An Assessment of the Usability of Cybercrime Datasets

Ildiko Pete
Cambridge Cybercrime Centre, University of Cambridge
ip358[@]cam.ac.uk

Yi Ting Chua ‡
Cambridge Cybercrime Centre, University of Cambridge
ytc36[@]cam.ac.uk

‡*Equal Contribution*

Abstract

Cybersecurity datasets play a vital role in cybersecurity research. Following the identification of potential cybersecurity datasets, researchers access and manipulate data in the selected datasets. This work aims to identify potential usability issues associated with dataset access and data manipulation through a case study of the data sharing process at the Cambridge Cybercrime Centre (CCC). We collect survey response from current users of the datasets offered by the CCC, and apply a thematic analysis approach to identify obstacles in the uptake of these datasets, and areas of improvement in the data sharing process. The identified themes suggest that users' level of technological competence, including previous experiences with other datasets, facilitate the uptake of CCC's datasets. Additionally, users' experiences with different stages of the data sharing process, such as dataset usage and its various aspects, including downloading and setting up the dataset, highlight areas of improvement. We conclude that addressing the identified issues would facilitate cybersecurity dataset adoption in the wider research community.

1 Introduction

A key element of cybersecurity research is publicly available datasets. There has been a recent effort in understanding dataset usage, creation and sharing [22], and classifying available datasets in specific areas of cybersecurity, such as network-based intrusion detection [19]. These studies are essential initial steps in encouraging more studies within the field, and in identifying relevant datasets. However, besides identification there exists another factor that affects dataset usage: *the usability of the dataset*. Thus, to facilitate the adoption of cybersecurity datasets we identify the analysis of how users interact with them as the next step. It has been established that non computer science experts face challenges in the adoption of big data technologies, such as the accessibility of big data tools [20]. Similarly, we identify usability as a factor that may have implications on the adoption of cybersecurity datasets and may pose entry barriers to the field for different disciplines.

In this preliminary work we conduct a case study with the aim of assessing the usability of datasets offered by the Cambridge Cybercrime Centre (CCC).[1] We use *thematic analysis* to identify potential usability issues with the CCC's datasets, which would allow for improvement in the sharing of cybersecurity datasets in general. The study also highlights challenges users face when accessing the datasets to performing data manipulation. Further, this work identifies barriers to dataset uptake and leads to the establishment of guidelines that will facilitate the adoption of cybercrime and cybersecurity datasets by the wider community. The rest of the paper is structured as follows. Section 2 presents relevant related work. Section 3 discusses the methodology and the case study approach. Section 4 describes the analysis of the preliminary results, while a discussion is presented in Section 5. We conclude by describing future work in Section 6.

2 Related Work

In 2008, Chris Anderson [1] argued that theories and methodologies are becoming obsolete with the shift towards the petabyte age where correlations can be derived from big data with statistical models and algorithms. This argument sparked subsequent discussions and debates among scholars, including the definition of big data and its impacts on research. Common features across definitions of big data are *volume*, *variability*, and *velocity* [5, 11, 12]. Volume refers the size of the datasets. Scholars often deem datasets as *big data* when their volume is beyond current technology's capabilities to store, manage, and analyze the data [5, 11] while others suggested a threshold of exabytes [11, 12]. Variability refers to multiple data sources, and velocity refers to the speed data are added and collected [5, 12]. These features are commonly used in the definition for big data in other fields, such as social sciences [3, 7, 13, 14, 16]. However, in the field of social sci-

[1] https://www.cambridgecybercrime.uk/

ences, there are also references to big data as a technique [4] or by data sources such as social media data [6].

Despite the variations in definition, the *challenges with big data* are similar across disciplines. These challenges are associated with various steps of big data usage. Jagadish and colleagues [9] identified five steps: “acquisition, information extraction and cleaning, data integration, modeling and analysis, and interpretation and deployment” (p. 86). Across these steps, technical challenges are consistently identified [9, 11, 12]. Some examples of these challenges include the *construction of datasets from various data sources*, the *query process* that makes the datasets accessible, and the *requirement of diverse skills* [5, 12, 17].

The technical challenges posed by big data are a concern to the study of cybersecurity and cybercrime due to increasing availability of big data in the field [14, 19, 21]. For example, Williams and colleagues [21] demonstrated the possibility to identify and monitor stages of tension surrounding the discussions of riots on social media site. The rising availability and popularity of big data points to its potential as the next natural step in the advancement of the field of cybersecurity and cybercrime [8, 10, 19].

With increasing popularity and availability, scholars are becoming more aware of the challenges big data poses and the limitations in current processing capabilities. The challenges are exacerbated for scholars from non-computer science backgrounds. In a report on current trends in computational social science and big data research, Metzler and colleagues [16] found that a third of the sample (n = 9412) from various social science disciplines engaged in research involving big data. Users involved in big data research cited funding, access to data and search for suitable collaborators as the main three challenges [16].

Some of these identified challenges contradict existing recommendations among scholars. The current advice given to researchers with little computer skills who wish to use big data in their research is to collaborate with computer and/or data scientists [6, 7, 10, 16]. Although plausible, it is not the most ideal solution. The most common barrier faced by social scientists who wish to engage in big data research is finding suitable candidates with the right skills and knowledge for collaboration. Further challenges faced by social scientists include time consumption and acquiring new analytic methods and software [16]. Thus, to properly address the issue of big data in research, it is necessary to expand beyond the suggestion of collaboration.

3 Methodology

3.1 Research Questions

The focus of this work is to assess the usability of the datasets offered by the CCC. The dataset sharing process is divided into dataset sharing and usage. *Dataset Sharing* is the process of informing potential users about the datasets, providing them instructions on the application process and how to proceed once access has been granted. *Dataset Usage* is the process of querying, manipulating and exporting data.

The usability question is twofold. First, we are interested in assessing current users’ experiences of the dataset sharing process. The results of the assessment can be used to improve the current process, including potential user interface changes where necessary. Second, we aim to discover whether users report any challenges and difficulties that may pose obstacles in using any of the datasets. It is of particular interest how dataset size and the need to utilise tools for data manipulation contribute to these challenges. For example, when working with large datasets, users potentially need to extract smaller subsets of the data.

The results of the analysis can be used to discover general challenges users may face when using cybercrime and cybersecurity datasets. The reason to study data sharing as part of assessing the usability of the CCC’s datasets is that due to the nature of the access process (described in 3.2), this step also shapes users’ perceptions of dataset usage. To address the above, we aim to answer the following research questions:

- What is users’ perception of the data sharing process of the CCC?
- What are the major obstacles in the uptake of the datasets offered by the CCC?
- What areas of improvement can we identify during data sharing and usage?

3.2 Case study: CCC Datasets

The aims of this work and the research questions lend themselves to a case study approach. The CCC offers various datasets to researchers who wish to research cybersecurity and cybercrime. These datasets encompass different areas of cybercrime and cybersecurity research ranging from UDP reflection attacks, spam, malware data to underground forum discussions. The current data sharing process starts with a formal application, in which researchers are required to state brief details of their research project to assess relevance. Once permission is granted, the researchers receive instructions and relevant files to set up the dataset they wish to use. Current users represent a range of disciplines: computer science, criminology, data science, law, to mention a few.

3.2.1 Surveys

We chose a qualitative approach to allow the design of open ended questions and the exploration of users’ experiences of the data sharing process in-depth. We opted for *survey questionnaires* with open-ended questions due to its affordances in flexibility and lessening constraints on locations and time zone differences to participants [18]. In addition, survey questionnaires allow for the collection of qualitative and quantitative data and automatically directing participants to relevant questions [18].

The survey questionnaire was administered via Qualtrics,[2] and it consists of three sections. Section 1 focuses on user background including data needs and previous experience with secondary data. Participants who are *not* currently using any of the datasets they applied for, are directed to Section 2, which presents questions related to initial experiences with the datasets. Users are also asked to rate relevant steps of the data sharing process. Answers from this section highlight challenges and factors that discourage the uptake of datasets offered by the CCC. Participants who are currently using at least one of the datasets are directed to Section 3. This section is aimed at exploring participants' experiences with the data sharing process, data usage, and their applications of the datasets.

Selection of Participants
Selected participants are current users of the CCC's datasets who have applied for access, and who may or may not be actively using the datasets. Accordingly, we group current users into the following three categories: A) users who are currently using at least one of the datasets they applied for, B) users who have applied for access and accessed the dataset, however are not using it currently, and C) users who have applied and not accessed the dataset. Respondents were invited to participate in a solicitation email sent through Qualtrics. An anonymous link to the survey questionnaire was included in the email along with a brief description and statement of purpose on the study. After clicking on the link, participants were brought to the informed consent page with detailed description on the aims of the study, confidentiality and anonymity measures, and contact information of the researchers. The page also states that the user study is independent from the CCC's current data sharing process to avoid priming effects.

Ethical considerations
After reading the informed consent page detailing the confidentiality, anonymity, and withdrawal measures, and by selecting the option "Agree", participants agree to participate and consent to the conditions set out. The conditions state that data will be de-identified, and stored securely in an anonymised dataset, and users may withdraw from the study any time.

4 Analysis and Results

The solicitation email was sent to 65 users. Following the collection of survey data, responses were transcribed for further analysis. At the time of writing this paper 16 responses were received, nine of which were fully completed. The partially completed responses provided insights to a subset of the research questions.

Next, *thematic analysis* was performed, which is aimed at identifying patterns of themes in qualitative data [2, 15]. This method suits this exploratory analysis particularly well as it serves the purpose of understanding the needs and requirements of the different users of the datasets. Drawing from the responses, a list of initial codes and corresponding responses were generated. The codes were evaluated and collated to form themes. Given the preliminary nature of the study and limited number of received responses, the results are exploratory and subject to change. The preliminary results are presented in the subsequent sections.

4.1 Theme 1: Level of Technological Competence

Most respondents demonstrate a high level of technological competence. This is reflected in three dimensions: ***respondents' research areas***, ***self-reported technical knowledge*** and ***previous experience***.

Most participants indicate experience with programming languages, such as *Python* and *R*, with the exception of one user where none of the programming languages were applicable. Of those who indicate familiarity with Python, five respondents describe their skills as "good", "decent" and "familiar", and five respondents label their skills as "intermediate", "competent", "proficient", "high" and "excellent".

Most respondents report a basic level of familiarity with *SQL*, with the exception of three users who describe their skills as "excellent", "intermediate" and "medium". Only one participant reports a "high" level of skills with relational database management tools, while the rest of the respondents either indicate "basic" familiarity or that this skill is not applicable in their research.

The survey also reveals that high proportions of respondents conduct research in various areas of computer science, such as "usable security", "malware", "NLP" and "cybersecurity". "Cybercrime" was mentioned by four respondents.

Nine users have prior experience with other datasets not offered by the CCC, eight of which are related to cybercrime or cybersecurity. These datasets were available in SQL, SQLite, JSON, and ZIP file formats, and range between 4.1GB to 100GB. Five participants report that using these datasets required new data management and programming skills, such as PostgreSQL, Apache Hadoop, Active Mediating Object System (AMOS), tcpdump parsing, and SQLite, while performing data extraction and analysis required the use of R and Python. Some of the datasets were collected by the participants through their own research, some were open access, while some required an application process. Several respondents demonstrate familiarity with datasets gathered from online platforms such as "online anonymous marketplace" (Respondent A6) and "Github malware datasets" (Respondent B5).

In an effort to understand the users' background, we were

[2] https://www.qualtrics.com/uk/

also interested in finding out what dataset formats users prefer most. According to the responses, the most preferred format is CSV, followed by TXT and SQL.

4.2 Theme 2: Users' Experiences

This theme encapsulates users' interactions with the data sharing process at different stages, and is further divided to four sub-themes.

The first sub-theme is ***learning about the CCC***. Some respondents learned of the CCC's datasets through direct connections: *"'I worked there before"* (Respondent A1), *"Through collaboration with members of the CCC"* (Respondent B1), and *"Directly through one of the members of the CCC"* (Respondent B3). More commonly, respondents learned through other academic contexts such as emails, collaborations, and publications. This sub-theme suggests success in the CCC's efforts in informing scholars via diverse channels.

The second sub-theme is ***users' initial expectations*** of the process. Respondents appear to have either no initial expectation or positive expectations, with the latter centered around ease of access: a) *"Easy to apply, access"* (Respondent B5) and b) *"website interface"* (Respondent B7).

The third sub-theme is ***users' discussions on their current use of the datasets***. All respondents have accessed the datasets at least once, and nine respondents indicate current use of at least one of the CCC's datasets. The most discussed dataset is the CrimeBB dataset, which is considered big data as it contains more than 48 million posts from various online forums. When asked about selection criteria for data extraction, forum size was the most cited criterion: a) *"The size and including of the forums"* (Respondent A6) and b) *"as well as the size of them"* (Respondent A7). Other criteria included forum structure and forum ranking. The extracted data from the CrimeBB dataset remained large. Respondent A1 indicated working with 60 million posts while Respondent A7's datasets included 10,436 users.

The fourth sub-theme is ***users' feedback on the process***. The study reveals positive interactions with the data sharing process. Specifically, respondents identified three positive features with the process – ease of access, speed of access, and frequent update. This is illustrated with Respondent A6's comment:*"Data access stage was pretty easy and fast. But I had difficulties in setting up and installing the database because of the postgreSQL version conflict. It definetely needs the same version to restore the database. This issue was not told anywhere from the dataset owners, that I configured it myself by trial and error. One other positive aspect is they update the dataset periodically."* (Respondent A6).

This feedback also highlights obstacle to dataset usage: downloading and setting up data. Two other respondents indicate similar issues: *"Problems to download big files"* (Respondent A1) and *"... the raw data needs some processing to be used in mysql"* (Respondent B6). The feedback points to an area of improvement and the need for clearer instructions for this specific step in the data sharing process. Nonetheless, seven respondents indicate that they were "somewhat" or "extremely likely" to recommend the CCC's datasets to other researchers.

5 Discussion

The study reveals generally positive users' experiences with the data sharing process of the CCC. Information on the CCC's datasets are disseminated through both direct and indirect connections with the CCC. The diversity in channels through which researchers can learn about the datasets indicate no major obstacles in the uptake of the datasets during the dataset sharing phase.

With the dataset usage phase, users' experiences remain positive. Most users are using the CrimeBB dataset that contains more than 48 million posts. Despite the high volume of data, users identify ease of access, speed of access and frequent update of data as positive features of the dataset sharing and usage process. Most users indicate some likeliness in recommending the CCC's datasets to other researchers. In spite of that, users report technical challenges with data setup and download during this phase of the data sharing process. These technical challenges are specific and highlight the need for clearer instructions.

A limitation to the findings is the issue of self-reinforcement. The study found that most users have a high level of technical competence. These users are less likely to experience generic technical obstacles associated with big data, such as data query and extraction [9, 11, 12]. It is therefore possible that the positive experience and the specificity and the stage of reported obstacles is applicable only to this group of users.

In any event, the current study suggests success with the CCC's data sharing process among a sub-set of cybercrime and cybersecurity researchers. It is part of our future work to establish different user profiles of current users as we receive more responses and to widen current user profiles to include researchers with varied technical skills and backgrounds.

6 Future Work

We intend to take this work forward in a number of ways. Firstly, the results of the thematic analysis will be applied to improve the data sharing process. Users will be grouped based on their needs and requirements to establish user profiles. Next, the main usage scenarios of the datasets will be established to identify and assess usability issues for each user group, and as the next step, to formulate suggestions on resolving them. This process will provide an insight into the barriers to adopting the dataset by different groups of users.

Finally, we aim to investigate applying standard usability evaluation methods, such as *remote testing*, and the *in-situ pop-up response* method to gauge users' interactions with the datasets and to reveal specific usability issues with the current data sharing process of the CCC.

References

[1] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

[2] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[3] Janet Chan and Lyria Bennett Moses. Is big data challenging criminology? *Theoretical Criminology*, 20(1):21–39, 2016.

[4] Janet Chan and Lyria Bennett Moses. Making sense of big data for security. *The British Journal of Criminology*, 57(2):299–319, 2017.

[5] CL Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

[6] Adam Edwards, William Housley, Matthew Williams, Luke Sloan, and Malcolm Williams. Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology*, 16(3):245–260, 2013.

[7] Sandra González-Bailón. Social science in the era of big data. *Policy & Internet*, 5(2):147–160, 2013.

[8] Thomas J Holt and Adam M Bossler. An assessment of the current state of cybercrime scholarship. *Deviant Behavior*, 35(1):20–40, 2014.

[9] HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014.

[10] Karuppannan Jaishankar. Cyber criminology as an academic discipline: History, contribution and impact1. *International Journal of Cyber Criminology*, 12(1), 2018.

[11] Stephen Kaisler, Frank Armour, J Alberto Espinosa, and William Money. Big data: Issues and challenges moving forward. In *2013 46th Hawaii International Conference on System Sciences (HICSS)*, pages 995–1004. IEEE, 2013.

[12] Avita Katal, Mohammad Wazid, and RH Goudar. Big data: issues, challenges, tools and good practices. In *2013 Sixth international conference on contemporary computing (IC3)*, pages 404–409. IEEE, 2013.

[13] Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):1–12, 2014.

[14] David Lazer and Jason Radford. Data ex machina: introduction to big data. *Annual Review of Sociology*, 43:19–39, 2017.

[15] Moira Maguire and Brid Delahunt. Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education*, 9(3), 2017.

[16] Katie Metzler, David A Kim, Nick Allum, and Angella Denman. Who is doing computational social science? Trends in big data research, 2016.

[17] Steven Miller. Collaborative approaches needed to close the big data skills gap. *Journal of Organization Design*, 3(1):26–30, 2014.

[18] Jenny Preece, Yvonne Rogers, and Helen Sharp. *Interaction Design*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 2002.

[19] Markus Ring, Sarah Wunderlich, Deniz Scheuring, Dieter Landes, and Andreas Hotho. A survey of network-based intrusion detection data sets. *Computers & Security*, 86:147–167, 2019.

[20] Muhammed Asif Saleem, Blesson Varghese, and Adam Barker. Bigexcel: A web-based framework for exploring big data in social sciences. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 84–91. IEEE, 2014.

[21] Matthew L Williams, Pete Burnap, and Luke Sloan. Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *The British Journal of Criminology*, 57(2):320–340, 2017.

[22] Muwei Zheng, Hannah Robbins, Zimo Chai, Prakash Thapa, and Tyler Moore. Cybersecurity research datasets: taxonomy and empirical analysis. In *11th USENIX Workshop on Cyber Security Experimentation and Test (CSET '18)*, 2018.