Most human-computer transactions generate data points. Data are generated whenever someone logs onto the internet, clicks on an ad, posts on social media, takes digital photos, calls using voice over IP (VOIP), uses digitally-connected services, or engages with internet-of-things (IOT). Consequently, human beings generate petabytes of data on a daily basis, and this rate of data generation means that more data are created in any given year than any previous year, a trend that is likely to continue (DOMO, 2020). All of these data provide opportunities not only for entrepreneurs, who use data analytics to understand their clients' behavior and preferences, but also for social scientists and practitioners, who study patterns of behavior of people who might be otherwise difficult to reach, including cybercriminals and their victims (An & Kim, 2018).

For the better part of the twenty-first century, social scientists such as criminologists have been using digital resources and methodologies to make their research more efficient and to explore new facets of deviant behavior (Powell et al., 2018; Smith et al., 2017). Digital strategies have allowed criminologists to assess deviance and crime in online and offline environments in terms of recruitment (Gundur, 2019; Wood et al., 2019), communication (Cheng, 2017), data collection (Dragiewicz et al., 2018; Giommoni & Gundur, 2018; Lavorgna & Sugiura, 2020; Lawson & Nesbit, 2013; Lynch, 2018; Poletti & Gray, 2019; Potter, 2017; Ramo & Prochaska, 2012), and criminal innovation with technology (Berry, 2018; Cross & Gillett, 2020; Décary-Hétu & Bérubé, 2018; Gillett, 2018; Moule Jr et al., 2013). Overwhelmingly, these methodologies focus on collecting and analyzing open source data, that is, "information derived from sources and by means openly available to and legally accessible and employable by the public" (Schaurer & Störger, 2013, p. 53). Some of these methodologies employ crowdsourcing to engage the public via the internet for the public's input for a defined problem. By aggregating the collective efforts of many, researchers can collect data and/or solve problems efficiently (Brabham, 2013; Solymosi et al., 2018).

Accordingly, this chapter discusses the collection and application of open source and crowdsourced data for criminological research and how researchers can collect and use these data to expand research capacity. This chapter proceeds as follows. First, it discusses the historic value of open source and crowdsourced data. Then, it describes common open source data collection tools, techniques, and technologies. Next, it discusses the analysis of open source and crowdsourced data. Finally, this chapter explores the potential to marry open source research with crowdsourcing as a means to further expand research capacity.

Why are crowdsourced and open source data valuable?

At their core, crowdsourced and open source data are readily available to and efficiently accessed by anyone. Although these terms were coined in the twentieth century, the concepts predate their coinage. In 1879, James A.H. Murray, the first editor of the *Oxford English Dictionary,* crowdsourced information in his global appeal for help in chasing definitions and etymologies of specific words (Winchester, 2018). An early notable example of open source intelligence (OSINT) is also British. In 1939, the British government realized that secret knowledge was useful knowledge and, accordingly, asked the BBC to monitor the public media communications of foes, which could provide insight, regarding key actors, events, and strategies, without having a human resource embedded on site (Schaurer & Störger, 2013). OSINT continues to be a staple of intelligence, military, and policing communities who value the richness of information that adversaries and targets put into the public domain (Akhgar et al., 2016; Trottier, 2015). Likewise, open source data have long been of value to the academic community (Schaurer & Störger, 2013). Social scientists routinely draw on information from public records and datasets to inform the basis of their knowledge.

For OSINT practitioners and academics, open source data are inexpensive compared to fieldwork and can provide access to spaces that would be impractical or difficult to access personally. Accordingly, intelligence, policing, and academic applications will continue to use open source data indefinitely and will expand its use to reach under-researched communities as members of those communities become more internet connected. Nonetheless, the increasing volume of data means how data can be collected and analyzed at scale will evolve. The digitization of communication has increased the volume of information that is publicly available, thereby requiring, in some cases, automated or technically-advanced techniques of data collection and analysis to engage in efficient OSINT (Hribar et al., 2014).

Besides OSINT applications, there are purposeful compilations of information presented for the public good without the expectation of direct monetary gain. Several examples exist which can aid cybercriminological and cybersecurity research, such as the cataloging of malware and ransomware (Roth, 2020), the documentation of scams in scamwatching forums (e.g., *ScamWarners*, 2021), and the sharing of public data sets (e.g. those provided by the Cambridge Cybercrime Centre (2020) or the CARE lab at Temple University (2020)).

Before proceeding, a quick aside on the ethics of open source data collection and analysis is necessary. The systematic collection and analysis of open source data are now easier than ever. This fact has raised some ethical concerns among institutional review boards (ethics committees), especially those that view the collection and analysis of open source information without the express consent of the posters to be ethically problematic (Gearon & Parsons, 2019; Hribar et al., 2014). However, the de-privatized nature of posting to public or ungated spaces on the internet assumes that posters should be, at a minimum, aware that whatever they post is public information and can be surveilled by state agents (Higgs, 2001; Reidenberg, 2014). The imposition of hurdles, which researchers must overcome to collect and use open source data, is antithetical to the production of knowledge.

Open source information is valuable to researchers because it is, by its very nature, not stolen, not classified at its origin, and not proprietary (except for copyright); it is information that is public and can be legally accessed freely without clandestine tactics. These characteristics are important especially as individuals who are currently underrepresented both in academic and practitioner circles – often as a result of resource limitations – make notable contributions to the understanding of crime and deviance within their communities (Carrington et al., 2018). At the same time, certain skills, techniques, and technologies keep this process possible as the booming volumes of digital data require more efficient and accurate assessments.

Skills, techniques, and tools for open source data collection in a digital age

The digitalization of the social world sometimes causes students, teachers, and researchers to forget that basic, less-technical strategies are often the most effective and that knowledgeable analysts are necessary to make sense of the vast amounts of information that OSINT can potentially collect (Hribar et al., 2014). OSINT predates internet communication technologies both in terms of existence and wide-spread usefulness. Certainly, the digitization of resources has made what used to be manual, labor-intensive processes easier and faster to execute. Thus, much information can be extracted from traditional media sources, such as media broadcasts and periodicals, and from administrative records, which include sentencing comments, and the National Registry of Exonerations, all of which are often cataloged in databases (Bright et al., 2012; Hassan, 2019; Lynch, 2018). The use of focused search terms on databases and search engines for text- and image-based information deployed using Boolean operators (e.g. AND, OR, NOT) allows for large-scale searching of pertinent information and continues to be a cornerstone of threat assessments, government reports, and academic research (Neri & Geraci, 2009; Williams & Blum, 2018).

Nonetheless, the contemporary digital world (and, for that matter, the digital world of the future) offers new data collection and analysis opportunities. Although textual information offered online across various services is often chaotic, unstructured, and vast, it is readily accessible. Non-text data, such as images, videos, geospatial data, and digital forensic data, given the advances in consumer electronics, can be readily collected, shared, and analyzed without special equipment or access. However, advances in technology and data generation will always pose new questions which will present their own, sometimes unforeseeable, difficulties in answering them. Accordingly, the fundamental research strategies that underwrite open source techniques to collect data must be adapted and expanded to investigate the vast volume and diverse types of open source and crowdsourced data. Two notable techniques that help in the collection of this vast data are data scraping and crowdsourcing, both of which have open-source applications and are capable of collecting various types of data.

Data Scraping

Data scraping, which involves using an automated program to harvest data that others have collected or posted to form a data set, is a technique commonly deployed to collect text-based data in digital spaces, such as clear and darkweb websites, forums, and social media accounts (Lynch, 2018; Mitchell, 2018; Turk et al., 2020); additionally, it can be used to collect any machine-readable data, such as geospatial or technical data (Ensari & Kobaş, 2018) – see also Chapters N & N. Data scraping may be achieved via automated scraping programs, many of which are open source or can be coded from open source materials (Lynch, 2018). Data can also be mined using shell scripts, a computer program designed to run in the command-line interpreter (Copeland et al., 2020). Scraping can provide a snapshot of a website at a given point in time, can be used to systematically document a website over time, or can monitor data leaks from a website (Ball et al., 2019; Décary-Hétu & Aldridge, 2015; Turk et al., 2020). While some companies, such as Twitter, provide tools to facilitate the collection of data within them, via their application programming interfaces (APIs), others expressly ban the scraping of their content (Burnap & Williams, 2015). Collection from sites that bar data scraping, by definition, would not result in open source data and may fall afoul of institutional review boards (Martin & Christin, 2016); those considerations, however, may not deter intelligence and law enforcement officials who may not face such constraints (Sampson, 2016).

Data scraping, nonetheless, has been used to collect significant amounts of open source data for academic studies. Its ability to collect large swathes of data, efficiently and quickly, has made it useful in examining several criminogenic problems. Moreover, data scraping has many applications and, when done well, results in the structured collection of data (Décary-Hétu & Aldridge, 2015). For instance, scraped data have provided several insights into illicit markets, which may be otherwise difficult, risky, or time consuming to obtain via fieldwork (de Souza Santos, 2018; Wong, 2015). Studies of illicit markets have used scraped open source law enforcement data and press releases to explore relationships between drug trafficking and serious organized crime (Hughes et al., 2020); job advertisements targeting women in Romania to identify possible human trafficking recruitment (McAlister, 2015); advertisements and listings on darknet marketplaces to illuminate drug pricing (Červený & van Ours, 2019; Frank & Mikhaylov, 2020); advertisements and images on the dark web to illuminate the dark web firearms trade (Copeland et al., 2020), and user posts and comments on carding forums to identify customer dynamics in those forums (Kigerl, 2018).

In addition, scraped data have been used to identify scam and fraud patterns, providing insights beyond victimization surveys or officially collected data, which result in time-delays in terms of reporting new problems and how known problems evolve (Schoepfer & Piquero, 2009). For instance, scraped data from Twitter and Instagram posts have been used to understand the sale of false COVID-19 products (Mackey et al., 2020). Scraped data from geotagged Tweets have identified location spoofing and spoofing strategies undertaken by possible trolls and bots (Zhao & Sui, 2017). And scraped data from forum posts have determined scam posts from advance fee scammers and illuminated the scams' mechanics (Mba et al., 2017).

Researchers have also used scraped social media data to evaluate communication trends and to identify the existence of poorly reported events or phenomena, thereby allowing researchers to engage with settings to which they may not have physical access. Scraped blog data have identified hate groups and their members (Chau & Xu, 2007). Scraped data from Google Trends and Twitter have allowed researchers to explore how the public perceives serious crime (Kostakos, 2018). Scraped Tweets have shown how hate speech spreads and influences audiences on Twitter (Ball et al., 2019; Burnap & Williams, 2015; Ozalp et al., 2020). Scraped data from Twitter and Facebook reports have been used to detect terrorist events in the developing world where there may be less reliable journalistic reporting (Kolajo & Daramola, 2017; Oleji et al., 2008).

Moreover, there are numerous tools that facilitate the bulk collection of various types of machine-readable data. Some of these tools are open source while others are paid or government proprietary solutions that use open source data. Among these tools are Foca, which finds metadata and hidden information in documents; Spiderfoot, which amalgamates information relevant for cyber threat intelligence assessments, such as suspicious IP and e-mail addresses and links to phishing campaigns; and 4NSEEK, an EU-funded tool, used by law enforcement practitioners, that scrapes and compares images of child sexual abuse to identify victims of child sexual exploitation (Al-Nabki et al., 2020; Pastor-Galindo et al., 2020). (N.B. Data, such as images, behind paywalls in illicit marketplaces do not constitute opensource data. However, this application is useful for law enforcement, who will develop digital assets in order to get behind paywalls of such businesses.) Common scraping strategies have also yielded results. These strategies include the scraping of geospatial data, such as location reports of neighborhood disorder, to create maps of crime hotspots or to document the geography of a problem (Solymosi & Bowers, 2018), and the scraping of technical data from the clearweb (Cascavilla et al., 2018) and the darkweb (see, for example: Lewis, n.d.) to reveal flaws in operational security by identifying information leaks despite attempts to keep that information private and secure.

Evidently, data scraping is a versatile and powerful strategy that can be used to collect large quantities of data systematically. It has clear advantages: it offers free or inexpensive research resources to the research community; it shortens the time arcs required to collect and process data, and it provides insights into communities that might be time-consuming or difficult to access. Nonetheless, it has limitations, especially in academic contexts: researchers cannot ethically scrape online properties that explicitly bar the practice in their terms and conditions, and scraping cannot be used with data not freely offered online nor with proprietary datasets to which researchers lack access. Researchers have sometimes overcome these limitations with crowdsourcing strategies.

Crowdsourcing Data and Solutions

Not all information of criminological interest is accessible via publicly available data. Consequently, some researchers have used crowdsourcing to solicit private data or to solicit help in conducting analysis or solving problems from the public (Estellés-Arolas, 2020). Crowdsourcing data or solutions allow researchers and practitioners to have participants opt into contributing data or participating in solving a problem or investigation (Powell et al., 2018). Crowdsourcing initiatives typically have a clear crowdsourcer who initiates the project, which has a clear goal, with an open call to a crowd who must carry out the task (Estellés-Arolas, 2020; Estellés-Arolas & González-Ladrón-de-Guevara, 2012). While crowdsourcing can be open to whomever wants to participate, some crowdsourcing efforts, such as those conducted via online contract labor portals, resemble surveys as they are geared towards specific populations (Behrend et al., 2011; Litman et al., 2017).

Within a criminal justice context, there are crowdsourcing platforms deployed by researchers and practitioners designed to respond to specific initiatives, namely collecting data, analyzing data, or solving problems (Estellés-Arolas, 2020). Data collection applications include crime reporting mechanisms, like the Australian Competition & Consumer Commission's (n.d.) Scamwatch and the UK's ActionFraud's (n.d.) online scam reporting tool. These mechanisms seek to collect reports of scams in the immediate aftermath of victimization, rather than relying on recollection as traditional victimization surveys do.

In addition, there are platforms that allow users to submit data. For example, *FixMyStreet* (Society Works, n.d.) allows UK-based users to report the geospatial data of potholes, broken streetlights, and other blights in their area. *Price of Weed* (2019), a website not affiliated with researchers or government agencies, allows cannabis users in the US, Canada, Europe, and Australia to submit the price of cannabis, in an effort to create a semi-global index of the street value of cannabis prices (Wang, 2016). Both *FixMyStreet* and *Price of Weed* register submissions on a regular basis, thereby providing data, apart from official estimates on these issues (local disorder and retail drug pricing), which are not always publicly available. The data collected by *FixMyStreet* have been used to assess locals' perceptions of disorder (Solymosi et al., 2018) and fear of crime (Solymosi et al., 2020). Likewise, researchers have used the data collected by the *Price of Weed* (often through scraping the website) to determine price responses to law enforcement and decriminalization (Larson et al., 2015; Lawson & Nesbit, 2013; Malivert & Hall, 2013), price shifts across a geographic market (Giommoni & Gundur, 2018), and demand and price elasticity (Halcoussis et al., 2017).

In addition to these open data collection mechanisms, there are online contract labor portals, such as Mechanical Turk, TurkPrime, or Qualtrics, which have become commonly-used participant pools for behavioral science researchers (Litman et al., 2017), including criminologists (Ozkan, 2019). These mechanisms have established pools of participants; through these pools, researchers can identify the participant attributes appropriate for completing the assigned task. Moreover, links to the collection tool can be independently distributed (Graham et al., 2020). These tools, however, are not necessarily accessed by relevant populations, particularly in underserved or underconnected communities. Thus, crowdsourced data may show only the direction of variables rather than their magnitude, thereby limiting the ability to generalize from the results (Thompson & Pickett, 2019).

Sometimes, researchers have established their own data collection tools, with variable success. To be successful, tools need to be known to a large crowd of willing participants. Connecting to that crowd, particularly in criminogenic settings, may be difficult, particularly if there is no immediate benefit for the participants. For instance, the now defunct drugsource.io was a platform set up in Europe to emulate the *Price of Weed;* in addition to cannabis, drugsource.io included several other controlled substances to report. The tool did not gain traction with potential contributors and failed to receive enough submissions to undertake any meaningful analysis. Likewise, one of the authors of this chapter set up an email address to have victims of scams forward their scams. That effort was only moderately successful due to a lack of visibility and temporal duration. Accordingly, it must be noted that new crowdsourcing efforts likely need resources, such as advertisements, to target a potential crowd, and those resources need to be commensurate to the geographic scope of the project (Estellés-Arolas, 2020). In cases where the generation of information through collective reporting is the primary benefit, compensation of early contributors likely needs to occur until a critical mass is achieved. Nonetheless, crowdsourcing platforms show that systems can be built to collect nearly any kind of data.

Crowdsourcing can be used to analyze data and solve problems (Estellés-Arolas, 2020). Analysis includes identifying people, objects, or patterns in images. Social media has been used to identify specific people responsible for crimes whose likenesses were captured by CCTV. Social media allows potentially millions of users to view these images and then identify the subjects in them. In the case of searching for perpetrators of crimes, these efforts often identify assailants, such as the Boston Marathon bomber or members of hate groups (Douglas, 2020; Gray & Benning, 2019; Nhan et al., 2017). However, these efforts sometimes become an exercise in vigilantism and deviate from true crowdsourced efforts as they lack a clear crowdsourcer or a clear criminal justice purpose; moreover, if the target is misidentified, there is no clear response to remedy the error (Chang & Poon, 2017; Douglas, 2020; Loveluck, 2019).

There are, however, dedicated operational platforms set up to solicit crowdsourced analysis. For instance, the UK's *Crimestoppers* platform seeks crowdsourced help to identify suspects or missing people (Estellés-Arolas, 2020). Another example of an operational crowdsourced analysis platform is GlobalXplorer°, a platform that invites the public to search through satellite imagery to identify images with evidence of looting of cultural property. The platform trains volunteers by establishing the task as a game and then provides the over 66,000 volunteers with snippets from areas at risk of looting (Yates, 2018). These platforms show that large analytical jobs to answer unknown questions (e.g., to solve problems) can be crowdsourced in various digital and digitized criminological contexts.

Beyond the platforms in use, there are several platforms that have been conceptualized, particularly for security and safety applications, to make use of crowdsourced data. These platforms reflect the potential crowdsourcing offers to problem solving. For instance, there is *MissingFound*, a platform that would use various open source datapoints to track missing people (Liu et al., 2016). There are also conceptualized platforms to crowdsource surveillance, by making certain CCTV footage public to allow for the real-time monitoring of spaces (Trottier, 2014); to crowdsource investigation, by providing collected evidence related to child abuse available to the public to help with its examination (Açar, 2018), and to crowdsource digital vulnerability identification, by having kernel reports served to analysts who can determine malware execution patterns to better identify threat vectors (Burguera et al., 2011).

Analysis of open source and crowdsourced data

A lot of analysis, however, requires professional knowledge. Subject expertise and methodological training are critical to making sense of data collected from public sources. Subject expertise helps researchers to identify points of interest, anomalies, and misdirection within the data; to make informed assumptions when the data is incomplete; and to interpret the data vis-à-vis its context (Nichols, 2017). No tool or technology can completely replace a researcher's ability to understand the implications of settings and context. Likewise, methodological training, which teaches the fundamentals of research design and analysis, allows researchers to ensure that the established questions are worth answering and to confirm that the answers offered correspond with the questions posed.

To make sense of the data collected, various approaches have been used, depending on the type and the quality of the data. Qualitative methods have been used to analyze open source and crowdsourced data much in the same way that qualitative methods are used to analyze qualitative data collected via fieldwork. Qualitative methods have been used to conduct content analysis on text-based data to analyze posters' perceptions and behaviors (Holt, 2010; Lynch, 2018; Williams et al., 2017).

Additionally, quantitative methodologies have been employed to analyze open source and crowdsourced datasets (Tompson et al., 2015). One notable analytical technique is Social Network Analysis (SNA) (see also Chapter N). Pioneered in criminology by the late Carlo Morselli (2009), SNA has been used to show how various actors are connected in a network (Bright et al., 2018) and how drugs flow across and within borders (Berlusconi et al., 2017; Giommoni & Gundur, 2018). Another technique involves the mapping of geospatial data to predict accident and disorder hotspots (dos Santos et al., 2017; Solymosi et al., 2020). Geospatial data are used extensively in crime mapping software and military geographical intelligence (GEOINT) to provide "actionable knowledge" on specific events (US Geospatial Intelligence Foundation, 2015, p. p11). The data can be used to identify crime hotspots, understand crime distribution, assess the impact of crime reduction programs, and communicate crime statistics to a wide audience (Chainey & Ratcliffe, 2013).

Researchers have also recognized that big data sets, with sometimes millions of entries, and non-text based data present their own challenges (Solymosi & Bowers, 2018; Williams et al., 2017). Accordingly, researchers have developed tools and methodologies to go beyond what traditional analytical processes have been able to illuminate using modest data sets, particularly when evaluating big data sets and non-text-based open source data. One notable example is with the analysis of malware or ransomware. Once malware is "released" to infect members of the public, its code becomes a piece of open source data which then can be subjected to digital forensics. Digital forensics involves the examination of digital devices and systems to assess the cause of a particular event, which may be criminal in nature. It uses a variety of scientific techniques on the "validation, identification, analysis, interpretation, documentation, and presentation of digital evidence" (Delp et al., 2009, p. 14). Various tools are available to help researchers collect data and conduct analysis of that data. Autopsy, for example, is a tool used to perform data analysis on imaged and live systems and to recover deleted data (Kävrestad, 2018). Other tools facilitate open source investigations into the actors behind cyberattacks. For instance, Maltego facilitates data mining and link analysis, using open source records, such as WhoIs, domain name system (DNS), and other digital records.

Open source data analysis – particularly when considering digital forensic data – often may be too complex for the computing power of one machine alone. As a result, data collection may require a networked computing system, like the Hapdoop ecosystem, to aggregate computing power, and a distributed, networked approach, including machine learning, to aid in the analysis (Landset et al., 2015). Machine learning, which is a subset of artificial intelligence, enables computers to improve automatically through experience. Computers can detect complex patterns that are often unpredictable to humans (Chhabra et al., 2020; Pastor-Galindo et al., 2020). Machine learning has been used in the field of cybersecurity to resolve vulnerabilities from hacking and malware. From a technical viewpoint, machine learning has been also used to detect vulnerabilities in systems by analyzing code mined from open source projects (Bilgin et al., 2020). Moreover, machine learning algorithms can facilitate various types of cybercrime detection and prevention efforts, such as identifying hate speech (Burnap & Williams, 2015; Ozalp et al., 2020) and responding to cyber bullying and cyber stalking (Ghasem et al., 2015). The potential for machine learning to speed up analysis and trigger real-time responses makes it an important analytical strategy for big data.

Collaboration: The Future Direction in Digital Data Collection and Analysis

The collection and analysis of open source data will continue, and the production of available digital data will increase. This is also true in criminogenic settings as deviant behavior will increasingly leverage technology (Berry, 2018). This rich and diverse digital data will provide researchers, studying deviant and other social behaviors, with opportunities to ask questions that have been previously difficult or impossible to answer. Human-computer interactions will create an increasingly large part of the data generated. Accordingly, social and technical research questions must be combined, and, to answer these questions, academics of different disciplines, researchers, practitioners, and the public must work together to find agile, collaborative solutions to data collection and analysis.

Collaboration will allow researchers with distinct strengths to develop research programs with the potential to better pose questions in the social world and to examine them. The increase in digital data, and digital connectivity, means that researchers will need to continue to innovate in their data collection techniques, particularly if various terms and conditions of content providers bar them from using presently successful strategies. The increased penetration of internet-connected devices provides an increasing opportunity to ask members of the public to collect and contribute to crowdsourced efforts. Relevant efforts are already underway with the collection of geospatial data via smartphone apps (Moore et al., 2017)

Finally, researchers may find marrying various types of data helpful in painting nuanced and accurate pictures of social phenomena. By pushing past the limitations of using only one or two data types, which, when used singularly, can produce abstract, static, and simplified pictures of criminal activity, researchers will be able to present a more complex analysis which portrays the often chaotic nature of criminal activity (Hobbs, 2014). The evolution of data collection, however, will necessitate careful consideration of the ethical implications of data collection and use, particularly when the data collected are identifiable.

References

Açar, K. V. (2018). OSINT by Crowdsourcing: A Theoretical Model for Online Child Abuse Investigations. *International Journal of Cyber Criminology*, *12*(1), 206-229. https://doi.org/http://dx.doi.org/10.5281/zenodo.1467897

Action Fraud. (n.d.). *Reporting Fraud*. Action Fraud. Retrieved January 4 from https://web.archive.org/web/20210104015305/https://reporting.actionfraud.police.uk/login

Akhgar, B., Bayerl, P. S., & Sampson, F. (2016). *Open Source Intelligence Investigation: From Strategy to Implementation*. Springer.

Al-Nabki, M. W., Fidalgo, E., Vasco-Carofilis, R. A., Jañez-Martino, F., & Velasco-Mata, J. (2020). Evaluating Performance of an Adult Pornography Classifier for Child Sexual Abuse Detection. *arXiv preprint arXiv:2005.08766*.

An, J., & Kim, H.-W. (2018). A data analytics approach to the cybercrime underground economy. *IEEE Access*, *6*, 26636-26652.

Australian Competition & Consumer Commission. (n.d.). *ScamWatch*. ACCC. Retrieved January 4 from https://www.scamwatch.gov.au/report-a-scam
Ball, M., Broadhurst, R., Niven, A., & Trivedi, H. (2019). Data capture and analysis of darknet markets. *Available at SSRN 3344936*.
Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior research methods*, *43*(3), 800.
Berlusconi, G., Aziani, A., & Giommoni, L. (2017). The determinants of heroin flows in Europe: A latent space approach. *Social Networks*, *51*, 104-117.
Berry, M. (2018). Technology and organised crime in the smart city: an ethnographic study of the illicit drug trade. *City, Territory and Architecture*, *5*(1), 16.
Bilgin, Z., Ersoy, M. A., Soykan, E. U., Tomur, E., Çomak, P., & Karaçay, L. (2020). Vulnerability Prediction From Source Code Using Machine Learning. *IEEE Access*, *8*, 150672-150684.
Brabham, D. C. (2013). *Crowdsourcing*. MIT Press.
Bright, D., Koskinen, J., & Malm, A. (2018). Illicit Network Dynamics: The Formation and Evolution of a Drug Trafficking Network. *Journal of Quantitative Criminology*, 1-22.
Bright, D. A., Hughes, C. E., & Chalmers, J. (2012). Illuminating dark networks: A social network analysis of an Australian drug trafficking syndicate. *Crime, Law and Social Change*, *57*(2), 151-176.
Burguera, I., Zurutuza, U., & Nadjm-Tehrani, S. (2011). Crowdroid: behavior-based malware detection system for android. Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices,
Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223-242.
Cambridge Cybercrime Centre. (2020). *Cambridge Cybercrime Centre: Description of available datasets*. Computer Laboratory, University of Cambridge. Retrieved January 4, 2021 from https://www.cambridgecybercrime.uk/datasets.html
Carrington, K., Hogg, R., Scott, J., & Sozzo, M. (2018). *The Palgrave Handbook of Criminology and the Global South*. Springer.
Cascavilla, G., Beato, F., Burattin, A., Conti, M., & Mancini, L. V. (2018). OSSINT-Open Source Social Network Intelligence: An efficient and effective way to uncover “private” information in OSN profiles. *Online Social Networks and Media*, *6*, 58-68.
Červený, J., & van Ours, J. C. (2019). Cannabis prices on the dark web. *European Economic Review*, *120*, 103306.
Chainey, S., & Ratcliffe, J. (2013). *GIS and crime mapping*. John Wiley & Sons.
Chang, L. Y., & Poon, R. (2017). Internet vigilantism: Attitudes and experiences of university students toward cyber crowdsourcing in Hong Kong. *International journal of offender therapy and comparative criminology*, *61*(16), 1912-1932.
Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, *65*(1), 57-70.
Cheng, F. K. (2017). *Using email and skype interviews with marginalized participants*. SAGE Publications Ltd.
Chhabra, G. S., Singh, V. P., & Singh, M. (2020). Cyber forensics framework for big data analytics in IoT environment using machine learning. *Multimedia tools and applications*, *79*(23), 15881-15900.
Copeland, C., Wallin, M., & Holt, T. J. (2020). Assessing the Practices and Products of Darkweb Firearm Vendors. *Deviant behavior*, *41*(8), 949-968. https://doi.org/10.1080/01639625.2019.1596465
Cross, C., & Gillett, R. (2020). Exploiting trust for financial gain: An overview of business email compromise (BEC) fraud. *Journal of Financial Crime*.
Cybersecurity in Application Research and Education Lab. (2020). *Downloads*. CARE Lab. Retrieved January 4 from https://web.archive.org/web/20210104012919/https://sites.temple.edu/care/downloads/

de Souza Santos, A. A. (2018). Risky closeness and distance in two fieldwork sites in Brazil. *Contemporary Social Science*, *13*(3-4), 429-443.
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., & Bérubé, M. (2018). *Délinquance et innovation*. Les Presses de l'Université de Montréal.
Delp, E., Memon, N., & Wu, M. (2009). Digital forensics. *IEEE Signal Processing Magazine*, *26*(2), 14-15.
DOMO. (2020). *Data never sleeps 8.0*. DOMO, Inc.
dos Santos, S. R., Davis Jr, C. A., & Smarzaro, R. (2017). Analyzing traffic accidents based on the integration of official and crowdsourced data. *Journal of Information and Data Management*, *8*(1), 67-67.
Douglas, D. M. (2020). Doxing as Audience Vigilantism against Hate Speech. *Introducing Vigilant Audiences*, 259.
Dragiewicz, M., Burgess, J., Matamoros-Fernández, A., Salter, M., Suzor, N. P., Woodlock, D., & Harris, B. (2018). Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms. *Feminist Media Studies*, *18*(4), 609-625.
Ensari, E., & Kobaş, B. (2018). Web scraping and mapping urban data to support urban design decisions. *A| Z ITU Journal of the Faculty of Architecture*, *15*(1), 5-21.
Estellés-Arolas, E. (2020). Using crowdsourcing for a safer society: When the crowd rules. *European Journal of Criminology*, 1477370820916439.
Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, *38*(2), 189-200. https://doi.org/10.1177/0165551512437638
Frank, R., & Mikhaylov, A. (2020). Beyond the 'Silk Road': Assessing Illicit Drug Marketplaces on the Public Web. In *Open Source Intelligence and Cyber Crime* (pp. 89-111). Springer.
Gearon, L. F., & Parsons, S. (2019). Research ethics in the securitised university. *Journal of Academic Ethics*, *17*(1), 73-93.
Ghasem, Z., Frommholz, I., & Maple, C. (2015). Machine learning solutions for controlling cyberbullying and cyberstalking. *J Inf Secur Res*, *6*(2), 55-64.
Gillett, R. (2018). Intimate intrusions online: Studying the normalisation of abuse in dating apps. Women's Studies International Forum,
Giommoni, L., & Gundur, R. V. (2018). An analysis of the United Kingdom's cannabis market using crowdsourced data. *Global Crime*, *19*(2). https://doi.org/10.1080/17440572.2018.1460071
Graham, A., Pickett, J. T., & Cullen, F. T. (2020). Advantages of matched over unmatched opt-in samples for studying criminal justice attitudes: A research note. *Crime & Delinquency*, 0011128720977439.
Gray, G., & Benning, B. (2019). Crowdsourcing criminology: social media and citizen policing in missing person cases. *SAGE Open*, *9*(4), 2158244019893700.
Gundur, R. V. (2019). Using the Internet to Recruit Respondents for Offline Interviews in Criminological Studies. *Urban Affairs Review*, *55*(6), 1731 –1756. https://doi.org/10.1177/1078087417740430
Halcoussis, D., Lowenberg, A. D., & Roof, Z. (2017). Estimating the Price elasticity of demand for Cannabis: a geographical and crowdsourced approach. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, *23*, 119-136.
Hassan, N. A. (2019). Gathering Evidence from OSINT Sources. In *Digital Forensics Basics* (pp. 311-322). Springer.
Higgs, E. (2001). The rise of the information state: the development of central state surveillance of the citizen in England, 1500–2000. *Journal of Historical Sociology*, *14*(2), 175-197.

Hobbs, D. (2014). Organised Crime as a Community of Practice. In C. Ellis (Ed.), *Disrupting Organised Crime: Developing the evidence base to understand effective action*. RUSI. https://rusi.org/sites/default/files/201411_stfc_disrupting_organised_crime.pdf
Holt, T. J. (2010). Exploring strategies for qualitative criminological and criminal justice inquiry using on-line data. *Journal of Criminal Justice Education*, *21*(4), 466-487.
Hribar, G., Podbregar, I., & Ivanuša, T. (2014). OSINT: a “grey zone”? *International Journal of Intelligence and CounterIntelligence*, *27*(3), 529-549.
Hughes, C. E., Chalmers, J., & Bright, D. A. (2020). Exploring interrelationships between high-level drug trafficking and other serious and organised crime: an Australian study. *Global Crime*, *21*(1), 28-50. https://doi.org/10.1080/17440572.2019.1615895
James, J. (2021). Geographies of the internet in rural areas in developing countries. In B. Warf (Ed.), *Geographies of the Internet* (pp. 93-114). Routledge.
Kävrestad, J. (2018). Open-Source or Freeware Tools. In J. Kävrestad (Ed.), *Fundamentals of Digital Forensics: Theory, Methods, and Real-Life Applications* (pp. 153-172). Springer International Publishing. https://doi.org/10.1007/978-3-319-96319-8_14
Kigerl, A. (2018). Profiling Cybercriminals: Topic Model Clustering of Carding Forum Member Comment Histories. *Social Science Computer Review*, *36*(5), 591-609. https://doi.org/10.1177/0894439317730296
Kolajo, T., & Daramola, O. (2017, 8-10 March 2017). Leveraging big data to combat terrorism in developing countries. 2017 Conference on Information Communication Technology and Society (ICTAS),
Kostakos, P. (2018). Public Perceptions on Organised Crime, Mafia, and Terrorism: A Big Data Analysis based on Twitter and Google Trends. *International Journal of Cyber Criminology*, *12*(1), 282-299. https://doi.org/http://dx.doi.org/10.5281/zenodo.1467919
Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, *2*(1), 24.
Larson, R. A., Rusko, C. J., & Secor, A. E. (2015). A Blunt Analysis: Marijuana Policy Liberalization and Market Prices in Colorado and Washington. *Centre College Empirical Analysis Paper*.
Lavorgna, A., & Sugiura, L. (2020). Direct contacts with potential interviewees when carrying out online ethnography on controversial and polarized topics: a loophole in ethics guidelines. *International Journal of Social Research Methodology*, 1-7.
Lawson, R. A., & Nesbit, T. M. (2013). Alchian and Allen revisited: law enforcement and the Price of Weed. *Atlantic Economic Journal*, *41*(4), 363-370.
Lewis, S. J. (n.d.). *Onion Scan*. Retrieved January 4 from https://web.archive.org/web/20201222060839/https://onionscan.org/
Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods*, *49*(2), 433-442.
Liu, W., Li, J., Zhou, Z., & He, J. (2016). MissingFound: An Assistant System for Finding Missing Companions via Mobile Crowdsourcing. *KSII Transactions on Internet & Information Systems*, *10*(10).
Loveluck, B. (2019). The many shades of digital vigilantism. A typology of online self-justice. *Global Crime*, 1-29.
Lynch, J. (2018). Not even our own facts: Criminology in the era of big data. *Criminology*, *56*(3), 437-454.
Mackey, T. K., Li, J., Purushothaman, V., Nali, M., Shah, N., Bardier, C., Cai, M., & Liang, B. (2020). Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales: Infoveillance Study on Twitter and Instagram. *JMIR public health and surveillance*, *6*(3), e20794-e20794. https://doi.org/10.2196/20794
Malivert, R., & Hall, J. C. (2013). The effect of medical marijuana laws on extralegal marijuana prices. *Atlantic Economic Journal*, *41*(4), 455-456.

Martin, J., & Christin, N. (2016). Ethics in cryptomarket research. *International Journal of Drug Policy*, *35*, 84-91.
Mba, G., Onaolapo, J., Stringhini, G., & Cavallaro, L. (2017). Flipping 419 cybercrime scams: Targeting the weak and the vulnerable. Proceedings of the 26th International Conference on World Wide Web Companion,
McAlister, R. (2015). Webscraping as an investigation tool to identify potential human trafficking operations in Romania. Proceedings of the ACM Web Science Conference,
Mitchell, R. (2018). *Web scraping with Python: Collecting more data from the modern web*. " O'Reilly Media, Inc.".
Moore, J., Baggili, I., & Breitinger, F. (2017). Find Me If You Can: Mobile GPS Mapping Applications Forensic Analysis & SNAVP the Open Sour VP the Open Source, Modular, Extensible Parser. *The Journal of Digital Forensics, Security and Law*, *12*(1), 15-29.
Morselli, C. (2009). *Inside criminal networks*. Springer.
Moule Jr, R. K., Pyrooz, D. C., & Decker, S. H. (2013). From 'What the F#@% is a Facebook?'to 'Who Doesn't Use Facebook?': The role of criminal lifestyles in the adoption and use of the Internet. *Social Science Research*, *42*(6), 1411-1421.
Neri, F., & Geraci, P. (2009). Mining textual data to boost information access in OSINT. 2009 13th International Conference Information Visualisation,
Nhan, J., Huey, L., & Broll, R. (2017). Digilantism: An analysis of crowdsourcing and the Boston marathon bombings. *The British Journal of Criminology*, *57*(2), 341-361.
Nichols, T. (2017). *The death of expertise: The campaign against established knowledge and why it matters*. Oxford University Press.
Oleji, C., Nwokorie, E., & Chukwudebe, G. (2008). Big data analytics of Boko haram insurgency attacks menace in Nigeria using DynamicK-reference Clustering Algorithm. *International Research Journal of Engineering and Technology*, *7*(1).
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), 2056305120916850.
Ozkan, T. (2019). Criminology in the age of data explosion: New directions. *The social science journal*, *56*(2), 208-219.
Pastor-Galindo, J., Nespoli, P., Mármol, F. G., & Pérez, G. M. (2020). The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends. *IEEE Access*, *8*, 10282-10304. https://doi.org/10.1109/ACCESS.2020.2965257
Poletti, C., & Gray, D. (2019). Good data is critical data: An appeal for critical digital studies. In A. Daly, S. K. Devitt, & M. Mann (Eds.), *Good data*. Institute of Network Cultures.
Potter, G. R. (2017). Real gates to virtual fields: Integrating online and offline ethnography in studying cannabis cultivation and reflections on the applicability of this approach in criminological ethnography more generally. *Methodological Innovations*, *10*(1), 2059799117720609.
Powell, A., Stratton, G., & Cameron, R. (2018). *Digital criminology: Crime and justice in digital society*. Routledge.
Price of Weed. (2019). *Price of Weed*. Retrieved January 4 from https://web.archive.org/web/20201223133241/http://www.priceofweed.com/
Ramo, D. E., & Prochaska, J. J. (2012). Broad reach and targeted recruitment using Facebook for an online survey of young adult substance use. *Journal of medical Internet research*, *14*(1), e28.
Reidenberg, J. R. (2014). The data surveillance state in the United States and Europe. *Wake Forest L. Rev.*, *49*, 583.
Roth, F. (2020). *Ransomware Overview*. Retrieved December 28 from https://www.google.com/url?q=http://goo.gl/b9R8DE&sa=D&ust=1609726179935000&usg=AFQjCNEa4yx_tweL7qwFDk3JYPHmpBhplQ

Sampson, F. (2016). Intelligent evidence: Using open source intelligence (OSINT) in criminal proceedings. *The Police Journal*, *90*(1), 55-69. https://doi.org/10.1177/0032258X16671031

ScamWarners. (2021). Retrieved January 4, 2021 from https://web.archive.org/web/20210104011455/https://www.scamwarners.com/forum/

Schaurer, F., & Störger, J. (2013). The evolution of open source intelligence (OSINT). *Comput Hum Behav*, *19*, 53-56.

Schoepfer, A., & Piquero, N. L. (2009). Studying the correlates of fraud victimization and reporting. *Journal of Criminal Justice*, *37*(2), 209-215. https://doi.org/https://doi.org/10.1016/j.jcrimjus.2009.02.003

Smith, G. J., Bennett Moses, L., & Chan, J. (2017). The challenges of doing criminology in the big data era: Towards a digital and data-driven approach. *British Journal of Criminology*, *57*(2), 259-274.

Society Works. (n.d.). *Fix My Street*. mySociety. Retrieved January 4 from https://web.archive.org/web/20201202190436/https://www.fixmystreet.com/

Solymosi, R., & Bowers, K. (2018). The role of innovative data collection methods in advancing criminological understanding. *The Oxford handbook of environmental criminology*, 210-237.

Solymosi, R., Bowers, K. J., & Fujiyama, T. (2018). Crowdsourcing subjective perceptions of neighbourhood disorder: Interpreting bias in open data. *The British Journal of Criminology*, *58*(4), 944-967.

Solymosi, R., Buil-Gil, D., Vozmediano, L., & Guedes, I. S. (2020). Towards a place-based measure of fear of crime: A systematic review of app-based and crowdsourcing approaches. *Environment and Behavior*, 0013916520947114.

Thompson, A. J., & Pickett, J. T. (2019). Are relational inferences from crowdsourced and opt-in samples generalizable? Comparing criminal justice attitudes in the GSS and five online samples. *Journal of Quantitative Criminology*, 1-26.

Tompson, L., Johnson, S., Ashby, M., Perkins, C., & Edwards, P. (2015). UK open source crime data: accuracy and possibilities for research. *Cartography and geographic information science*, *42*(2), 97-111.

Trottier, D. (2014). Crowdsourcing CCTV surveillance on the Internet. *Information, Communication & Society*, *17*(5), 609-626.

Trottier, D. (2015). Open source intelligence, social media and law enforcement: Visions, constraints and critiques. *European Journal of Cultural Studies*, *18*(4-5), 530-547.

Turk, K., Pastrana, S., & Collier, B. (2020). A tight scrape: methodological approaches to cybercrime research data collection in adversarial environments. 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW),

US Geospatial Intelligence Foundation. (2015). State of GEOINT 2015. Retrieved 21/12/2020, from http://usgif.org/system/uploads/3661/original/SOG_FINAL.pdf

Wang, M. (2016). Crowdsourcing the landscape of cannabis (marijuana) of the contiguous United States. *Environment and Planning A*, *48*(8), 1449-1451.

Williams, H. J., & Blum, I. (2018). *Defining second generation open source intelligence (OSINT) for the defense enterprise*.

Williams, M. L., Burnap, P., & Sloan, L. (2017). Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *The British Journal of Criminology*, *57*(2), 320-340.

Winchester, S. (2018). *The meaning of everything: The story of the Oxford English Dictionary*. Oxford University Press.

Wong, R. W. (2015). A note on fieldwork in 'dangerous' circumstances: interviewing illegal tiger skin suppliers and traders in Lhasa. *International Journal of Social Research Methodology*, *18*(6), 695-702.

Wood, M., Richards, I., Iliadis, M., & McDermott, M. (2019). Digital public criminology in Australia and New Zealand: results from a mixed methods study of criminologists' use of social media. *International Journal for Crime, Justice and Social Democracy*, *8*(4), 1.

Yates, D. (2018). Crowdsourcing antiquities crime fighting: a review of GlobalXplorer. *Advances in Archaeological Practice*, *6*(2), 173-178.

Zhao, B., & Sui, D. Z. (2017). True lies in geospatial big data: detecting location spoofing in social media. *Annals of GIS*, *23*(1), 1-14. https://doi.org/10.1080/19475683.2017.1280536