Too Much Data? Opportunities and Challenges of Large Datasets and Cybercrime [Accepted Version]

Jack Hughes, Yi Ting Chua, Alice Hutchings
Department of Computer Science & Technology
University of Cambridge
15 JJ Thomson Avenue
Cambridge CB3 0FD

Abstract

Never before have criminologists have had such rich data about the communications of a wide variety of individuals involved at various stages of crime. We now have records of discussions held between cybercrime offenders going back 20 years. Indeed, given we now have over 70 million posts by almost two million users, we are encountering a different type of problem: we have too much data. Although the datasets potentially allow us to answer questions we never before thought were possible, we also face unique challenges such as categorisation of large datasets and temporal shifts in users, topics, ideas and ways of communications. One answer to this problem may lie in automation: using machine learning to classify and label posts and interactions at scale. In this chapter, we will outline some of the opportunities and challenges associated with using such large datasets, some of the ways we are currently addressing these challenges, and potential ways forward.

Keywords

Cybercrime; big data; large datasets; data collection; databases; toolkits; data analysis

Introduction

The term "big data" is gaining popularity and attention among criminologists, ever since its introduction in a McKinsey report (Manyika et al., 2011). However, its definition and conceptualisation varies across disciplines, with some computer scientists viewing it as a marketing term or buzzword. Some may view big data by unit of measurement, such as datasets that exceed a petabyte in size (Anderson, 2008), while others conceptualise big data in terms of the number of rows and columns in a dataset (Burrows & Savage, 2014; Lazer & Radford, 2017). In addition to volume, big data is characterised by other features such as exhaustivity (capturing an entire population), relationality (features allowing for integration with other datasets), and velocity (the speed of data collection and utilisation) (Kitchin, 2014; Ozkan, 2019). Despite such variations, a shared notion of big data is that the size of the datasets creates challenges, with existing software and tools unable to manage the logistics of data collection, analysis, and management (Chan & Moses, 2016; Chen, Mo, & Liu, 2014; Lynch, 2018;

Manyika et al., 2011; Snijders et al., 2012). Given the complexities with "big data", the chapter will use the phrase "large datasets" in place of big data to minimise confusion.

As technology and digital devices become highly integrated into society, it creates tremendous amounts of digital data and information ranging from user-generated content to personal identifiable information (Burrows & Savage, 2014; Lazer & Radford, 2017; Ozkan, 2019). This data may not only become the target of cybercrime attacks (Porcedda & Wall, 2019; Yar, 2005), but also of academic research. Increasingly, social scientists are using these large datasets to understand human behaviour. The shift towards digital data presents unique opportunities and challenges to criminologists and cybercrime scholars. The intersection of large datasets, data science, and social science has resulted in the birth of new fields such as computational social science (Lazer et al., 2009), digital criminology (Smith et al., 2017), and social computing (Wang et al., 2007). In addition, there are increasingly new tools and datasets previously inaccessible to scholars (Metzler et al., 2016; Moore et al., 2019). Such advancements require new developments in research, to process large datasets in a short span of time, and the use of new techniques such as natural language processing and machine learning (Benjamin et al., 2015; Chan & Moses, 2016; Chen et al., 2014; Lazer & Radford, 2017; Li, Chen, & Nunamaker, 2016). This creates new challenges in research methodology, such as sampling biases and missing data (Edwards et al., 2013). Current discussions on these impacts tend to revolve around criminology as a field, and do not account for unique issues faced by cybercrime scholars. The goal of this chapter is to provide an overview of the opportunities and challenges of large datasets in the context of cybercrime.

The chapter begins with a background on the emergence of large datasets in the field of criminology and recent trends of research in cybercrime using large datasets. The discussion is followed by detailed accounts of opportunities and challenges cybercrime scholars may encounter during the research process, ranging from data collection to data analysis. In addition, ethical issues unique to the use of large datasets in research are considered.

Background

An early concern with the rise of large datasets is the decreased necessity for theoretical framework and discussions. Arguing that the era of large datasets is the "end of theory", Anderson (2008) states the sufficiency of correlation with large volume of data and cited Google as a demonstration of the power of algorithms and large datasets. Criminologists dismiss such notions, arguing that the expertise of scholars is necessary to make sense of data-driven findings. There has been a trend in criminology towards using large datasets as research data, as well as using data science tools (i.e. modelling and algorithms) to analyse them (Chan & Moses, 2016). These allow researchers to perform more complex and comprehensive modelling and analyses, and at faster pace due to automation (Chen et al., 2014; Lazer & Radford, 2017; Ozkan, 2019; Snaphaan & Hardyns, 2019). However, the necessity of theories is found in the ineffectiveness of data science methods to identify underlying causal mechanisms, or to distinguish between noises and true signals within large datasets (Chan & Moses, 2016; Ozkan, 2019).

Some common types of large datasets used in criminology and criminal justice include administrative data, social media datasets (i.e. Facebook, Twitter, Reddit, etc.), and survey data (Burrows & Savage, 2014; Lynch, 2018; Metzler et al., 2016; Ozkan, 2019), as well as leaked data of online communities (i.e. underground forums) (Holt & Dupont, 2019; Motoyama et al., 2011) and scraped data of online communities (Benjamin et al., 2015; Pastrana et al., 2018a; Westlake & Bouchard, 2016). Recent research in criminology shows social scientists' keenness in incorporating such powerful items into their toolboxes. One example is the application of artificial intelligence and big data in the context of predictive policing and surveillance (Gerritsen, 2020; Hayward & Maas, 2020). Another example is in the advancement of environmental criminology, with a recent review by Snaphaan and Hardyns (2019) finding a shared recognition on the availability, scalability, and temporal and spatial granularity afforded by large datasets, allowing for greater insight on the time-place dynamics in crime. Lynch (2018) recognises similar benefits with open-source data and administrative records in the advancement of criminology research in the Presidential Address to American Society of Criminology in 2017. As the awareness of the benefits of large datasets grows, there is undeniable evolution in criminology research.

The rise of large datasets, however, cannot be fully embraced without discussion on its unique limitations and challenges. Large datasets are a double-edged sword, especially in the context of cybercrime. In addition to being a data source and methodology, large datasets are potentially targets and means of cyber-attacks (Newman & Clarke, 2003; Motoyama et al., 2011). For cybercrime and cybersecurity scholars, this shift in attacks dictates changes in the nature and format of available datasets necessary for research. This shift raises a series of challenges at various stages of research, including data collection, data management, and data analyses. Three common challenges from the reported experiences of 9,412 social scientists who work with large datasets include: 1) gaining access to commercial data; 2) finding collaborators with appropriate skills and knowledge; and 3) learning new software and analytical methods. Social scientists without large-dataset experience stated similar barriers, with the exception of accessibility (Metzler et al., 2016).

In addition, there are general issues and challenges due to the nature of large datasets. First, there is the issue of generalisability (Lazor & Radford, 2017; Edwards et al., 2013). For example, social media platforms differ in regulations, format of organisations, and user demographics, with some platforms being more attractive to specific populations (Lazor & Radford, 2017). There is also the issue of authenticity of collected information, such as distinction between bot-generated content and real user information or the absence of demographic information from individual users on social media platforms (Lazor & Radford, 2017; Edwards et al., 2013). Second, there is the issue of data quality and biases in large datasets (Lynch, 2018; Gonzalez-Bailon, 2013). With open-source data, there are no standards or quality assurance in place (Lynch, 2018; Lee & Holt, 2020). With large datasets provided by third-parties or via application programming interfaces (APIs), there are sometimes limits and sampling filters imposed which researchers may not necessarily be aware of (Gonzalez-Bailon, 2013, Lee & Holt, 2020). Both issues can affect the replicability of findings. This is further exacerbated when examined alongside with the accessibility of datasets (Lee & Holt, 2020). Although not exhaustive, these challenges and issues with large datasets merit further considerations for cybercrime scholars. In the following section, we will discuss in detail how some of these issues appear during cybercrime research.

Large datasets and cybercrime research

Collecting Data

Cybercrime is a hidden activity, due to its illicit nature. One of the problems facing researchers is access to good quality data. Biases in datasets can sometimes tell us more about how the data were collected than about cybercrime itself. While some data may be available from industry sources through non-disclosure agreements, as researchers cannot disclose where it was obtained or share it with others, the data and research are non-reproducible. Research students, who are often time-limited, may find that a considerable duration is spent trying to access and collect data than actually analysing it. The Cambridge Cybercrime Centre was established in 2015 in recognition of these issues. The Centre collects cybercrime-related datasets and makes them available to other academic researchers through data sharing agreements (Cambridge Cybercrime Centre, 2019). Some of these datasets are collected directly by the Centre, and others are provided by industry through incoming agreements.

The datasets being collected constantly evolves due to new criminal opportunities, as well as the methods available to us. As the datasets are continually being collected, they also tend to grow over time. One example is the various types of honeypots we operate. Honeypots mimic vulnerable machines, and collect data about attacks. We operate honeypots that pretend to be vulnerable to certain types of malware (such as Mirai, which predominantly affects Internet of Things devices), devices that can be remotely accessed, as well as misconfigured servers that are used for amplifying denial of service attacks (Vetterl & Clayton, 2019).

In relation to the denial of service attack sensors, we operate around 100 sensors in various locations around the world (Thomas et al., 2017). These sensors pretend to participate in 'reflected' or 'amplified' denial of service attacks, in which the attacker sends the misconfigured server a small packet, but pretends the request is coming from the IP address they want to attack. The server replies to the request with a much bigger packet, but to the spoofed victim IP address. This means that the traffic being sent to the victim is considerably larger than the traffic initially sent by the attacker. Our sensors pretend to be a misconfigured server, but instead of sending on the response, captures data about the attack, including the victim IP address. In 2020, we had collected over 4 trillion packets, and this is likely to continue to increase. This dataset allows for research into attacks over time, such as measuring the impact of interventions (Collier et al., 2019).

The dataset requested most frequently by academic researchers is the CrimeBB dataset (Pastrana et al., 2018a). This includes posts scraped from publicly available underground forums that discuss cybercrime and related activities. Some of these forums have been operating for many years and we have now amassed a collection of over 90 million posts, some dating back more than 10 years. We have English, Russian, German, and Spanish-language forums, and are actively expanding to include other languages. We hold nearly complete collections of these forums, allowing for analysis across the whole dataset. We have also been collecting data from additional platforms, and currently hold over three million Telegram messages (from 50+ channels) and 2.5 million Discord messages (from over 3,000 channels) that relate to cybercrime. Furthermore, we are expanding the scope beyond cybercrime, with the ExtremeBB dataset including extremist forums related to hate groups, extremism and radicalisation, as well as incel communities.

These datasets are collected by scraping the sites, using a custom script to crawl the forums for automatic collection (Pastrana et al. 2018a). There are also off-the-shelf tools for scraping, but all methods require continual monitoring and maintenance, as these can break when page layouts change. Additionally, some websites may be hostile to scraping methods, requiring the researcher to take steps to avoid detection (Turk et al., 2020). This can include limiting the rate of pages collecting, and solving CAPTCHAs either manually or automatically.

Other data sources, such as Twitter, often have an API available, which provide a standard way to collect data, used by libraries in R and Python. This simplifies data collection. Furthermore, CrimeBB is not the only dataset available on cybercrime and underground forum data. Pushshift.io collects data from Twitter, Reddit and Gab, and additionally provides a hosted Elasticsearch service for researchers to search, filter, and aggregate data without needing to download the entire dataset (Bevensee et al., 2020). This minimises the need for researchers to transform and prepare the dataset.

Using the Data

In addition to the time-consuming data collection process of large datasets, there are additional challenges when it comes to their handling and analysis. Usability of these types of datasets is explored in an evaluation study by Pete and Chua (2019), who found that large datasets pose issues even for technology-savvy researchers.

Where the data has been self-collected, the researcher needs to convert the raw data into a format that is useful to them. CrimeBB is built upon PostgreSQL, which is useful for querying and aggregating data. Tools such as PgAdmin provide a visual interface for querying data and downloading CSV files, although this can also be carried out using the command line if preferred. Difficulties reported by Pete and Chua (2019) include challenges with downloading and setting up datasets due to version conflicts for PostgreSQL and compatibility of raw datasets with other tools such as mySQL, despite overall satisfaction with the data sharing and usage.

If the dataset is in the form of a database, some skill may be required in writing SQL queries, although there are off-the-shelf tools available for working with these either in R or Python, or in standalone programs. Scholars also face the decisions of sampling, if only using a subset of the larger datasets. In some cases, the researcher may wish to convert the database into a different format, such as CSV files or edgelists (e.g. for social network analysis). However, it may not be useful to transform the entire dataset immediately, as the analysis software used may not be able to handle files of that size. It may be useful to use scripts in R or Python to convert between different formats when working with large datasets, rather than by hand.

Before using the data for analysis, if the data are raw and unprocessed, data cleaning and preparation needs to be carried out. This may involve checking records contain data where required, checking that descriptors of variables make sense, noting limitations with the dataset (e.g. there may not be a full historical dataset of forum posts as some may have been deleted before collection), and removing unrelated data (e.g. removing link and image tags in posts). Dataset filtering can take place either before or after the cleaning step has taken place, depending on the method. Filtering includes selecting data required for analysis, such as by date range.

Toolkits

Toolkits are a set of analysis techniques, provided to either connect with your own dataset, or use a dataset provided with the toolkit. Examples of these toolkits include SMAT (Bevensee et al., 2020), which provides searching and visualisation tools to explore the Pushshift database, and OSoMe (Davis et al., 2016) for analysing the OSoMe Decahose dataset with tools for bot detection, trend tracking, and social network visualisation. Toolkits vary in flexibility: those which are linked with an existing dataset are useful in reducing the time taken in carrying out analysis, allowing for early exploration of the dataset.

Other types of toolkits, such as Gephi (Bastian et al., 2009), provide flexibility with the type of data used, with a trade-off in ease-of-use, and require the researcher to transform the data into a format suitable for the tool, and setup the software. While Gephi runs as a single program, some toolkits like SMAT use a collection of programs and can be accessed using a web browser. This adds some complexity, but the toolkit developers may implement newer techniques, such as using Docker to support reproducibility, by defining what programs need to run and how they connect, to automatically setup the software, isolated from the researcher's own computer programs.

Analysing Large Datasets

There have been numerous studies working with large datasets relating to cybercrime. Some examples include the OSoMe project for Twitter data (Davis et al., 2016) and the IMPACT program with a wide range of data on the issues of cybercrime and cybersecurity (Moore et al., 2019). We specifically focus on our own work within this area, specifically on underground forums, working with CrimeBB.

Pastrana et al. (2018b) use the CrimeBB dataset to analyse key actors within HackForums, the largest and most popular underground hacking forum. The subset of CrimeBB containing HackForums data included more than 30 million posts created by 572k users. They analysed key actors, defined as those of interest to law enforcement, using ground truth data obtained from various sources. For their analysis, they use different methods including natural language processing for text content, social network analysis to analyse the relations between users, machine learning for clustering (grouping) forum members based on their activity, and explore the evolving interests of key actors.

Some of the types of approaches we have used are covered in Chapter XX, in relation to machine learning and natural language processing (NLP). We provide a brief overview here, related to their use for analysing large, unstructured, datasets. NLP is a field of computer science that focuses on analysing and understanding text content. NLP techniques are therefore useful for analysing large datasets of text, where qualitative coding is too time consuming. Caines et al. (2018) automatically classify posting intent and types using NLP and machine learning. They identify a set of words frequent in a small set of posts, but not frequent across all of CrimeBB ("TF-IDF"), as their dependent variables ("features" in machine learning terminology), to predict intent and type (dependent variable). They use a support vector machine (SVM) model for prediction. Given the large number of posts available on CrimeBB, this approach can scale a small set of manually curated annotations to automatically annotate all posts. However, care must be taken for validation to make sure that the automatic annotations are similar to the human annotations. In machine learning, the set of annotations is usually divided up into three sets: a training set, testing set, and validation set. The training set is used to create the model, and the testing set is used to evaluate the fit of the model and to tweak parameters. The validation set is used to create evaluation metrics, and should not be used to tweak parameters or re-fit the model.

Pastrana et al. (2018b) also use machine learning for clustering: grouping of similar members based on a set of variables. Specifically, they use k-means clustering. Clustering techniques can be found in statistical packages, as well as within Python and R libraries for machine learning. Other techniques used in these studies include social network analysis, with social networks based upon a user B replying directly to user A, or user C replying in a thread created by A. As the HackForums data includes positive and negative reputation voting, Pastrana et al (2018b) were able to annotate the relations between users. Pastrana et al. (2018b) also explore the evolution of interest of members on the forum. They define a member's interest for a given time period based upon the number of posts created in a category added to three times the number of threads created in a category. This is used to explore the changing interests of key actors over time. For predicting key actors overall, the authors combine the predictions of different models, filtering out those selected by only one model. These predictions are validated by checking the similarity of hacking-related terms used by key actors to those predicted key actors. This metric based approach supports validation of the predictions, although a qualitative sampling approach could be used instead.

Hughes et al. (2019) use a similar hybrid-model approach, outlined in Figure 1, with a gaming-specific underground hacking forum. As ground truth data on actors of interest to law enforcement was not available relating to users of this forum, this research defined key actors as users who have a central role in overtly criminal activities, or activities which could lead to later offending, and hence might benefit most from interventions. This work explored creating a systematic data processing pipeline for the analysis of unstructured forum datasets. Predictive models used included logistic regression, random forest and neural network models. Results were cross-validated using topic analysis. Furthermore, Hughes et al., (2019) extended the types of methods used for analysing forum members to include group-based trajectory modelling, a statistical technique developed by Nagin (2005) that takes time-series data and groups these into different trends over time (‘trajectories’). For validation, in addition to checking the similarity of terms used, Hughes et al. (2019) used methods to explore model predictions. For logistic regression, observing odds ratios is straightforward, but for some machine learning models, this is non-trivial. Partial dependency plots and permutation importance were used to identify important features (independent variables) used in prediction. These state-of-the-art techniques work by changing different input parameters to detect changes in the prediction, rather than inspecting the whole model directly.

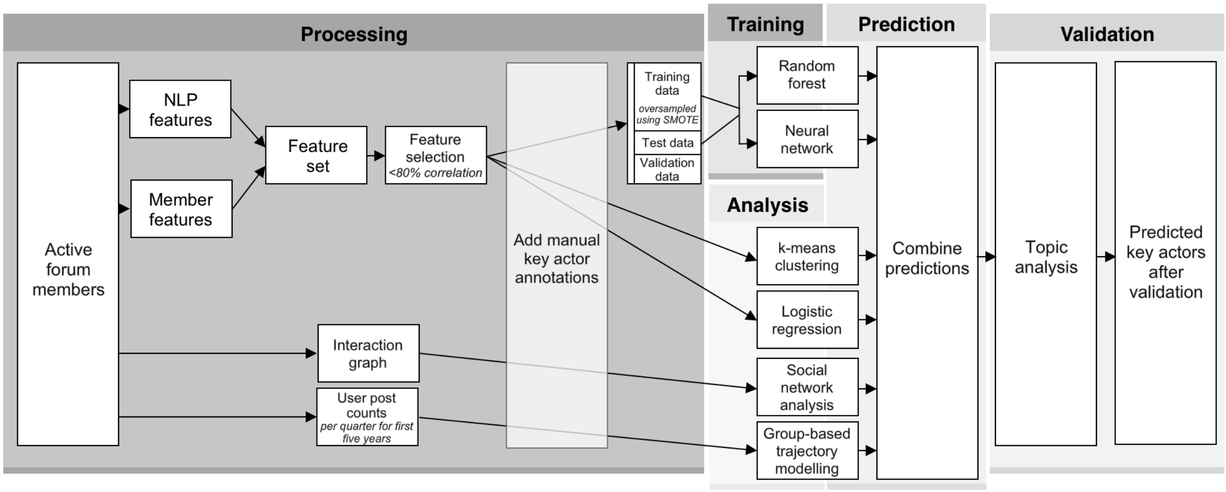

Figure 1: Pipeline used to process data and analyse forum members (adapted from Hughes et al., 2019)

The development of NLP tools to aid the analysis of cybercrime forums is ongoing. For example, to aid in the discovery of new topics that arise within underground forums, Hughes et al. (2020) developed a tool for detecting trending terms. Just like Twitter allows its users to view items being discussed at a particular time, the idea behind this tool is that researchers can uncover important topics within the noise of constant chatter.

Other studies have used CrimeBB to understand and quantify specific crime types, such as eWhoring, a type of online scam where offenders obtain money from customers in exchange for online sexual encounters (Pastrana et al., 2019). As outlined in Figure 2, eWhoring-related images shared on the forum were extracted and processed. There were significant ethical issues to address, as many of the images are pornographic, and there was a risk they may include child sexual abuse material. Therefore, the researchers worked with the Internet Watch Foundation, and were able to use their PhotoDNA service, which computes a hash of a given image and matches it against a database of known child abuse material. Indeed, some offending images were detected, which were subsequently reported for takedown and removed from the Cambridge Cybercrime Centre’s servers.

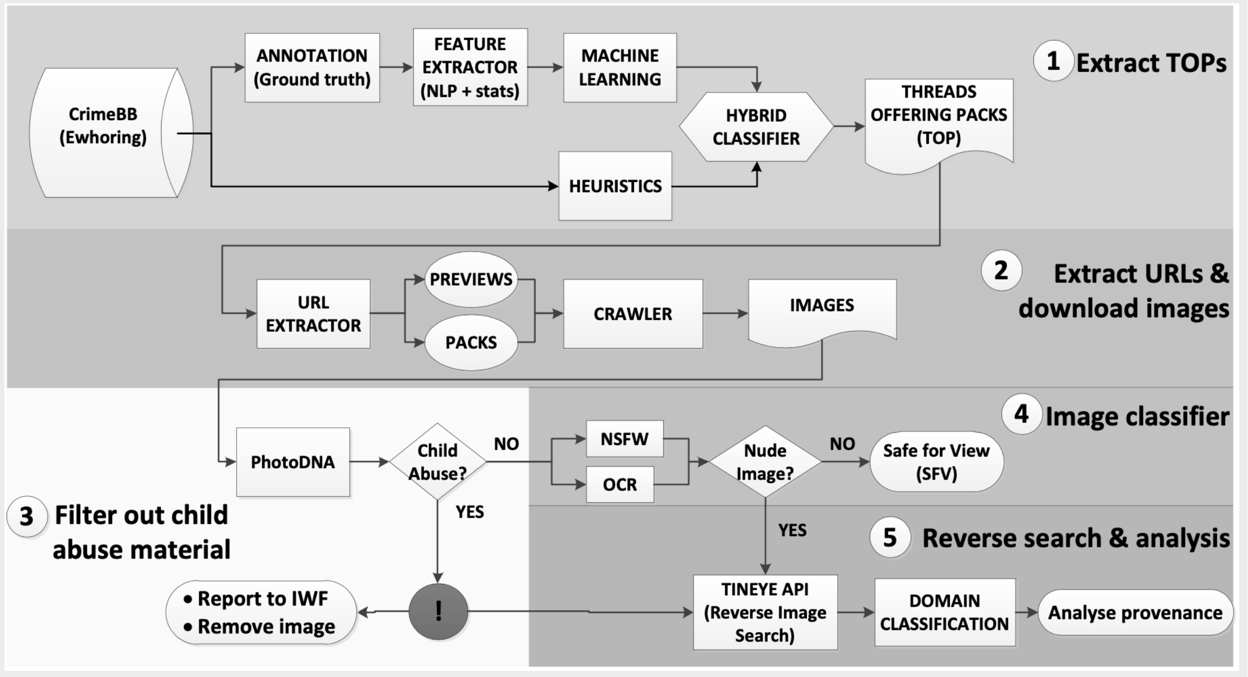

Figure 2: Pipeline used to identify, download and analyse packs of images used for eWhoring (adapted from Pastrana et al., 2019)

The images were also processed to categorise them as 'nude images' (e.g. pornographic photographs which are sold and shared on the forum), or 'safe to view'. This processing involved the use of a 'NSFV' ('not safe for viewing') classifier, a machine learning model that provides a probability of an image containing indecent content, and an optical character recognition classifier, to identify images containing text. The nude images were then reverse image searched using the TinEye API, to identify where else on the Internet they could be found. This revealed most images were likely being obtained from adult and pornography sites, as well as social networking sites, blogs, photo sharing sites, and other online forums. Pastrana et al. (2019) also manually analysed the images containing text, as these predominantly contained 'proofs of earnings', namely screenshots of payment platform dashboards which were shared as a way of boasting and gaining kudos within the community. This allowed for the estimation of the amount of money being made per customer, popular payment platforms (PayPal and Amazon gift cards), and currency (predominantly USD).

We have also explored specific segments of forums, with Vu et al. (2020) focusing specifically on the evolution of HackForums' contract system, which was introduced to build trust on the marketplace by creating a partially public record of transactions between users. The use of the system was enforced for all members using the marketplace, and members could optionally make details of the transaction public. The researchers collected contracts across a two year period, from the setup of the system to June 2020, which we examined using Tuckman's (1965) stages of group development, identifying three eras: "setup" (forming and storming), "stable" (norming), and "COVID-19" (performing). The longitudinal dataset supported measurements of how the contract activity changed through different stages of the development of the system, and while the researchers were not able to identify the purpose of all contracts, they were able to measure the types of public contracts. They found most public contracts were linked to currency exchange and payments, indicating the marketplace was being used for cashing-out. Over time, the evolution of the types of products offered was stable, and Bitcoin remained the most preferred payment method, followed by PayPal.

The researchers also used Latent Transition Modelling to identify latent classes in the dataset, for modelling activity over time by different types of users, based upon making/accepting of contracts, power-users, and counts of contracts. This supported analysis of identifying flows between types of users, to find which were the most common types of makers and takers on the marketplace, for each of the three eras. Finally, the researchers investigated the cold start problem on the marketplace, in which new members starting in the second ("stable") era face the challenge of getting started on the market to build up reputation against members that have already built up reputation during the "setup" era. Using clustering and regression, followed by qualitative analysis, the researchers found groups of users of the contracts system would overcome the cold start problem by starting with low-level currency exchange. The results of clustering and regression were used to select a group of users who overcame the cold start problem, for further qualitative analysis to understand the types of products and services they were providing.

Comparative studies compare and contrast different forums. The research by Pete et al. (2020) takes this approach with six forums that are hosted as hidden services on anonymity networks. They use a social network approach to model interactions between members of forums, for identifying structural characteristics and patterns, and identifying a group of members of importance in communities. The six forums varied in size, from 1,127 members to 40,763 members, and the researchers built an interaction network between users to observe these communities of members. Interactions on forums are not explicit, compared to Twitter mention data, and therefore some assumptions were needed including creating an interaction between members posting in the same thread. The researchers used scripts to automatically pre-process and run the network analysis given the size of the dataset, using social network analysis techniques including community detection, and network centrality. They used qualitative analysis to analyse and characterise the types of posts created by the central members of each community, who typically discussed general topics related to the content of each forum, and some members were administrators of forums.

Due to the size of the datasets we use, much of our work has focused on quantitative analysis methods, using social network analysis, natural language processing, and machine learning techniques. However, qualitative research methods come into their own when it comes to analysing the richness of the available data. Our research using qualitative analysis relies on sampling techniques, such as stratified random sampling and snowball sampling, to obtain a dataset of suitable size. This uses an initial set of keyword searches, to find relevant posts and similar keywords, with the set of keywords increased until a suitable number of posts have been found. Limitations to be aware of include not collecting all relevant posts, particularly due to the use of informal language, variations of spelling, and the changing meaning of terms over time. Examples of research where we have used such qualitative research methods include a crime script analysis of eWhoring (Hutchings & Pastrana, 2019), and an exploration into masculinity and perceptions of gender (Bada et al., 2020).

Conclusion

Nils Christie (1997) differentiated between ‘near data’, which includes information pertaining to a small number of participants while providing thousands of insights, and ‘distant data’, such as large datasets from official records, which may contain thousands of cases but provides little in-depth understanding. With large datasets, particularly with the ability to match data and add value by combining with other data sources, we are perhaps entering a time of distant-yet-near data, with rich insights ascertainable from large numbers of cases. Criminologists have never before had access to the conversations of a large number of individuals who are starting to become involved in crime, covering their learning, interaction, benefits and rewards, concerns and fears, initiation, persistence, and desistance. We are just starting to collect such data, and the insights we can glean will be of benefit to a large body of criminological research, not just cybercrime, including critical criminology, policing studies, and crime prevention.

What will make this possible, however, is not just access to the data, but novel ways to interpret and analyse it. While the Cambridge Cybercrime Centre has spent many years collecting data, we are now also turning our efforts to develop better tools to enable use by researchers in the social sciences. Here, the techniques used by other disciplines, such as computer science, are valuable. Machine learning and natural language processing allow tasks previously undertaken manually (such as coding of large volumes of text data), to be automated.

Such progress will bring new challenges to the fore, not least ethical concerns. Much of cybercrime research involves processing data created by or of individuals. This opens up issues around informed consent, anonymity, and confidentiality, and whether the data was obtained as intentionally public data, or data which has been obtained from a leak and would otherwise be private. At the infrastructure-level, there needs to be consideration in protecting the anonymity and confidentiality of collected data across time and against evolving attack techniques (Moore et al., 2019). Another concern is anonymity where conventional de-anonymisation techniques, such as removal of personal information, may no longer be sufficient. For example, Sweeney (1997) describes a case where the removal of identifiable information such as names and addresses from a hospital database does not guarantee anonymity since one’s unique history of diagnosis remains and therefore identifiable through queries. Similarly, Narayanan and Shmaikov (2008) demonstrate the effectiveness of de-anonymisation algorithm with the Netflix Prize dataset (containing movie ratings from users) by cross-referencing the anonymised data with publicly available information (e.g. the Internet Movie Database (IMDB). Moving forward, discussion on ethical considerations with the use of large datasets require an interdisciplinary insight. For a more detailed discussion about ethical issues in cybercrime research, please see Chapter XX.

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7), 16-07.

Bada, M. Chua, Y. T., Collier, B., Pete, I., (2020). Exploring masculinities and perceptions of gender in online cybercrime subcultures. In *Proceedings of the 2nd Annual Conference on the Human Factor in Cybercrime*.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the Third International ICWSM Conference* (pp. 361-362).

Benjamin, V., Li, W., Holt, T., & Chen, H. (2015). Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 85-90).

Bevensee, E., Aliapoulios, M., Dougherty, Q., Baumgartner, J., McCoy, D., & Blackburn, J. (2020). SMAT: The Social Media Analysis Toolkit. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*.

Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*, *1*(1), 1-6.

Caines, A., Pastrana, S., Hutchings, A. & Buttery, P. (2018). Automatically identifying the function and intent of posts in underground forums. Crime Science, 7(19), 1-14.

Cambridge Cybercrime Centre. (2019). *Process for working with our data*. https://www.cambridgecybercrime.uk/process.html.

Chan, J., & Moses, B.L. (2016). Is Big Data challenging criminology? *Theoretical Criminology*, *20*(1), 21-39.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171-209.

Christie, N. (1997). Four blocks against insight: Notes on the oversocialization of criminologists. *Theoretical Criminology, 1*(1), 13-23.

Collier, B., Thomas, D. R., Clayton, R., & Hutchings, A. (2019). *Booting the Booters: Evaluating the effects of police interventions in the market for denial-of-service attacks*. Proceedings of the ACM Internet Measurement Conference, Amsterdam.

Davis, C.A., Ciampaglia, G.L., Aiello, L.M., Chung, K., Conover, M.D., Ferrara, E., Flammini, A., Fox, G.C., Gao, X., Gonçalves, B., Grabowicz, P.A., Hong, K., Hui, P., McCaulay, S., McKelvey, K., Meiss, M.R., Patil, S., Peli Kankanamalage, C., Pentchev, V., Qiu, J., Ratkiewicz, J., Rudnick, A., Serrette, B., Shiralkar, P., Varol, O., Weng, L., Wu, T., Younge, A.J., & Menczer, F. (2016). OSoMe: The IUNI observatory on social media. *PeerJ Computer Science 2*:e87.

Edwards, A., Housley, W., Williams, M., Sloan, L., & Williams, M. (2013). Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology*, *16*(3), 245-260.

González‐Bailón, S. (2013). Social science in the era of big data. *Policy & Internet*, *5*(2), 147-160.

Gerritsen, C. (2020). Big data and criminology from an AI perspective. In B. Leclerc & J. Calle (Eds.), *Big Data* (pp. 29-39). Routledge.

Holt, T. J., & Dupont, B. (2019). Exploring the factors associated with rejection from a closed cybercrime community. *International Journal of Offender Therapy and Comparative Criminology*, *63*(8), 1127-1147.

Hughes, J., Aycock, S., Caines, A., Buttery, P., & Hutchings, A. (2020). Detecting trending terms in cybersecurity forum discussions. *Workshop on Noisy User-generated Text (W-NUT)*, virtual event.

Hughes, J., Collier, B., & Hutchings, A. (2019). From playing games to committing crimes: A multi-technique approach to predicting key actors on an online gaming forum. In *Proceedings of the APWG Symposium on Electronic Crime Research (eCrime)*, Pittsburgh.

Hutchings, A. & Pastrana, S. (2019). Understanding eWhoring. In Proceedings of the 4th IEEE European Symposium on Security and Privacy, Stockholm.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1-12.

Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology*, *43*, 19-39.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational Social Science. *Science (New York, NY)*, *323*(5915), 721.

Lee, J. R., & Holt, T. J. (2020). The challenges and concerns of using big data to understand cybercrime. In B. Leclerc & J. Calle (Eds.), *Big Data* (pp. 85-103). Routledge.

Li, W., Chen, H., & Nunamaker Jr, J. F. (2016). Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. *Journal of Management Information Systems*, *33*(4), 1059-1086.

Lynch, J. (2018). Not even our own facts: Criminology in the era of big data. *Criminology*, *56*(3), 437-454.

Manyika, J. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Available at *http://www. mckinsey. com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation* (Accessed August 25, 2020).

Metzler, K., Kim, D. A., Allum, N., & Denman, A. (2016). Who is doing computational social science? Trends in big data research.

Moore, T., Kenneally, E., Collett, M., & Thapa, P. (2019). Valuing Cybersecurity Research Datasets. In *18th Workshop on the Economics of Information Security (WEIS)*.

Motoyama, M., McCoy, D., Levchenko, K., Savage, S., & Voelker, G. M. (2011, November). An analysis of underground forums. In *Proceedings of the 2011 ACM SIGCOMM Internet Measurement Conference* (pp. 71-80).

Nagin, D.S. (2005). *Group-Based Modeling of Development.* Harvard University Press.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the IEEE Symposium on Security and Privacy (sp 2008)* (pp. 111-125).

Newman, G. R., & Clarke, R. V. (2003). *Superhighway Robbery: Preventing E-commerce Crime.* Cullompton: Willan.

Ozkan, T. (2019). Criminology in the age of data explosion: New directions. *The Social Science Journal*, *56*(2), 208-219.

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

Pastrana, S., Hutchings, A., Caines, A., & Buttery, P. (2018b). Characterizing Eve: Analysing cybercrime actors in a large underground forum. In *Proceedings of the 21st International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, Heraklion.

Pastrana, S., Hutchings, A., Thomas, D. R., & Tapiador, J. (2019). Measuring eWhoring. In *Proceedings of the ACM Internet Measurement Conference*, Amsterdam.

Pete, I., & Chua, Y. T. (2019). An assessment of the usability of cybercrime datasets. In *12th USENIX Workshop on Cyber Security Experimentation and Test (CSET 19)*.

Pete, I., Hughes, J., Bada, M., & Chua, Y.T. (2020). A social network analysis and comparison of six dark web forums. In *IEEE European Symposium on Security and Privacy (EuroS&PW) Workshop on Attackers and Cyber Crime Operations (WACCO)*.

Porcedda, M. G., & Wall, D. S. (2019, June). Cascade and chain effects in big data cybercrime: Lessons from the talktalk hack. In *IEEE European Symposium on Security and Privacy (EuroS&PW) Workshop on Attackers and Cyber Crime Operations (WACCO)* (pp. 443-452).

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. *The British Journal of Criminology*, *57*(2), 259-274.

Snaphaan, T., & Hardyns, W. (2019). Environmental criminology in the big data era. *European Journal of Criminology*, 1-22.

Snijders, C., Matzat, U., & Reips, U. D. (2012). "Big Data": big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, *7*(1), 1-5.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, *25*(2-3), 98-110.

Thomas, D. R., Clayton, R., & Beresford, A. R. (2017). 1000 days of UDP amplification DDoS attacks. In *Proceedings of the 2017 APWG Symposium on Electronic Crime Research (eCrime)* (pp. 79-84). IEEE.

Tuckman, B.W. (1965). Developmental sequence in small groups. *Psychological Bulletin 63*(6), 384.

Turk, K., Pastrana, S., & Collier, B. A tight scrape: methodological approaches to cybercrime research data collection in adversarial environments. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshop on Attackers and Cyber-Crime Operations (WACCO)*.

Vetterl, A., & Clayton, R. (2019). Honware: A virtual honeypot framework for capturing CPE and IoT zero days. In *Proceedings of the 2019 APWG Symposium on Electronic Crime Research (eCrime)* (pp. 1-13). IEEE.

Vu, A.V., Hughes, J., Pete, I., Collier, B., Chua, Y. T., Shumailov, I., & Hutchings, A. (2020). Turning up the dial: The evolution of a cybercrime market through set-up, stable, and COVID-19 eras. In P*roceedings of the ACM Internet Measurement Conference*, Pittsburgh.

Wang, F. Y., Carley, K. M., Zeng, D., & Mao, W. (2007). Social computing: From social informatics to social intelligence. *IEEE Intelligent Systems*, *22*(2), 79-83.

Westlake, B. G., & Bouchard, M. (2016). Liking and hyperlinking: Community detection in online child sexual exploitation networks. *Social science research*, *59*, 23-36.

Yar, M. (2005). The Novelty of 'Cybercrime' An Assessment in Light of Routine Activity Theory. *European Journal of Criminology*, *2*(4), 407-427.