https://doi.org/10.1142/9781800614079_0015

Chapter 5.1

Identifying and Collecting Public Domain Data for Tracking Cybercrime and Online Extremism

Lydia Wilson, Viet Anh Vu, Ildikó Pete and Yi Ting Chua

Abstract

Collecting and making use of publicly available data is not always straightforward, particularly for interdisciplinary researchers who often lack skills to deal with technical issues that arise during the process. This chapter gives an overview of the challenges involved in identifying and collecting materials, and outlines a general technical framework for building effective and sustainable computer programmes to scrape, process and store online open source materials into structured datasets for research purposes. We also discuss the data licensing process, which is essential for experiment reproducibility, along with ethical considerations when working with the data to protect both researchers and the general population. We demonstrate, as a case study, how we collect and handle cybercrime and extremist resources at the Cambridge Cybercrime Centre – an interdisciplinary initiative combining diverse expertise at the University of Cambridge.

Introduction

Vast amounts of data are now publicly available online, free to download and store. This might suggest that we are working in a golden age of open source research,[1] but in fact there are numerous barriers to overcome –

[1] In line with most of the rest of this book, we use 'open source' in the social science sense of freely available data rather than the computer science concept of intellectual property and reusability.

technical, ethical and analytical – before using such data to carry out robust studies. This chapter shows the process from identification of material to interpretation of data, taking as a case study the process of creating databases of cybercrime and extremist content at the Cambridge Cybercrime Centre (CCC), within the University of Cambridge's Department of Computer Science and Technology.[2]

The CCC has been collecting data on cybercrime since 2015, from data traces of DDoS attacks[3,4] to scraping[5] conversations on underground forums discussing crimes, such as hacking and illicit marketplaces. These forums form the basis of CrimeBB, a large-scale dataset consisting of more than 99M posts and 11.6M threads made by over 4.6M users on 34 cybercrime forums in 5 different languages, English, Russian, German, Arabic and Spanish.[6] In 2019, the group expanded its collection to include extremist material, starting a new structured dataset, ExtremeBB, for this content.[7] Areas of focus have been extremist ideologies including white supremacy, manosphere such as incels (involuntary celibates) and lookism,[8] and online forums dedicated to trolling and doxxing.[9] Scraping has been expanded to collect data on far-right ideologies more broadly, and in 2021, jihadi material started to be added. As of April 2022, ExtremeBB contains nearly 48M posts in 3.5M threads from more than 390K active members on 12 extremist forums. These forums are scraped – and results are systematically processed and stored – on an ongoing basis, with the long-term aim of providing data to researchers looking at online extremism in the early-mid 21st century. As of the writing date (2022), access to ExtremeBB has been

[2]The Cambridge Cybercrime Centre [online]. Available from: https://www.cambridgecybercrime.uk/ [Accessed 12 July 2023].

[3]DDoS stands for Distributed Denial-of-Service, a type of attack on computer systems that makes them unavailable to intended users.

[4]Thomas D. R., Clayton R., and Beresford A. R. 1000 Days of UDP Amplification DDoS Attacks. *Proceedings of the IEEE Symposium on Electronic Crime Research (eCrime)*. 2017, pp. 79–84.

[5]The process of automatically collecting and extracting data from a website.

[6]Pastrana S., Thomas D. R., Hutchings A., and Clayton R. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale. *Proceedings of the World Wide Web Conference (WWW)*. 2018, pp. 1845–1854.

[7]Vu A. V., Wilson L., Chua Y. T., Shumailov I., and Anderson R. ExtremeBB: A Database for Large-Scale Research into Online Hate, Harassment, the Manosphere and Extremism. *The ACL Workshop on Online Abuse and Harms (WOAH)*. 2023.

[8]The term refers to techniques for enhancing men's physical attractiveness to women.

[9]The action of digging out and publishing information to expose identities, or finding personal, and previously private, information such as addresses to threaten individuals.

granted for 39 researchers in 12 groups from 10 universities and institutions around the world (excluding the team at Cambridge), while the figures are 170, 50 and 39 for CrimeBB, respectively.

There were various motivations for building these databases. First, many people do not have the skills to collect big datasets, putting such activities out of reach for many non-technical researchers. If people do have the necessary skills, it is still time-consuming for them to build datasets. Using large-data approaches without a pre-existing dataset would be impossible for a year-long MSc project, for example, but if data have already been collected, a researcher can bypass the collection step and focus on analysis. Further, such datasets are in general not widely available, making it difficult for others to check results, or for single teams to interrogate them via different techniques and analytical tools in order to compare research methods. Finally, complete longitudinal datasets are valuable for spotting how trends emerge and change over time, which is at odds with the current academic model of project-based funding. The databases developed by the CCC resist such short-term pressures, and will be useful to the wider academic community now and in the future.

This chapter shows the process of building the cybercrime and extremist databases, which are made freely available to researchers (subject to agreements to prevent misuse), and the further steps necessary to interpret the data. We start by considering the many ethical issues that need to be addressed for any work in this area. The chapter then broadly follows Jagadish *et al.*'s five steps in big data usage: 'acquisition, information extraction and cleaning, data integration, modelling and analysis, and interpretation and deployment',[10] describing our data identification, collection and storage methods, and discussing technical challenges and our processes for overcoming these. We then look at how the data are processed and made available to the research community, first by cleaning the data, and then providing the tools and expertise for a non-technical researcher to interrogate them. Finally, we look at interpreting the data, and demonstrate how interdisciplinary research works best for this work. Throughout the chapter, we show the complexities and possibilities of big data research. We encourage

[10] Jagadish H. V., Gehrke J., Labrinidis A., Papakonstantinou Y., Patel J. M., Ramakrishnan R., and Shahabi C. Big data and its technical challenges. *Communications of the ACM*. 2014, 57 (7), pp. 86–94.

interdisciplinary work to better understand the problem of online extremism and cybercrime in our societies.

Ethical Considerations

Discussions on ethical considerations and impacts from such data use are increasingly relevant.[11] Fundamentally, there is a balance to be struck between expectations of privacy on the side of users and the valuable information and understanding that can be gained from the research. There are no clear-cut guidelines to achieve this balance, as different contexts bring different considerations of potential harm and risk. Discussions on ethics can be broadly categorized into two groups for most research: (1) the general population and research subjects, and (2) the researchers themselves.

Ethical considerations: General population and research subjects

Although traditionally considered separately, the distinction between a general population and research subjects is increasingly blurred due to the nature of open source big data. For all research on people, informed consent is an unavoidable topic. In general, academic best practice stipulates that research subjects must voluntarily agree to participate, having been given details of the research, including its purpose, any potential risks associated with taking part and participants' rights.[12] However, under specific conditions, informed consent is not required. These conditions include (a) the use of secondary data where research subjects are in life-threatening situations where interventions by researchers are necessary before any possible consent; (b) circumstances in which research subjects cannot be identified; (c) cases where the research cannot be practically conducted with consent; and (d) times when the research poses no more than minimal risks to research subjects.[13]

[11] Other chapters in this volume consider ethical dilemmas within open source research and how practitioners approach these. See, for example, the chapters by Wilson, Samuel & Plesch, Duke, Freeman & Koenig, Ahmad, Michie *et al.*, and Bedenko & Bellish (Chapters 1, 2.2, 2.5, 3.1, 5.3 and 5.4).

[12] Bachman R. D., Schutt R. K., and Plass P. S. *Fundamentals of Research in Criminology and Criminal Justice: With Selected Readings*. Newbury Park CA: 2016.

[13] *ibid.*

With open source data, the issue of informed consent is further complicated by the public versus private nature of the data sources. Some argue that online platforms are publicly accessible and thus data collection without consent can be justified.[14] This view is reasonable for research that primarily involves observation only. However, researchers also need to take members' perceptions of the selected online community into account. For some communities, members may consider their publicly accessible postings to be private while other communities welcome the sharing of personal information.[15] Unless the research is conducted with care, research subjects may feel that their rights are being violated, which might prompt them to move towards more private or closed platforms, and thereby distort the composition and characteristics of the groups that they leave.

Leaked datasets, especially those containing classified data, also raise ethical dilemmas.[16] Some argue that it is ethical to use such datasets once they have been leaked. However, if leaked datasets provide access to otherwise personal data (e.g. mental health records), their use can harm individuals. For much publicly available data, additional measures are necessary to ensure the anonymity and confidentiality of individuals in the datasets, since secondary data analyses could compromise these aspects.[17] A de-anonymization algorithm can reveal personal information, as Narayanan and Shmatikov demonstrated[18] when they cross-referenced an anonymized Netflix Prize dataset with publicly available information, e.g. the Internet Movie Database (IMDB).

The first step to address these ethical considerations is to appoint a Review Board to ensure that human subjects are protected, and to help

[14] Holt T. J. Exploring Strategies for Qualitative Criminological and Criminal Justice Inquiry using On-Line Data. *Journal of Criminal Justice Education*. 2010, 21, p. 466.

[15] Garcia A. C., Standlee A. I., Bechkoff J., and Cui Y. Ethnographic Approaches to the Internet and Computer-Mediated Communication. *Journal of Contemporary Ethnography*. 2009, 38 (1), pp. 52–84.

[16] Thomas D. R., Pastrana S., Hutchings A., Clayton R., and Beresford A. R. Ethical Issues in Research Using Datasets of Illicit Origin. *Proceedings of the ACM Internet Measurement Conference (IMC)*. 2017, pp. 445–462.

[17] *ibid.*

[18] Narayanan A. and Shmatikov V. Robust De-anonymization of Large Sparse Datasets. *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*. 2008, pp. 111–125.

devise systems that mitigate potential harms to subjects,[19] (see also footnote 14), including, for example, refraining from naming particular websites to help ensure the anonymity of research subjects[20,21] (see also footnote 6). Minimizing and/or avoiding the use of long quotes can also lower the traceability of users and thus protect the participants' anonymity. Ultimately, when using open source data, researchers always need to consider ethical issues, especially with regard to potential harms, while weighing these against potential benefits (see footnote 16).

Ethical considerations: Researchers

In addition to research subjects, researchers also need to consider, and protect, their own safety. Researchers examining violent extremism and cybercrime have a higher chance of witnessing, encountering and/or being asked to participate in illegal or criminal activities. For example, interviewees in a study by Holt and Copes[22] were asked about their intellectual property violations (e.g. illegal media downloads), which meant that the authors had knowledge of interviewees' illegal behaviour. In addition, the fieldworker for the study had to actively participate in the forums dedicated to intellectual property violations and demonstrate her knowledge in order to gain the trust of other forum members. In the example of extremism research, downloading terrorist content can be a crime, and it is essential that researchers engage with their research institutions to make sure that they have suitable protections.

Although there is wide variation in the severity and ramifications of crimes – e.g. drug offences and illegally downloaded material have different impacts and victim footprints – researchers exploring any illegal actions need to decide whether, when and how to report their findings to law enforcement authorities. Researchers may be tempted to report anything considered

[19] Franklin J., Perrig A., Paxson V., and Savage S. An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants. *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*. 2007, pp. 375–388.
[20] *ibid.*
[21] Holt T. J. and Copes H. Transferring Subcultural Knowledge On-line: Practices and Beliefs of Persistent Digital Pirates. *Deviant Behavior*. 2010, 31 (7), pp. 625–654.
[22] *ibid*.

a crime, but this might actually not be the right approach for several reasons. For example, in the case of terrorist content, intelligence agencies may be monitoring the same activities and might not want the extra burden of responding to researchers. For common crimes, local police may not have the capacity to pursue every report of minor infractions. Meanwhile, the relevant laws vary from country to country, and so researchers' decisions on whether to report will be affected by the jurisdiction they are based in. In many jurisdictions there is no legal obligation to report particular crimes, and so decisions must be made on a case-by-case basis.

One oft-posited rule is to report criminal behaviours when researchers have knowledge of a serious crime in which innocent third parties can be harmed, although this rule is rarely followed in reality.[23] For offline studies, researchers can tell research subjects to refrain from discussing illegal activity, and clarify reasons for this. But for open source intelligence where there is no direct interaction with research subjects, researchers need to establish guidelines and rules about when and what to report before data collection and analyses.

To protect themselves, researchers should also consider certain measures for the hardware and software used for research. These technologies could be compromised when visiting sites of online groups and communities that are infected with malicious software (see footnote 14). Such malware could also risk the anonymity of research subjects, if sensitive information is stored on affected devices. To minimize such occurrences, researchers should use different computers for storing and analyzing data.

In the context of our work compiling the CCC datasets, these ethical issues have been explicitly addressed. While discussing the CrimeBB dataset, Pastrana and colleagues (see footnote 6) delve into the ethical considerations of using web crawlers[24] to collect forum data, noting the challenges of breaking terms and conditions, bypassing CAPTCHA[25] and working to ensure that the research does not harm individuals. Before

[23] Sandberg S. and Copes H. Speaking with Ethnographers: The Challenges of Researching Drug Dealers and Offenders. *Journal of Drug Issues*. 2013, 43 (2), pp. 176–197.

[24] A bot that systematically trawls the internet.

[25] CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a challenge to determine whether the actor interacting with the content is a real human, for example, by requiring them to recognize distorted text. In the rest of the chapter, we will use the lowercase term 'Captcha' for ease of reading.

deploying their web crawlers, the researchers submitted an ethics application, seeking permission for their work from the departmental Review Board. In the ethics application for the ExtremeBB dataset, harms to researchers were addressed by plans to hold regular meetings on these matters and to follow institutional guidelines. Information about institutional resources such as counseling services is also made available to researchers, to protect researchers by minimizing the effects of working closely with potential violent and extreme content.

Material Identification

The increased integration of technology into society has resulted in the creation of tremendous amounts of digital data. These range from user-generated content to personal identifiable information, in the form of websites, e-mails, blogs, forums, instant messaging, social media, and accounts on services such as Netflix[26–28] (see also footnotes 14 and 18). These data sources allow researchers to observe and examine behaviours and attitudes of online underground communities across platforms and over time[29] (see also footnote 6). This section provides an overview of the methodological challenges this increase in data presents.

One major issue is identifying and selecting representative data from the ocean of available sources.[30] For example, researchers need to determine whether the conclusions derived from one platform are applicable to another platform dedicated to the same topic, or consider the generalizability of results derived from a subset of the whole platform population. The issue is further complicated by the reach of the internet, as there may be differences in digital rights and uses based on cultures, geographic locations, legal

[26]Burrows R. and Savage M. After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*. 2014, 1 (1). Available from: https://doi.org/10.1177/2053951714540280.

[27]Lazer D. and Radford J. Data ex Machina: Introduction to Big Data. *Annual Review of Sociology*. 2017, 43, pp. 19–39.

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

[29]Bada M., Chua Y. T., Collier B., and Pete I. Exploring Masculinities and Perceptions of Gender in Online Cybercrime Subcultures. *Cybercrime in Context: The Human Factor in Victimization, Offending, and Policing*. 2021, pp. 237–257.

[30]Hughes J., Chua Y. T., and Hutchings A. Too Much Data? Opportunities and Challenges of Large Datasets and Cybercrime. *Researching Cybercrimes*. 2021, pp. 191–212.

provisions, and/or languages (see footnote 14). Another big challenge is the rapid migration of online communities, which could be a result of de-platforming or attempts to evade attention. A famous example is the de-platforming of Parler in January 2021 when it lost its hosting service, and was removed from Apple and Google.[31] This resulted in the migration of users from Parler to other well-known platforms that emphasized free speech, such as Gab.[32]

The CCC's recent effort in creating ExtremeBB illustrates the challenges presented by such migrations. We collect data on all aspects of extremism, irrespective of the direct research interests of the team, with the expectation that the resource will open up new avenues of research in the future. We began collecting data from extremist forums in 2019, starting with a range of sites with far-right ideologies, and in 2021 expanded this to include extremist Islamist data. The resulting database therefore comprises a wide range of research material, lumped together under the common rubric of 'extremism'. It is likely that researchers will use the data to assess one ideology at a time, although the database also allows for 'compare and contrast' analyses, which may yield some interesting results.

In collecting data from far-right online communities, the CCC drew on in-house expertise to compile a list of known sites, which is added to regularly. However, a different approach was needed to collect data on extreme Islamist communities, as these tend to be more fractured and shorter-lived because multiple actions are taken against them. Accordingly, the CCC found that experts who constantly monitor Islamist online communities were needed. These experts had different views about CCC's data collection and storage plans. Many welcomed the initiative, and began to collect and send on sources, beginning with Telegram and Discord channels that the CCC systems then scraped. Others, who were manually collecting material

[31] Fung B. Parler has now been booted by Amazon, Apple and Google. *CNN Business*. 11 January 2021. Available from: https://edition.cnn.com/2021/01/09/tech/parler-suspended-apple-app-store/index.html [Accessed 12 May 2022].

[32] Ray S. The Far-Right Is Flocking To These Alternate Social Media Apps — Not All Of Them Are Thrilled. *Forbes*. 14 January 2021. Available from: https://www.forbes.com/sites/siladityaray/2021/01/14/the-far-right-is-flocking-to-these-alternate-social-media-apps—not-all-of-them-are-thrilled/?sh=3c2cf25655a4 [Accessed 13 July 2023].

from official channels, preferred to stick to their own approach. Yet others were automatically collecting data but couldn't share for proprietary reasons.

With the help of some external experts, the ExtremeBB database is slowly building up, and a useful feedback loop has been created, whereby the team scraping the data can alert experts when channels are closed down. This type of collection requires continuous attention given the speed of emerging sub-groups of supporters on a variety of different platforms.

Data Collection

Open source, public domain data are in many cases readily available on the internet and require no special privilege to access (subject to local jurisdictions). Gathering these data at scale in the long term on a sustainable basis, however, can be tricky and time-consuming as most web administrators do not intentionally offer Application Programming Interfaces (APIs) that allow researchers to fetch data directly from their servers. Additionally, some data sources are only available through a special access mechanism; for example, using Tor[33] with an anonymous communication channel is a prerequisite to enter hidden websites.

However, such websites are basically still *public* and data can be collected from them in one way or another, albeit a manual collection process might take months or even years to complete. Although some data only need to be downloaded once (e.g. some documents or images), others are compiled continuously over time (e.g. chat and forum discussions) and thus require a long-term collection plan. Similarly, while a number of data sources are just a single file (e.g. a database) which can be downloaded by just one click, some are large, not well structured and non-trivial to gather (e.g. forum chats). In such cases, manual approaches would never be able to capture a useful sample of material.

Computer programmes can help to automate the processes of fetching, extracting, parsing and storing data, including, for example, web scrapers. Despite the existence of protection mechanisms on some websites, which

[33] Dingledine R., Mathewson N., and Syverson P. *Tor: The Second-Generation Onion Router*. Technical Report, Naval Research Lab Washington DC. 2004.

may significantly slow down and restrict the access of such automated bots, as long as the data are still public and can be seen by humans, they can also be seen by bots. This section outlines some of the technical challenges involved and suggests a general framework to build sustainable and efficient computer bots to automate data collection.

Challenges

To a technician, building a web scraper sounds simple and obvious: It involves using a web driver (most popularly Selenium and Puppeteer)[34] to access an identified webpage, then find and save relevant content. However, it may not be that straightforward in practice. The automated bot often mimics human behaviour by clicking and viewing the webpage, which, at scale, may generate a significant amount of requests towards the targeted server. The increased traffic may thus attract attention and be detected by administrators, who tend to protect their data from being crawled.

As a result, websites often adopt anti-crawling protection mechanisms, such as: limiting the number of requests clients can send within a time period; using Captcha to prevent bots; blacklisting suspicious IP addresses, a range of IP addresses or the whole Autonomous System hosting these IPs; and more sophisticated techniques such as measuring timing between clicks and introducing non-visible malicious links to trap automated bots. Some websites also use DDoS protection mechanisms, limit sensitive content for registered users only, and require a paid (or reputed) account to access.[35] More sophisticated protections are offered by third-party providers (e.g. Cloudflare and DDoS Guard)[36] to block suspicious traffic, such as bot actions and DDoS attacks. These can detect and block web scrapers, for example, through infinite Captcha attempts.

While many defences can be bypassed with ease and do not impact scraping tools, some combined techniques may effectively slow down web

[34] Selenium [online]. Available from: https://www.selenium.dev/ [Accessed 13 July 2023]. Puppeteer [online]. Available from: https://pptr.dev/ [Accessed 13 July 2023].

[35] Benjamin V., Samtani S., and Chen H. Conducting large-scale analyses of underground hacker communities. *Cybercrime Through an Interdisciplinary Lens*. 2016, pp. 26, 56.

[36] Cloudflare [online]. Available from: https://www.cloudflare.com [Accessed 13 July 2023]. DDoS Guard [online]. Available from: https://ddos-guard.net [Accessed 13 July 2023].

crawlers.[37] It is thus challenging to make automated scrapers stealthy, in the sense that they can mimic human behaviour to avoid being detected, while still being effective, sustainable and not causing negative consequences e.g. bandwidth congestion or a denial of services. Some sites often update their HTML structures (e.g. changing theme, adding new features or switching to new frameworks), thus the scrapers may need to be tailored regularly. Some data are short-lived, for example, chat channels and forum threads that only appear for a short period before being permanently deleted, and so real-time collection is sometimes necessary.

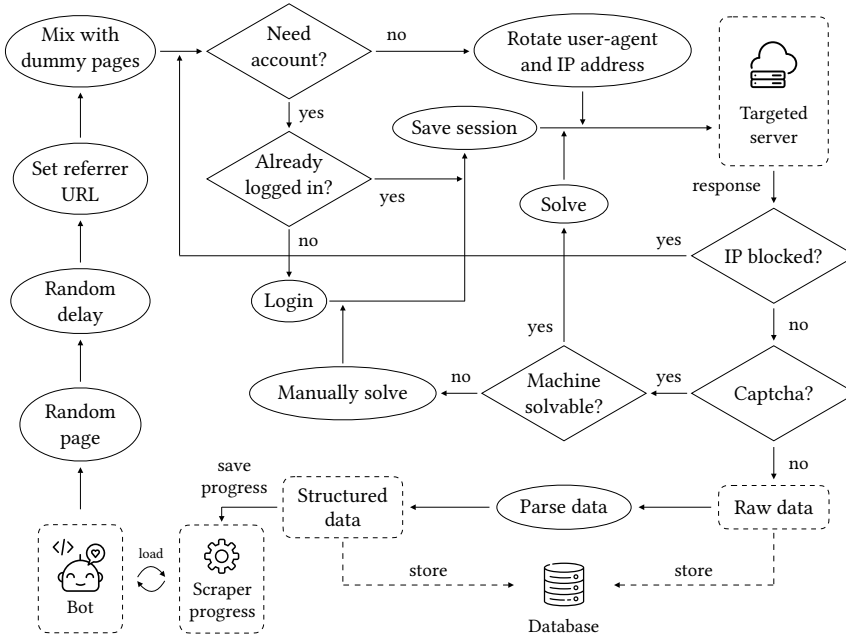

Figure 1: A general framework to build web scrapers for open source data collection.

[37] Turk K., Pastrana S., and Collier B. A tight scrape: methodological approaches to cybercrime research data collection in adversarial environments. *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. 2020, pp. 428–437.

Technical solutions

Prior work has introduced some automated bot architectures for scraping online forums (see footnote 6). Here, we outline a more general framework to develop a web scraper on public domain sources, as depicted in Fig. 1. We do not describe technical details such as which programming languages and which programming libraries should be used, but instead overview some essential rules to bear in mind in order to make a scraper effective and appear natural. Note that the scrapers should be designed with ethical considerations in mind, as discussed above: The aim is not to flood targeted servers to gather the desired data as quickly as possible, but to automate tedious data collection.

First of all, it is necessary to make sure that the bot is set up with the correct web driver and communication tunnel. For instance, if the targeted website is only available on hidden webs (typically found with .onion domain), using a Tor browser with an anonymous communication channel (or setting up an onion routing)[38] is to be expected. Second, although timezone information is critical for longitudinal analyses, many websites do not clearly specify such information. Some display the date and time dynamically corresponding to the location of users, while others just show a fixed timezone. Thus, before starting the collection, manual effort is needed to figure out the actual timezone of the targeted websites.

When the bot starts, it chooses a page P of the targeted website to send requests to. Suppose that we need to visit a number of targeted pages, hitting them in a fixed order one by one (e.g. timestamp descending) is not a good idea as this would look like a robot's behaviour. Instead, P should be picked at random. After identifying P, the next important step is to add a random delay. The delay should never be a constant, as this could reveal repeated timing patterns that are easily detected. Then, a referrer URL should be set, which indicates the link that the bot has visited right before accessing P. This should look as genuine as possible; for example, by setting it to google.com, it will look like P has been discovered through a search and not by direct access. For different requests, the referrer URL should be also rotated periodically and appropriately; for example, setting

[38] Reed M. G., Syverson P. F., and Goldschlag D. M. Anonymous connections and onion routing. *IEEE Journal on Selected Areas in Communications*. 1998, 16 (4), pp. 482–494.

it to the homepage or navigation URLs of the targeted website would be good choices. Next, it is important to mix P with a number of non-targeted pages, for example, by navigating to the site's homepage, clicking on some random adverts or unrelated URLs and then eventually visiting P. This will potentially hide the actual intention of the bot and thus mitigate the chance of being detected.

The original IP address of scrapers can be mapped to a network and geographic location. Using a generic proxy service or setting up dedicated cloud-based servers to work as Virtual Private Networks (VPNs) will reduce the likelihood of the bot being linked to the researcher's institution. Similarly, when the targeted website does not require a registered account, it is beneficial to rotate the user-agent[39] and IP address frequently and randomly (sometimes, rotating the browser window's size is also recommended as this can be used to track users' behaviour). They should, however, be rotated together, as using the same user-agent from a particular IP for a long period makes the bot look more like a human. The rotation should not be done for every request but at an appropriate (and also, random) rate. If the targeted website requires logging in with a registered account, it is important to *not* rotate the IP as well as the user-agent for every request because it is unusual for one account to engage with many different IPs and user-agents. In this case, multiple accounts should be created to log into the site, so that each account can collect a subset of the targeted URLs, which should be divided randomly. For each account, the IP address and user-agent should remain constant throughout the collection process. IP addresses should be chosen from different geolocations, hosted by different Autonomous Systems (AS) and Internet Service Providers (ISP). After successfully logging in, the scraper should save the logged session by appropriate means (typically using cookies)[40] to make sure that the bot will not be asked to login again. Once this process has been done, the request to P is sent to the targeted server.

[39] User-agent is an intermediate software between servers and end-users helping them interact with websites. Not rotating the user-agent identification for a long period may lead to the bot being detected. A list of popular user-agents is available from a quick search, and it is better to use the common ones.

[40] Cookie, or HTTP cookie, is a small piece of data stored in the client to identify the session of a user accessing a website, which tells the web server who is using the service so that the user does not need to log in repeatedly. It also helps web servers deliver personalized content better, as it knows who is accessing the websites.

Servers' responses may vary. If an IP address is banned, another IP and user-agent should be chosen to resend the request and the blocked IP should not be used again (if the site asks users to log in, also rotate the account). Even when an IP address is allowed, the server may require Captcha solving to determine if the request has been made by robots. Captcha is perhaps the most challenging protection to bypass. While some Captchas are rather simple and can be cracked by modern machine learning algorithms (e.g. distorted text Captcha),[41] others are more challenging and typically hard for machines to bypass (e.g. reCaptcha, hCaptcha). Some Captcha solving services are available for a small fee;[42] however, one possible approach worth trying is using cookies (if the site allows) by (1) manually solving the Captcha, (2) saving the cookie session and then (3) attaching it back to subsequent requests. Some sites adopt a third-party DDoS defence layer (popularly Cloudflare and DDoS Guard), which requires additional effort to bypass. However, the same strategy as bypassing Captcha, plus incorporating a long enough delay (to wait for the DDoS check), may be effective to address this. Some third-party providers have started to offer sophisticated mechanisms to detect bot traffic, which aim at distinguishing 'good bots' (e.g. Google's bots for web indexing) and 'bad bots' (e.g. bots spying and stealing data for commercial uses). Fortunately, if research involves 'good bots', it is possible to get the bot's IP addresses whitelisted by contacting the third-party providers and explaining the research purposes.

After completing these steps, raw data are fetched from P. In any situation, a copy of the raw data (in HTML or other formats) should be stored locally. The scale of the data collection means that it is impossible to anticipate the fetched layout of P in advance, which makes data parsing (processing the raw data to store it in a format that is more readable and therefore suitable for analysis) cause unexpected errors. It is also necessary to parse the content offline later, if further information is required. In such cases, it will be critically useful to avoid sending a bunch of requests again, which

[41] Ye G., Tang Z., Fang D., Zhu Z., Feng Y., Xu P., Chen X., and Wang Z. Yet Another Text Captcha Solver: A Generative Adversarial Network Based Approach. *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*. 2018, pp. 332–348.

[42] Motoyama M., Levchenko K., Kanich C., McCoy D., Voelker G. M., and Savage S. Re: Captchas-Understanding Captcha-Solving Services in an Economic Context. *Proceedings of the USENIX Security Symposium (USENIX Security)*. 2010, pp. 435–462.

takes time, causes unnecessary traffic towards the website and may increase the chance of the scraper being detected. After successful parsing, the structured data are stored in a database (or other system). Finally, the progress of the scraper should be recorded and persistently stored so that it can resume from the broken point if unexpected incidents happen, for example, the bot crashes, the internet drops out or there are other server errors. The scraper then repeats with a new page P' – normally, the next randomly selected page. It is always worth noting that each single page should be visited only once, and only what is exactly needed should be collected, to prevent flooding the targeted server with unnecessary traffic.

Once the scraper is running smoothly, setting a low rate limit (viz. a small number of requests per hour) is recommended to keep it indistinguishable from human users. This again helps ensure that the scraper will not be detected and the target site's administrators thus will not adopt additional protection layers or change their access policy which can impede or prevent data collection, such as making some of the site only available to 'premium' accounts. Boosting the processes by crawling in parallel may be feasible in some cases, but a completely different user profile and browser settings should be used for each bot. The choice of request rate heavily depends on how large the traffic of the targeted website is; thus, it is worth looking at the website traffic before increasing the rate limit or running bots in parallel. A rule of thumb is to build a scraper that keeps you under the radar, be patient and not greedy!

Data Usability

Data licensing and accessibility

A key component of research is reproducibility. When research is open, and based on open data, it can be interrogated by other scholars who can check conclusions, and thereby build confidence in findings and make the process more robust. Further, sharing collected data can be useful for multidisciplinary studies addressing different aspects of a problem. Moreover, making available the data collected through automated processes can mitigate obstacles faced by social scientists who often lack the technical backgrounds to build such tools themselves, enabling them to concentrate on analyzing the data in line with their own expertise.

The Cambridge Cybercrime Centre has robust ethical procedures to deal with scraped data that may contain sensitive personal information, and long experience in making such data available across multiple jurisdictions including the USA, the EU and China. Access is given to a data-sharing web platform from which authorized users can download the shared datasets. Along with the platform, we also provide detailed instructions on how to import and make use of the data to ensure that researchers from different backgrounds can get started with ease.

Before access is granted, users are required to complete legal paperwork to protect the data from misuse. The licensing regime was carefully developed in conjunction with legal academics, university lawyers and specialist external counsel. Once licensed, access to the most recent data snapshots will be automatically granted as they are published, without any further action by the licensee. The agreement includes requirements to inform the CCC about publications that draw on the data, and who is accessing them.

Data use by non-technical researchers

Open source data collection and sharing can ease the collective burden of identifying and gathering datasets, and save months or even years of researcher time.[43] Online cybercrime and extremist forums in particular lend themselves to collaborative, interdisciplinary research. However, translating the high-level research questions and the desired goals of research projects into actionable steps is a non-trivial task, given the complexity and size of the datasets alongside the technical requirements needed to work with them. The characteristics of the forum data render manual analysis infeasible, and often necessitate the application of data science methods and tools.[44–47]

[43] See also the chapter by Withorne in this volume (Chapter 5.2).

[44] Pastrana S., Hutchings A., Caines A., and Buttery P. Characterizing Eve: Analysing Cybercrime Actors in a Large Underground Forum. *Proceedings of Research in Attacks, Intrusions, and Defenses – 21st International Symposium (RAID)*. 2018, vol. 11050, pp. 207–227.

[45] Motoyama M., McCoy D., Levchenko K., Savage S., and Voelker G. M. An analysis of underground forums. *Proceedings of the ACM Internet Measurement Conference (IMC)*. 2011, pp. 71–80.

[46] Caines A., Pastrana S., Hutchings A., and Buttery P. J. Automatically Identifying the Function and Intent of Posts in Underground Forums. *Crime Science*. 2018, 7 (1), pp. 1–14.

[47] Portnoff R. S., Afroz S., Durrett G., Kummerfeld J. K., T. Berg-Kirkpatrick, McCoy D., Levchenko K., and Paxson V. Tools for Automated Analysis of Cybercriminal Markets. *Proceedings of the World Wide Web Conference (WWW)*. 2017, pp. 657–666.

The most immediate aspect that might pose an impediment to analyzing such data is size, exacerbated for researchers from non-technical backgrounds. We surveyed existing users of CCC's datasets to understand this aspect; they reported technical challenges with data exploration and download prior to analysis.[48] Thus, research aimed at automating the overall process or individual steps of data analysis, for example to identify posts on cybercrime forums related to transactions, is a highly valuable contribution for both technical and non-technical scholars (see footnote 47).

Stemming from these insights and the desire to develop tools for interdisciplinary analysis of underground forums, the Cybercrime-NLP (CC-NLP)[49] project was created. One of the aims of CC-NLP is to develop a web application, PostCog, that provides a user interface allowing licensees to explore CrimeBB and ExtremeBB with ease and without the need to develop substantial new technical skills.[50] To support longitudinal data analysis and understand underground forums at scale while taking into account the unique characteristics of the language used and interactions taking place on these forums, CC-NLP aims to create tools to allow automatic analysis of posts in the datasets. These tools, which will be integrated with PostCog, will provide answers to questions around generalizability, and will contribute to social scientists being able to discover useful and interesting themes within the data. Finally, the project involves engaging with research communities in various disciplines to understand and address their data analysis needs.

Data preparation

A necessary step for using open source datasets, like those offered by CCC, is data preparation, regardless of methodology to be used to analyze it (e.g. quantitative versus qualitative). In order to apply quantitative techniques such as natural language processing and machine learning, the datasets require further preparation. For example, in the study of masculinity and hacker forums by Bada and colleagues (see footnote 29), several preparatory steps were performed on posts extracted from an underground hacking

[48] Pete I. and Chua Y. T. An Assessment of the Usability of Cybercrime Datasets. *Proceedings of the USENIX Workshop on Cyber Security Experimentation and Test, (CSET)*. 2019.

[49] The project title refers to the application of Natural Language Processing techniques.

[50] Pete I., Hughes J., Caines A., Vu A. V., Gupta H., Hutchings A., Anderson R., and Buttery P. PostCog: A tool for interdisciplinary research into underground forums at scale. *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. 2022.

forum, before analyzing the dataset with a natural language processing programme. These steps included (a) removing unique content such as quotations, website links, programming codes, images or references, (b) removing 'normal' content such as stop words (a, the, of, in, etc.), numbers and punctuation, (c) removing capitalization of words, (d) text lemmatization, and (e) converting text into tokens (smaller chunks than the whole posts extracted from the dataset).

For qualitative methodologies and techniques, it is also necessary to adjust the data in order to efficiently and feasibly perform data analysis. An obstacle for qualitative methodologies is often the time required to perform in-depth analysis on large volumes of data. One possible solution is to incorporate sampling techniques. For example, while performing a modified grounded theory approach to identify key gender-related concepts in the underground hacker forum, Bada and colleagues (see footnote 29) included new samples of posts at each stage of coding to ensure that the categories identified at these points are consistently found throughout the data.

Data interpretation

Researchers can also encounter challenges when interpreting the output of analyses. Given the ongoing debate in the social sciences between quantitative and qualitative approaches,[51] there needs to be a shift in discussions towards how these methodologies are complementary and not mutually exclusive, especially with the emergence of online and open source data as well as improved technical knowledge and tools.

Here, we give examples of how such a mixed methods approach has been applied to research using underground forum data. In the paper mentioned above, Bada and colleagues (see footnote 29) applied both data science and qualitative approaches to examine the construct of masculinity and its relationship to the hacker subculture. Through the use of natural language processing techniques, the authors performed an exploratory analysis of the entire sample, which consisted of more than 490,000 posts. The outputs were then compared with the qualitative results derived from a

[51] Buckler K. The Quantitative/Qualitative Divide Revisited: A Study of Published Research, Doctoral Program Curricula, and Journal Editor Perceptions. *Journal of Criminal Justice Education*. 2008, 19 (3), pp. 383–403.

modified grounded theory approach. When comparing findings from both approaches, the authors discovered overlaps in perceptions of gender, as well as how gender is discussed in a specific context, such as social engineering.

Quantitative methods can also be accelerated by qualitative theories to provide insights into the way online communities (often considered as groups of users) are established and develop over time. This method has been used in our recent work on the evolution of a cybercrime marketplace, in particular how it responded to the COVID-19 pandemic.[52] The findings suggested a stimulus of trading activities in this marketplace at that time, explained by the fact that people spent more time online during lockdowns due to missing school or being jobless, and faced the boredom of being confined to their room.

Conclusion

This chapter has presented an automatic data collection framework for cybercrime and extremist resources, and ways to maximize their use. With appropriate care, the vast amounts of data that are freely available online can be used by researchers in a multitude of disciplines to answer many questions about online crime and extremism. Collaboration is key, and not just because of the number of skills required for the separate stages of data identification, collection, sharing and interpretation, but also because without sharing datasets it is impossible to replicate research, a cornerstone of scientific activity. Open source research is at its heart a shared endeavour, and datasets from the Cambridge Cybercrime Centre contribute to this, providing resources that can be used both for original research and also to verify the findings of others. It is an ongoing process, requiring constant attention to keep the sources and tools updated – one of enormous value to a wide research community.

[52] Vu A. V., Hughes J., Pete I., Collier B., Chua Y. T., Shumailov I., and Hutchings A. Turning Up the Dial: The Evolution of a Cybercrime Market Through Set-up, Stable, and COVID-19 Eras. *Proceedings of the ACM Internet Measurement Conference (IMC)*. 2020, pp. 551–566.

Acknowledgement

We are grateful to Richard Clayton and our colleagues at the Cambridge Cybercrime Centre for their useful feedback and valuable comments on an early draft of this chapter. Icons used in the figures are designed and provided for free by Freepik, Darius Dan, Vitaly Gorbachev and Pixel Perfect.