*Technical Report*

UCAM-CL-TR-1003
ISSN 1476-2986

Number 1003

**UNIVERSITY OF CAMBRIDGE**

**Computer Laboratory**

# A ten-year review of the Cambridge Cybercrime Centre

Hannah Pankow, Alice Hutchings,
Richard Clayton

November 2025

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

*https://www.cl.cam.ac.uk/*

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

*https://www.cl.cam.ac.uk/techreports/*

ISSN 1476-2986
DOI *https://doi.org/10.48456/tr-1003*

**Abstract**

This report presents a comprehensive ten-year review of the Cambridge Cybercrime Centre (CCC), evaluating its impact on the academic research landscape. Employing a mixed-methods approach, we combine a scoping review of 203 publications with a survey of 44 dataset users and ten in-depth interview sessions with twelve individuals, including the Centre's leadership.

The findings reveal a consistent and accelerating uptake of the Centre's datasets, with 131 of the reviewed papers making explicit use of the data. This growth is driven by a broad, international, and interdisciplinary community. We find the Centre's primary contribution is the enablement of new research, particularly for postgraduate students and early-career researchers who gain access to large-scale data that would otherwise be infeasible to collect.

While technical setup difficulties present a barrier for some users, the development of the PostCog analysis tool has been a critical success, serving as both an accessibility and efficiency tool for the community. However, its impact is currently limited by an awareness gap. We conclude that the Centre's success lies in its evolution from a data provider to a holistic research ecosystem. Its future relevance will depend on its ability to adapt its data collection to evolving online harms and to ensure the long-term viability of its critical infrastructure.

# Contents

# Chapter 1

# Executive summary

This report marks the tenth anniversary of the Cambridge Cybercrime Centre by providing a comprehensive evaluation of its impact, challenges, and future directions. Through a mixed-methods approach combining a review of 203 publications, a survey of 44 dataset users, and ten interviews with twelve individuals, this report assesses the Centre's success in fulfilling its mission to significantly broaden the scale and scope academic cybercrime research. The findings are overwhelmingly positive, demonstrating that the Centre has evolved from a data provider into an essential research ecosystem.

## 1.1 Key findings

- **Significant and growing impact:** The Centre's datasets have seen a consistent and accelerating uptake, evidenced by 131 academic publications making explicit use of the data. This impact is international, with data sharing agreements spanning 473 scholars in 31 countries, and with 39% of publications now originating from non-affiliated researchers.
- **Enablement of an interdisciplinary community:** The Centre's most profound contribution is its role as an enabler of new research. It serves a diverse community of computer scientists, criminologists, psychologists, and others. For postgraduate and early-career researchers in particular, access to the Centre's large-scale, longitudinal data is transformative, allowing them to conduct empirical work that would otherwise be infeasible.
- **Barriers and successful interventions:** One of the main barriers some of the users encounter are administrative delays in the data access process and the technical complexity of setting up raw data dumps. The development of the PostCog analysis tool has been a critical and successful intervention, serving as both an accessibility tool for social scientists and an efficiency tool for computational experts. However, an awareness gap currently limits its full potential.
- **Strong alignment of vision and need:** The Centre's strategic vision to adapt its data collection to address evolving online harms on modern platforms aligns with the expressed needs of the research community.

## 1.2 Key recommendations

Based on these findings, we offer five key recommendations to guide the Centre's next phase of work:

1. **Streamline the data licensing process** to reduce the administrative bottleneck for users.

2. **Enhance dataset discoverability** on the Centre's website with clear summaries and data examples to reduce user uncertainty.

3. **Adapt data collection** to continue the pivot towards emergent harms on modern chat-based platforms.

4. **Invest in the continued development and dissemination of PostCog**, launching a targeted outreach campaign to close the awareness gap.

5. **Secure long-term funding for infrastructure maintenance** to ensure its critical data collection and analysis tools remain available to the international research community.

# Chapter 2

# Ten years of the Cambridge Cybercrime Centre

The Cambridge Cybercrime Centre was established on 1 October 2015. The Centre is based in the Department of Computer Science & Technology at the University of Cambridge. It was originally established through a research grant from the UK's Engineering and Physical Sciences Research Council (EPSRC). As well as doing our own research, the Centre collects and shares cybercrime data. The Centre's objective is to generate a step change in the quantity and quality of academic cybercrime research. The Centre's current primary funding is an ERC grant (iCrime) to research cybercrime offenders, cybercrime types, cybercrime places, and evaluating the effects of interventions.

In this report, we provide an overview of the Centre's achievements over the past ten years. To document these achievements, we surveyed and interviewed current and former members of the Centre and academic researchers who use our datasets. We also reviewed publications that include the Centre's datasets as a key contribution. In this review, we focus not only on what the Centre has been able to achieve, both directly and indirectly, but also on what we could be doing better and potential areas to focus on in the future.

## 2.1 Why a review now?

This year marks the tenth anniversary of the Cambridge Cybercrime Centre. This milestone provides a natural opportunity to evaluate the Centre's contribution and to reflect on the development of cybercrime research more broadly. Over the past decade, the Centre has established itself as a key hub for the systematic collection and distribution of cybercrime data.

The timing is also significant because the cybercrime landscape has shifted considerably in the past ten years. The expansion of Cybercrime-as-a-Service models, the role of cryptocurrencies, the increasingly profit-driven attacks, the effects of global crises such as COVID-19, and high-profile takedowns of illicit platforms have all reshaped online offending [1, 2]. Therefore, evaluating the Centre's first decade of data collection and research provides a timely opportunity to assess how academic research has captured and explained these changes.

As new datasets are added and existing ones continue to grow, there is value in consolidating what has been achieved, identifying gaps, and setting priorities for the next phase of work.

## 2.2 Research questions

This review is organised around four research questions:

### RQ1: dataset growth and uptake

As the Centre's datasets have evolved over time, our first question asks: To what extent have they been taken up by the research community? Some datasets, like our honeypot data (§3.2.4), reach back over a decade, while others, such as ExtremeBB (§3.1), were only introduced relatively recently. We therefore look at uptake in context, drawing on timelines, data-sharing agreements, and published outputs to show how different resources have spread across the research community.

We find that the academic uptake of the Centre's datasets has increased consistently over time, demonstrating a growing influence and relevance in the research community. The Centre's impact extends significantly beyond its own researchers, with a substantial and growing proportion of publications coming from our community of licensees.

### RQ2: research questions and contributions

Our second question relates to impacts and types of research problems scholars have addressed, namely: What kinds of research have the Centre's datasets enabled and how have they contributed to scholarship and researcher development?

We find the Centre's datasets have enabled significant interdisciplinary research, attracting a diverse user base where Computer Science and Social Sciences are both represented. For the majority of researchers, the Centre's datasets do not merely supplement existing work but enable new research projects that would otherwise be infeasible.

### RQ3: barriers and enablers of use

Access to large-scale cybercrime data is never straightforward, therefore our third question asks: What challenges do researchers face in working with the Centre's datasets, and what infrastructures have helped to lower these barriers? Here the emphasis is on practical obstacles, such as technical complexity, discoverability, ethical considerations, and the tools that have been developed to address them. PostCog (§3.2.3) and the Centre's natural language processing (NLP) classifiers (§3.2.2) are central, but so too are broader questions about usability and researcher support.

We find researchers from less computational disciplines such as criminology, sociology, and psychology, more frequently report issues related to data setup and environment (e.g., importing SQL, encoding issues). Conversely, researchers from more computational disciplines (e.g., computer science) report fewer issues overall, but are more likely to report issues related to data scale or specific formats.

We also find that while PostCog acts as a critical accessibility tool for researchers from non-computational disciplines to overcome technical barriers associated with handling raw data dumps, it has also readily been adopted by more technical users.

### RQ4: strategic lessons and future directions

Finally, we ask: What lessons can be drawn from the Centre's first decade, and how might its vision be carried forward?

We consider what the Centre has achieved relative to its founding aims, and how this informs the next phase of work. This includes questions of sustainability, expanding

datasets, and maintaining the balance between technical innovation, ethical responsibility, and interdisciplinary accessibility.

We reflect on how the nature of online harms has evolved since the Centre's inception, shifting from forum-based cybercrime towards more diverse, chat-based platforms and "emergent harms" (like extremism and harassment), requiring a strategic pivot in data collection. A major lesson from the Centre's first decade is that providing raw data is insufficient; sustainable impact requires building an accessible ecosystem that includes user-friendly tools (like PostCog) and robust support structures.

# Chapter 3

# Background

Obtaining reliable data on criminal activity has long been a central challenge for researchers. Studies of traditional offline offending typically rely on small-scale qualitative interviews, surveys, official records, or retrospective self-reports, each of which provides only a partial (and usually delayed) picture of criminal behaviour. Longitudinal data on trajectories into and out of crime are particularly difficult to come by and expensive to collect.

Cybercrime presents both new opportunities and new challenges. Unlike many offline settings, online forums and marketplaces leave behind large volumes of written interactions between offenders and aspiring cybercriminals. These glimpses offer a unique window into deviant subcultures, revealing learning processes, business models, and career pathways at a scale rarely possible in conventional criminological research [3].

Yet, systematic study remains difficult. Technical barriers such as CAPTCHAs, changing domain names, and access restrictions complicate data collection, while legal and ethical concerns constrain what can responsibly be stored and shared [4, 5]. In practice, individual researchers scraping their own datasets have often ended up with fragmented and outdated snapshots that are difficult to compare across studies.

An additional challenge arises from the dynamic nature of online communities. Forums close, reappear under new guises, or shift to new platforms; participants migrate between them; and the language and tools they use rapidly evolve [2, 5]. It requires large-scale sustained collection of data to make longitudinal analysis possible and the provision of datasets that enable reproducible research as opposed to one-off case studies. While some researchers work with leaked datasets or data provided under non-disclosure agreements from law enforcement or security companies, this raises further concerns around consent, legality, and reproducibility, thereby limiting their overall usability and value [4].

The Centre was established in response to these challenges. It collects and curates substantial datasets relating to cybercrime, while also negotiating access to additional feeds from external partners.[1] Leveraging its academic status, the Centre has built some of the largest and most diverse cybercrime datasets available to any research organisation. Its aim is to create a sustainable and internationally competitive hub for academic cybercrime research.

A key part of this mission is data sharing. Rather than keeping our extensive data collections to ourselves, the Centre makes them available to academic researchers across multiple disciplines under legal agreements designed to prevent misuse and to uphold ethical standards [4]. This model allows researchers to bypass the often prohibitive obstacles of negotiating access, building data collection tools, and maintaining their own infrastructure, enabling them to instead focus on substantive questions around online

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

offending.

## 3.1 Datasets

The flagship dataset, **CrimeBB**, contains discussion threads from dozens of underground forums. Since its initial release in 2018, CrimeBB has grown into the largest collection of its kind, expanding from an early version of approximately 48 million posts across four forums [3] to more than 129 million posts by 2025, spanning 40 forums in six languages (English, Russian, German, Arabic, Spanish, and Vietnamese) [6–8]. The historical depth is equally significant as the archive includes material dating back over two decades, to 2002 [6]. The forums include both surface and dark web pages, and cover a broad spectrum of activity, from malware development and vulnerability trading to online game cheating and discussions of grey area money-making schemes [3, 9]. Among these, Hack Forums stands out as the largest English-language hacking forum on the surface web [3, 10].

Building on this model, the Centre also established **ExtremeBB**, a data collection designed for the study of online extremism. By 2021, the dataset contained around 46 million posts from twelve extremist forums, including misogynistic, inceldom, and far-right spaces [11, 12]. In 2025, the collection has grown to 71.5 million posts on 17 forums. To reduce risks associated with illegal content, ExtremeBB currently preserves URLs for images and memes but does not download the media itself [11].

We have noticed that the level of engagement on cybercrime forums has declined over time. There is evidence of displacement to more ephemeral chat channels. Therefore, we pivoted our data collection efforts to collect chat data from a wide range of Telegram and Discord groups. The **CrimeCC** dataset includes 14 million messages and replies from 1,521 cybercrime-related channels. **ExtremeCC** contains 17.5 million message and replies from 883 extremist channels. In addition to forum-based resources, the Centre maintains extensive sensor data infrastructures. Since 2014, a geographically distributed **honeypot network** has captured data on Distributed Denial of Service (DDoS) activity, with around 80 active sensors across 62 IP prefixes and 15 autonomous systems [13, 14]. This dataset includes records of over four trillion packets providing a unique longitudinal view of the DDoS ecosystem. Related collections track Mirai scanning traffic and capture malware binaries, yielding around 15,000 samples to date.

Other datasets include **investment scam websites** (150,000+ URLs), archives of **phishing emails** and 419 "advance fee fraud" scams dating back to the early 2000s, **blog spam**, and **domain name registrations**. More recent additions encompass datasets on **modded Android applications** [15] and more than 550,000 **defaced websites** [16].

## 3.2 Infrastructure and tools

To ensure that datasets can be collected, processed, and made accessible at scale, the Centre has developed a set of tools and infrastructure to support continuous and comprehensive data collection.

### 3.2.1 CrimeBot

CrimeBot is the Centre's purpose-built crawler specifically designed to capture and regularly update data from underground forums in order to enable large-scale and longitudinal analysis of cybercrime forums [3, 17]. The scraper has been built iteratively over time, to deal with anti-scraping mechanisms such as login processes and CAPTCHAs as well

as site-specific defences [18]. CrimeBot systematically collects forum content, including threads, posts, and associated metadata, while archiving raw HTML for re-parsing if needed. It supports both initial crawls to capture an entire forum's historic content, and incremental crawls, which update datasets with newly posted material [3].

Furthermore, CrimeBot annotates posts with embedded elements such as images, videos, links, and attachments, thereby enabling richer forms of analysis [7]. While it detects and annotates these elements, it does not currently download non-textual data to reduce risks to researchers [3]. CrimeBot is the backbone of the Centre's forum and marketplace datasets such as CrimeBB and ExtremeBB.

### 3.2.2 Classifiers

The scale and linguistic complexity of underground forum data make manual coding or "off-the-shelf" NLP approaches infeasible. Posts are often short, filled with jargon, slang, and deliberate obfuscation, which means that conventional models trained on standard corpora perform poorly in this domain [19]. To address this, the Centre has invested in adapting NLP methods specifically to its datasets, most notably CrimeBB. These efforts have focused on automatically classifying and labelling forum posts to make them analysable at scale. Categories include post type (e.g. question, tutorial, comment), author intent (e.g. aggressive, helpful, neutral), addressee (general audience vs. specific user, and crime type (e.g. malware, DDoS, credential trading, identity theft, or other) [20, 21]. The classifiers label the CrimeBB datasets, and the labelled posts are integrated into PostCog for filtering [6].

### 3.2.3 PostCog

While the Centre's datasets open unprecedented opportunities for large-scale cybercrime research, they also present significant barriers to entry: working directly with raw SQL dumps requires technical expertise in databases, scripting, and NLP approaches [6]. To address this, the Centre developed PostCog, a web-based interface designed to lower these barriers and make CrimeBB, ExtremeBB, and chat channel datasets accessible to researchers from diverse disciplines.

PostCog provides several core functions. First, it enables keyword search across post content and thread titles, with results displayed alongside metadata such as forum, sub-forum, and date. Second, it allows users to apply filters not only by forum or date range but also by NLP-derived labels, including post type, author intent, and crime type–the classifiers developed through Centre's NLP pipeline (§3.2.2).

The interface further supports thread-level navigation for contextual analysis, a dash-board overview for high-level dataset statistics, and the ability to export filtered results as CSV files for offline analysis. Importantly, PostCog integrates a feedback mechanism, allowing users to flag mislabelled posts or hate speech, which in turn supports iterative improvement of the classifiers [6].

### 3.2.4 Honeypot network

Alongside its forum datasets, the Cambridge Cybercrime Centre operates a large-scale honeypot infrastructure to monitor and analyse DDoS activity, and UDP-based reflection and amplification attacks in particular. This system has been active since 2014. It uses a custom tool, Hopscotch, which was developed from scratch to emulate services commonly exploited in reflective UDP amplification attacks. Hopscotch mimics various vulnerable

UDP services, including Quote of The Day (QOTD), CHARGEN, DNS, NTP, SSDP, MS SQL Monitor (SQLMon), Portmap, and Multicast DNS (mDNS). The system is designed to attract malicious traffic by responding to protocol-compliant incoming traffic [13].

The Centre maintains a median of around 65 active honeypots that are geographically and topologically diverse, are deployed across multiple countries, and are located within 31 IP prefixes and eight Autonomous Systems (ASes). At its peak, the network has included 80 active sensors across 62 prefixes in 15 ASes. Each of these sensors typically receives around 200GB of inbound traffic per month, providing the Centre with insights into the frequency, scale, and targets of reflection attacks [14, 16, 22].

Ethical safeguards are a fundamental part of Hopscotch's design and operation – Hopscotch limits the number of packets it reflects, sending just enough packets to appear as a viable reflector before suppressing responses for 30 minutes to ensure it does not meaningfully contribute to ongoing attacks. It also maintains an exclusion list of known vulnerability discovery scanners such as Shodan to avoid skewing their measurements or providing them with misleading results [13].

## 3.3 User base

Since its inception, the Cambridge Cybercrime Centre has built a broad user base across disciplines and institutions. To date, the Centre's datasharing agreements span 473 scholars at 109 institutions in 31 countries across five continents. We note that while our datasets cannot be shared openly, our existing datasets are used significantly more than most open access datasets available the academic security community [23].

The reach of the Centre extends well beyond computer science. Criminologists, sociologists, legal scholars, economists, and psychologists have all drawn on datasets such as CrimeBB and ExtremeBB to address questions around offender learning [3, 24, 25], criminal marketplaces [1, 26, 27], cultural practices [12, 28–30], and regulatory responses [14, 22, 31, 32]. The Centre's commitment to accessibility has been particularly important for non-technical users: tools like PostCog allow researchers without programming or database expertise to explore and filter forum data, significantly lowering the barrier to entry [6].

By providing curated datasets under a robust legal and ethical framework, the Centre not only supports advanced computational projects but also empowers social scientists to carry out research that would have been left undone without such infrastructure in place. The interdisciplinarity has been central to the Centre's value, ensuring that cybercrime is not confined to a single discipline but instead benefits from multiple theoretical and methodological approaches.

The range of scholarship that has already been carried out using the Centre datasets, spans technical, cultural, economic, and policy domains.

## 3.4 Capability building and interdisciplinary training for research students and staff

Over the past ten years, the Centre has hosted, trained and collaborated with over 50 research staff, PhD students, research visitors, and undergraduate interns. Our team is highly interdisciplinary, with members coming from a wide variety of academic backgrounds, including computer science, criminology, law, anthropology, mathematics, linguistics, psychology, and economics. Team members who have left have gone on to have successful careers in academia and industry. Destination universities include Birkbeck

University London, Deakin University, King's College London, Queen Mary University London, Technische Universität Hamburg, Universidad Carlos III de Madrid, the University of California, Los Angeles, the University of Edinburgh, the University of Ljubljana, the University of Strathclyde, and the University of Tulsa.

# Chapter 4

# Methodology

## 4.1 Literature mapping

### Data collection

To assess the breadth and impact of the Cambridge Cybercrime Centre, we identify academic research outputs that make use of the Centre's datasets since its inception in 2015. The goal is to generate a comprehensive collection of relevant publications and to understand more about the researchers using our datasets and their research.

Our initial approach relied on internal records of dataset recipients – researchers who had signed a datasharing agreement with the Centre. We had planned to locate their academic profiles (e.g. on Google Scholar, OpenAlex, Scopus), extract a list of their publications, and then filter these for usage of the Centre's datasets. However, this proved more challenging than expected due to widespread disambiguation issues: many researchers had multiple or incomplete profiles, had changed institutions, or shared names with unrelated researchers in other fields. This makes it difficult to reliably match dataset recipients to their publications across platforms.

We also experimented with dataset-related keyword searches in academic databases, but these yielded limited results as dataset usage is rarely mentioned in the title, abstract, or keyword fields of publications – the fields indexed by most APIs. Instead, dataset acknowledgments typically appear in the methodology section or in footnotes, which are not captured by standard metadata or programmatic searches.

Given these limitations, we adopt a different strategy. Our data licensing agreement requires authors to acknowledge when they have used our datasets in their research, so we searched Google Scholar using the keyword "Cambridge Cybercrime Centre" (including variants such as "Cambridge Cybercrime Center" and the occasionally confused "Cambridge Cybersecurity Center") and extracted all the publications this located. We then augmented this list with the Centre's internal record of publications. This combined approach yields a working list of 203 unique published works. A small number of known papers that used the Centre's datasets but did not explicitly name the Centre (instead citing only a dataset release paper, such as the CrimeBB or ExtremeBB paper) are also added manually based on prior knowledge from earlier citation-based discovery efforts.

### Data processing and coding

Each publication was reviewed manually. Wherever possible, the full text was retrieved and the introduction and/or methodology sections were checked to verify and code dataset usage. Where multiple datasets were named or described in the text, each of them was recorded.

In a second coding step, we determine whether a publication was affiliated with the Centre. This was done by cross-referencing the publication list with the Centre's internal record of Centre-authored or co-authored research outputs. Publications appearing on this internal list are marked as Centre-affiliated, indicating authorship by a current or former Centre researcher. We also cross-referenced the list of authors against the Centre's internal recipient list to identify which researchers had previously received dataset access. This allows us to highlight overlaps between data recipients and published authors, examine patterns of dataset reuse, and better understand the relationship between dataset availability and research output.

### Thematic analysis

To map the scholarly landscape and understand the type of research enabled by the Centre's datasets, we conduct a thematic analysis of the 131 publications that use the Centre's data. Given the scale of the literature, we employ a hybrid approach that involves some AI elements next to human review. First, we conduct a close reading of a sample of key papers to develop an initial set of research themes. Following this, we use Google's NotebookLM to accelerate the process of identifying other relevant papers from within our list of publications and to generate initial summaries of their key findings. This AI-assisted workflow allows for a broad survey of the literature in a compressed timeframe.

### Analysis and visualisation

Following data collection and coding, we use a Google Colab notebook for analysis and generating visual summaries of the literature collection and dataset usage. This includes calculating overall dataset usage frequencies, identifying trends in publication output over time, and visualising author contributions and collaboration patterns. Charts and tables generated through this process are presented in the findings section. All code was written in Python.

### Limitations

We note several limitations. A small number of publications could not be retrieved in full due to broken links or restricted access. This applies primarily to a few dissertations (with expired institutional links) and some book chapters behind paywalls. These are excluded from the analysis unless dataset usage could be confirmed via abstracts or other descriptions.

There is also a degree of inconsistency in how datasets are referred to across publications. While "CrimeBB" and "ExtremeBB" are usually named explicitly and consistently, others – particularly chat-based datasets – are referred to using overlapping labels (e.g. "ExtremeCC", "Telegram", "chat channels"), which complicates fine-grained matching. As a result, broad usage patterns are reliable, but more granular distinctions between overlapping datasets should be interpreted with caution.

A further limitation arises from the hybrid human-AI approach used for thematic analysis. While we review all AI-generated summaries for coherence and consult a subset of the original papers for verification, it is possible that the AI tool did not always perfectly capture the main contributions of every paper. The resulting thematic analysis should therefore be seen as a broad, good-faith overview of the scholarly landscape rather than an exhaustive reading of every publication.

Finally, the literature mapping dataset is likely incomplete. Some researchers may fail to mention the Centre in their publications or cite only the intermediary flagship papers;

others have used the datasets for student theses, preprints, or literature not captured by our methods. As such, the set of 203 publications represents a conservative estimate of the Centre's academic footprint to date.

## 4.2 Survey

To complement our literature overview and provide a user-centred assessment of the Centre's impact, we conducted an online survey targeting researchers who had previously signed a data sharing agreement for one or more of the Centre's datasets. The goal was to collect detailed feedback on the usability, relevance, and impact of the datasets and associated tools (such as PostCog), as well as insights into how Centre resources could be improved. The survey design was informed by earlier work, particularly Pete and Chua's [33] assessment of the usability of cybercrime datasets, to ensure comparability with prior insights and to capture both technical and experiential dimensions of dataset usage.

### 4.2.1 Recruitment and timing

All individuals with an active or past data sharing agreement held in the Centre's internal records were invited to participate. The survey was launched on 28 July 2025 and remained open until 12 September 2025. A reminder email was sent on 1 September 2025. In total, 452 invitations were sent. Of these, 74 bounced or returned as undeliverable. The final number of completed survey responses is 44.

### 4.2.2 Survey structure and content

The survey is implemented in Qualtrics and structured into four thematic sections, along with an introductory participant information and consent form. In total, it comprises 47 questions, combining multiple-choice, matrix, and open-ended items.[1] Where relevant, logic branching is used to tailor follow-up questions. Open-text responses are used throughout to allow for elaboration. The survey was designed to take approximately 15–20 minutes to complete.

The survey is structured as follows:

#### Section 1: user background and dataset awareness

This section gathers information about respondents' academic roles, familiarity with relevant technical skills (e.g. Python, SQL, social network analysis tools), and prior knowledge of and engagement with the Centre's datasets. It includes items on preferred data formats and awareness of other cybercrime datasets, mirroring dimensions explored in Pete and Chua's [33] evaluation.

#### Section 2: experiences with access and process

Participants are asked to reflect on how they first discovered the Centre, their expectations of the data access process, and their experience with the formal application process. This section also asks about any technical barriers encountered during dataset download or setup.

[1] The full list of survey questions can be found in Appendix A.

**Section 3: dataset usage and research outcomes**

This section explores how participants use specific Centre datasets, their relevance to current or future research, and whether any requested datasets remain unused. Respondents are invited to suggest new types of cybercrime data or platforms that the Centre could consider collecting.

**Section 4: research value and follow-up**

The final section focuses on the perceived value of the datasets, the influence of the Centre's resources on respondents' research trajectories, and their likelihood of recommending the Centre's tools. Participants are asked if they would like to receive a summary of the findings or take part in a follow-up interview.

### 4.2.3 Data processing and analysis

The survey responses were exported from Qualtrics into CSV, and Python was used for cleaning and analysis. Multiple-choice and matrix questions are analysed using descriptive statistics to identify overall patterns of dataset usage, user satisfaction, and perceived barriers. Open-text responses are coded and reviewed to surface recurring themes and more nuanced feedback, particularly around tooling, access difficulties, and areas for improvement.

Where possible, we examine associations between user background and dataset usage patterns – for example, whether technical skill levels influenced the perceived ease of use of specific tools such as PostCog. Respondents who indicated willingness to participate in a follow-up interview were flagged and contacted separately. All survey data were analysed anonymously. For confidentiality reasons, no attempt is made to link responses to individual researchers or institutions.

**Limitations**

As all questions are optional, many respondents chose to skip open-text fields or leave them only partially completed. This limits the depth of insight for some topics, particularly when trying to connect technical or procedural challenges to specific research contexts, disciplines, or user profiles. As a result, some thematic findings remain suggestive rather than definitive.

## 4.3 Interviews

To enrich the survey data and capture more in-depth perspectives on the Cambridge Cybercrime Centre and its impact, we conducted a series of semi-structured interviews with a subset of individuals who had expressed interest in participating. In addition, we interviewed the founding director and the current director of the Centre to gather insights into the Centre's origins, development, and strategic vision.

### 4.3.1 Recruitment and participation

Interviewees are primarily recruited through the survey, where participants could voluntarily provide an email address if they were open to a follow-up conversation. A total of 13 survey respondents did so. As the survey is anonymous, no survey response data is linked to interview invitations.

Interviews were scheduled based on availability and project time constraints. In total, 10 interview sessions with 12 individual researchers are conducted. Due to time limitations, a small number of volunteers could not be contacted to schedule interviews. These individuals were invited to share any additional reflections via email if they had specific contributions they wished to voice.

### 4.3.2 Participant overview

The 12 interviewees represent a diverse range of roles, career stages, and affiliations, providing a multi-faceted perspective on the Centre's work and impact. The participants are summarised in Table 4.1. To protect anonymity, external and early-career researchers are referred to using an identifier.

Table 4.1: Overview of interview participants

| ID | Primary Role | Discipline(s) |
|---|---|---|
| **Leadership** | | |
| P1 | Founding Director | Computer Science |
| P2 | Current Director | Criminology / Computer Science |
| **Internal Researchers** | | |
| P3 | Postdoctoral Researcher | Computer Science |
| P4 | Former PhD Student | Computer Science |
| **External Professors** | | |
| P5 | Professor | Psychology |
| P6* | Professor | Data Science / Computer Science |
| **External PhD Students** | | |
| P7 | PhD Student | Computer Science / Human-Centred Security |
| P8 | PhD Student | Economics / Computer Science / Usable Security |
| P9 | PhD Student | Computer Science |
| **External Students** | | |
| P10 | Master's Student | Business Informatics |
| P11* | Undergraduate Student | Computer Science / Cybersecurity |
| P12* | Undergraduate Student | Computer Science / Cybersecurity |

*Note:* *Participants marked with an asterisk were interviewed together in a single group session.

### 4.3.3 Interview design and format

The interviews follow a semi-structured format with a shared set of core questions and group-specific prompts depending on the interviewee's role (e.g. external survey respondent, affiliated researcher involved in data collection, directors/founders). Topics covered include dataset discovery and usage, barriers to access or use, research outcomes, career development, and perceptions of the Centre's current role and future trajectory.

Interviews lasted between 20 and 65 minutes and were conducted between 30 July 2025 and 16 September 2025. Four interviews took place in person at the University of Cambridge and the remainder were held online via Google Meet.

### 4.3.4 Recording and transcription

All interviews were recorded with participant consent. In-person interviews were transcribed using Microsoft Word's automatic transcription feature and manually corrected. Online interviews were recorded via Google Meet, transcribed using its auto-transcription service, and subsequently reviewed and corrected for accuracy. The median interview length is 27 minutes, with an average of 32 minutes.

### 4.3.5 Analysis approach

Interview transcripts are analysed using thematic analysis, guided by Braun and Clarke's [34] widely adopted framework. Coding was carried out in Microsoft Word and Excel, using an iterative, question-driven process to identify patterns relevant to the project's aims. The codebook was developed inductively, guided by the interview data as well as the core research questions.

While not employing a fully reflexive thematic analysis approach, the researcher remained aware of their own disciplinary background and interpretive position throughout the coding process. Attention was given to participants' own language, and care was taken to represent a range of perspectives without forcing consensus.

Two rounds of coding were conducted, with themes developed by grouping related codes and refining them over time. The analysis focused on both recurring concerns (e.g., barriers to access, perceptions of dataset quality) and more nuanced reflections on the Centre's evolving role in the cybercrime research ecosystem.

### 4.3.6 Limitations

Given the limited number of interviews and the diversity of roles represented – including the Centre's Directors, affiliated researchers, and external dataset users – the analysis was not designed to reach thematic saturation. Rather, it aims to provide a structured snapshot of the range of experiences and insights shared by participants. Furthermore, interviews were conducted, transcribed, and analysed by a single researcher, which may limit interpretive triangulation. However, consistency was maintained through a transparent coding process and iterative theme development.

## 4.4 Ethical considerations

This research project received ethical approval from the Department of Computer Science and Technology's Ethics Committee at the University of Cambridge. All data collection and analysis were conducted in accordance with this approval.

The survey was designed to ensure respondent anonymity. To achieve this, a strict data separation protocol was followed. Any potentially identifying information, such as email addresses provided for follow-up interviews or lists of publications included in free-text responses, was removed and stored separately from the main analytical dataset. This process was completed prior to any analysis, ensuring that the survey responses were treated as fully anonymous. For confidentiality reasons, no attempt was made to link survey responses to individual researchers or institutions.

For the interviews, all participants received a Participant Information Sheet in advance and signed a Consent Form before taking part. No incentives were offered, though participants could opt to receive a summary of the findings. Interview transcripts were

anonymised prior to analysis, with all identifying information removed unless explicit consent for attribution was granted. This applies in particular to the interviews with senior Centre members, who were interviewed in their professional capacities.

# Chapter 5

# Results

## 5.1 Usage patterns of datasets

This section addresses our first research question: How have the Centre's datasets evolved over time, and to what extent have they been taken up by the research community? The analysis reveals a significant and growing academic footprint, driven by the Centre's philosophy of open data sharing. Out of 203 unique publications identified as relevant to the Centre's work, a majority – 131 publications (64.5%) – make explicit use of one or more of the Centre's datasets. This uptake is underpinned by a broadening international reach and the central importance of key "flagship" datasets.

### 5.1.1 A decade of growth in academic output

The primary indicator of the Centre's growing influence is the clear upward trajectory of academic publications that utilise its datasets. Figure 5.1 illustrates the total number of publications using the Centre's data published per year since 2017. The trend shows a consistent rise, peaking with 18 publications in 2024. While the data for 2025 appears lower, this is due to an incomplete year at the time of analysis. Evidence from our survey of dataset users suggests that the growth trajectory is likely to continue: roughly one-third of respondents (14 of 44) reported having publications or other works currently in the pipeline that use the Centre's data. These forthcoming outputs include journal articles and conference papers, indicating ongoing steady research based on the Centre's datasets.

Notably, the data shows a temporary dip in output in 2021 and 2022 – which may reflect the delayed impact of the COVID-19 pandemic on academic research cycles. The disruption to university operations in 2020 could have slowed the progress of long-term research projects, leading to a dip in publications. The subsequent sharp increase in 2023 and 2024 suggests a strong recovery and potentially a surge in research, analysing data collected during the lockdown periods.

### 5.1.2 Expanding reach: a broadening research community

Crucially, this growth is not merely an internal phenomenon. The data demonstrates the Centre's success in disseminating its resources to the broader research community. Of the 131 publications using the Centre's datasets, a substantial 51 (39%) originate from researchers with no current affiliation with the Centre. This global reach is further evidenced by the scale of the Centre's data sharing program. To date, data sharing agreements have been established with 473 scholars at 109 institutions in 31 countries.

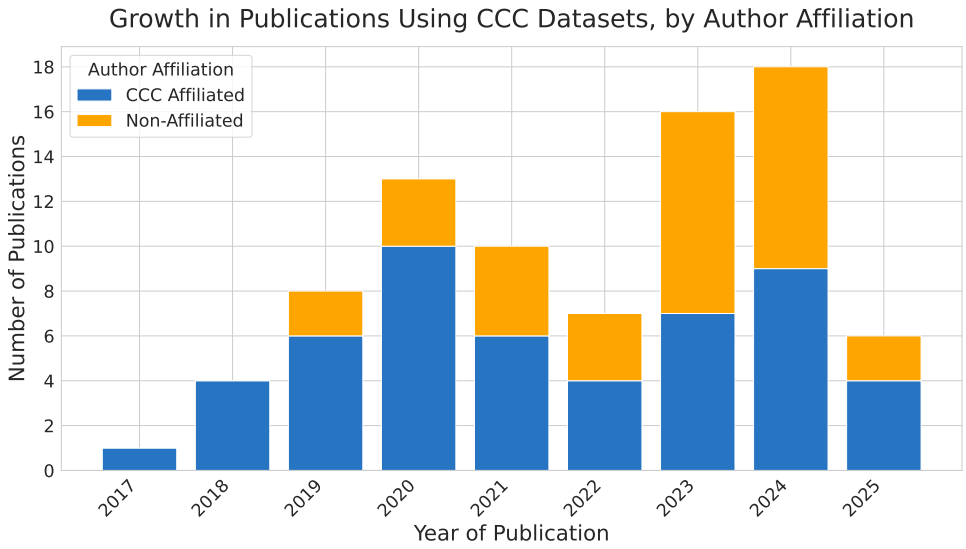


Figure 5.1: Growth in publications that use the Centre's datasets, broken down by author affiliation.

Figure 5.1 visually represents this, with the contribution from non-affiliated researchers (yellow segments) becoming a significant and consistent part of the annual output. This confirms that the Centre is effectively fulfilling its mission as a data provider for the wider field, enabling research globally.

### 5.1.3 The open data philosophy

This trend of widespread adoption can be directly attributed to the Centre's foundational philosophy, which was designed as a response to two critical problems in cybercrime research: data timeliness and the lack of reproducibility.

The Founding Director identified a common academic practice in which researchers retain exclusive access to a dataset until their own analysis is complete and published. This standard procedure often means that, by the time data is shared with the wider community, it is already significantly dated. As the Founding Director notes, "in cybercrime, two-year-old data is not as much use" [P1]. To combat this, the Centre adopted a radical policy of sharing data immediately, even before conducting its own analysis, ensuring the community receives resources at their point of maximum relevance. As the Founding Director put it: "we collect data and we give it to people and we haven't necessarily looked at it first ourselves" [P1].

The second, equally important problem was the fragmented nature of existing research data. Much of prior work, while valuable, is based on siloed, one-off data collections. This creates a landscape of disconnected studies that are difficult to compare or build upon. As the Founding Director observed, this resulted in "all of these papers with little snapshots of what was going on" [P1]. This fragmentation also created a scientific challenge, as these datasets were often "collected in a way which was not reproducible" [P1], making it impossible for findings to be independently verified. The Centre's model directly addresses this by providing common, longitudinal datasets, which allows the community to work from the same ground truth and, in principle, to "check other people's work" [P1], thereby fostering reproducibility, leading to a more robust and verifiable scientific process.

### 5.1.4 The cornerstones of the portfolio: dataset usage and user behaviour

While the Centre maintains a diverse portfolio of datasets, the analysis of publications, data-sharing agreements, and user surveys reveals two clear "flagship" resources that anchor its collection: CrimeBB and ExtremeBB.

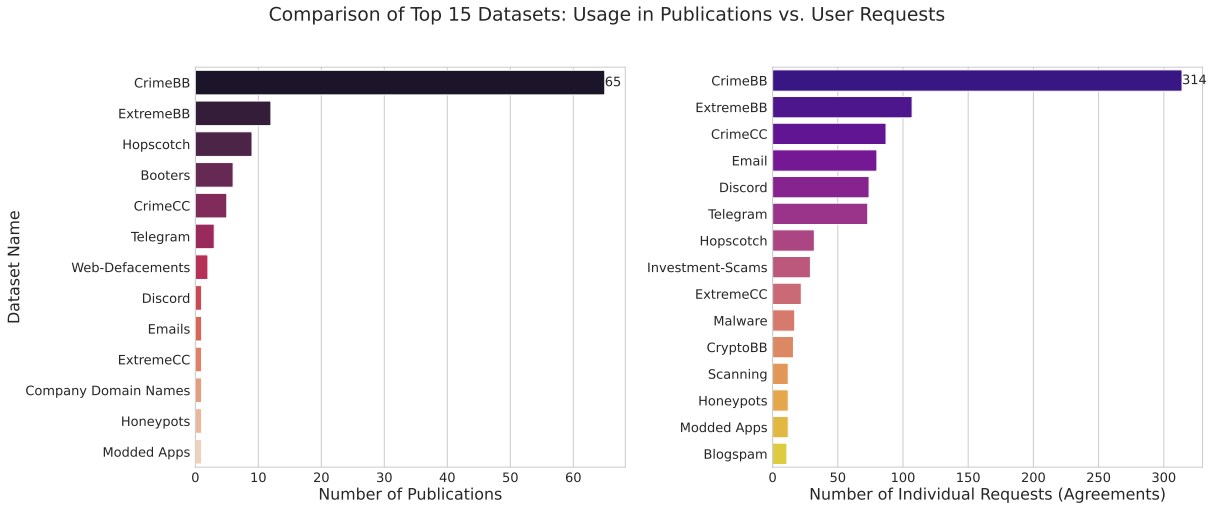


Figure 5.2: Comparison of Top 15 Datasets: Usage in Publications vs. User Requests.

Figure 5.2 provides a direct comparison of dataset usage. CrimeBB is overwhelmingly the most dominant dataset, cited in 65 publications and requested in 314 individual agreements. ExtremeBB, for which collections began much more recently, comes second and is cited in twelve publications and requested 86 times. This finding is strongly corroborated by the survey data, where 27 out of 44 (61%) respondents reported using CrimeBB (Figure 5.3).

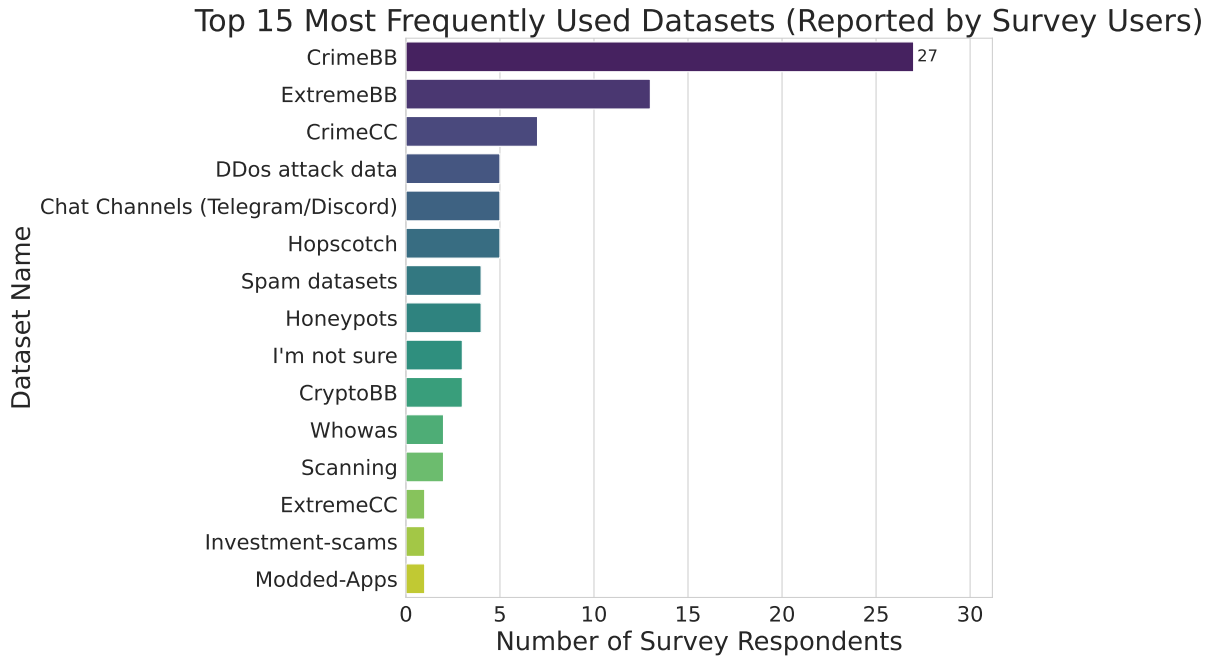


Figure 5.3: Top 15 Most Frequently Used Datasets as Reported by Survey Respondents.

The comparison between requested and published usage reveals valuable insights into

the data access and research process. Datasets such as CrimeCC (84 requests vs. 5 publications) and emails (76 requests vs. 2 publications) show a significant gap between demand and published output. The data from our survey and interviews suggests this discrepancy is not due to a single cause, but rather a combination of several contributing factors.

For newer resources like CrimeCC, a natural time lag in the academic lifecycle means that many projects using this data are likely still in the pipeline and have not yet reached publication. In addition, a significant portion of research outputs is not published or indexed; the survey data indicates that 36% of respondents used the data for a "MPhil/PhD thesis or student project(s)" and 24% for "Technical report(s)," which are less likely to be captured in our literature mapping. This is compounded by clear evidence of initial user uncertainty. As one survey respondent noted, it would be "very helpful to have a short description of the data of each specific dataset" before applying, a sentiment echoed by an interviewee whose supervisor "kind of didn't have any idea what it consisted of" before the research began [P10]. This suggests a pattern of speculative requesting, where researchers acquire multiple datasets but ultimately focus on one. Finally, as noted in the methodology, inconsistent naming conventions between project names (e.g., ExtremeCC) and platform names (Telegram) can complicate a precise mapping between requested and cited datasets.

## 5.2 Research questions and contributions

### 5.2.1 An interdisciplinary community of researchers

The Centre's datasets do not serve a single academic field; rather, they act as a hub for a broad, interdisciplinary community of researchers. This community comprises researchers at all career stages who often work across traditional disciplinary boundaries, a characteristic that mirrors the non-traditional pathways of the Centre's own leadership. The Founding Director [P1] came to academia after a career in the private sector, where he ran a software house and later became an "Internet expert" at a major ISP in the UK. Similarly, the current Director's [P2] interest in the field "predates my academic career", having worked in "private investigation and... brand protection" before pursuing a PhD in Criminology.

This cross-disciplinary nature is reflected in the Centre's user base. Figure 5.4 illustrates the academic diversity of survey respondents. While computer science (47.2%) represents the largest single group, there is strong representation from social and interdisciplinary sciences, including criminology/crime science (13.9%), cybersecurity (13.9%)[1], and psychology (5.6%). A significant portion of respondents (19.4%) fall into an "Other" category that includes fields as varied as economics, governance, sociology, and science & technology studies.

This interdisciplinary nature extends beyond researchers' formal backgrounds to the very practice of their research. While computer science represents the largest single group among survey respondents, our in-depth interviews revealed a consistent theme: even those who identify primarily with this field often describe their work as inherently interdisciplinary. This perspective is clearly articulated by one PhD student [P8], who notes that their professor's view is that their work in usable security "is 40% technical and 60% is more like psychology...". This sentiment was echoed by another computer science PhD

[1]Cybersecurity is listed as a separate discipline as some respondents' understanding is that it is inherently interdisciplinary and therefore not primarily computer science or criminology.

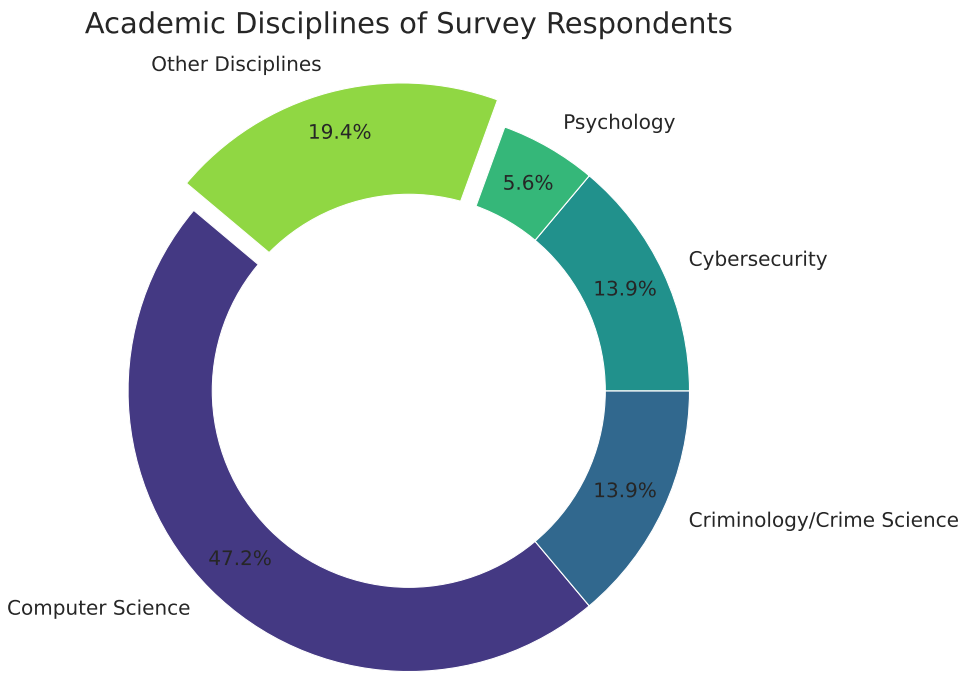


Figure 5.4: Academic Disciplines of Survey Respondents.
*Note:* 'Other Disciplines' (7 respondents) include: economics, computer science/criminology, governance, sociology, data science, information systems, and science & technology studies.

student, who described his research as "combining a technical background with the social side of things and using quite a lot of qualitative methods" [P9]. Another internal researcher [P3], despite a computer science background, described his work as evolving into "computer science with elements of criminology", adding that having a purely technical perspective means you "miss some of the theories behind why these things happen or some of the cultural aspects". This pattern, consistent across nearly all interviews, strongly suggests that the nature of the Centre's datasets naturally encourages researchers to adopt interdisciplinary perspectives, regardless of their primary training.

Figure 5.5 shows that this community is a mix of established and emerging researchers. While established Professors/Lecturers (36%) form the largest group, early-career researchers, including PhD Students (18%) and Postdocs (11%), make up a critical mass of the user base. This mix of disciplines and career stages creates a vibrant ecosystem for knowledge exchange. Many of these researchers explicitly identify as "translators" or "bridge-builders". One Master's student [P10] with a background in business informatics, described her role as being "the bridge between" technical and management stakeholders, feeling "more like a translator than a computer scientist." This reflects a common experience of navigating the"language barrier" [P7] and methodological differences between fields. We note that academic positions are often precarious at the early career stage, with research students and postdoctoral researchers often moving institutions. Given our high number of undelivered survey invitations, participants may also be biased towards more senior academics with more established positions.

The growth of this community appears to be driven primarily by academic networks. As seen in Figure 5.6, the most common ways researchers learn about the Centre are through "Colleagues" (24 mentions) and "Academic publications" (13 mentions). This "word-of-mouth" dissemination within and between universities is a strong indicator of the datasets' perceived value and utility in the field.

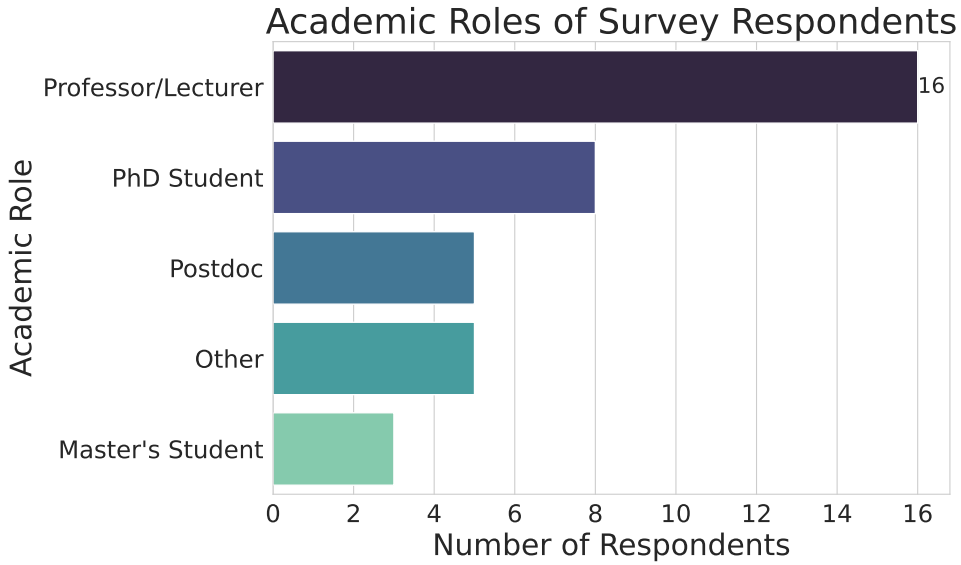


Figure 5.5: Academic Roles of Survey Respondents.
*Note:* The 'Other' category (5 respondents) includes roles such as Researcher, Undergraduate Student, and Junior Lecturer.

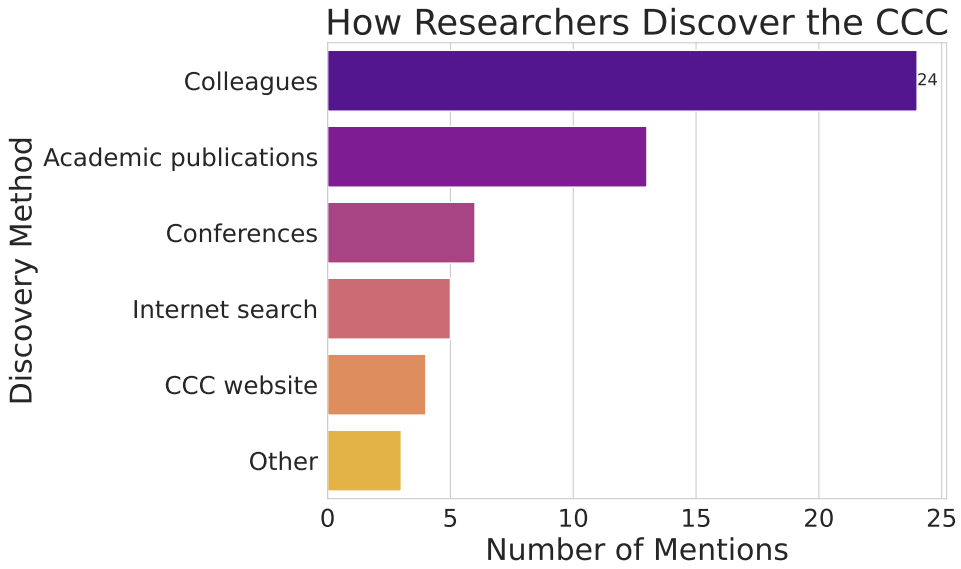


Figure 5.6: How Survey Respondents Discovered the Datasets.

A key function of this community is the development of the next generation of researchers. Our analysis of the publications using the Centre's data identifies six PhD theses that have used at least one of the datasets. The survey data points to at least one more completed PhD thesis that is currently not publicly accessible. While Master's dissertations are less likely to be formally published, we identified two in our search, and the survey data suggests this is a significant undercount, with many student projects using the data. This highlights the Centre's crucial role in providing the foundational data for postgraduate research.

### 5.2.2 A thematic review of centre-enabled research

This diverse, interdisciplinary research community has leveraged the Centre's datasets to produce a rich body of scholarship. Our thematic analysis of the publications that use

the Centre's data reveals five major clusters of inquiry, which collectively constitute the landscape of research enabled by these resources. The following subsections provide a narrative synthesis of some of the contributions within each theme.

### Mapping underground economies & markets

The datasets provided by the Cambridge Cybercrime Centre, particularly the extensive CrimeBB dataset, have been instrumental in advancing the study of underground economies. This body of scholarship provides data-driven insights into the structure, function, and economic mechanisms of illicit markets, progressing from initial descriptions of specific criminal enterprises to more sophisticated analyses of the entire ecosystem.

Foundational research in this area dissects specific illicit business models to understand their core mechanisms. An example is the work on 'eWhoring', where qualitative crime-script analysis is used to understand the fraudulent business model [35], which then directly informed subsequent measurements of the sources of materials and estimated offender profits [36]. This descriptive work is complemented by research into the critical financial and social infrastructures that underpin the ecosystem. Work in this area examines the financial instruments that facilitate illicit trade [10, 21] and investigates social mechanisms like appraisal and feedback systems that are essential for building trust [27].

Building on this foundational understanding, a second stream of research leverages the scale and longitudinal nature of the data to measure and analyse these markets over time. These studies use quantitative methods and machine learning to map broad market dynamics, most notably the commoditisation of cybercrime through Cybercrime-as-a-Service (CaaS) models. Researchers have developed classifiers to conduct large-scale analyses of the supply and demand for CaaS offerings, finding that the market is dominated by services for bots, reputation escalation, and illicit traffic [26]. Interestingly, this data-driven work can also challenge common assumptions; while the CaaS model is often perceived as rapidly expanding, the longitudinal analysis by Akyazi et al. [26] found no evidence for such growth over time within Hackforums, suggesting a more stable and resilient market segment than is often portrayed. The datasets also enable the mapping of other vast economies, such as the trade in fake social media engagement [37], and have proven particularly useful for understanding market evolution, as demonstrated by Vu et al. [1] in their detailed analysis of how trading activities and payment methods shifted during the COVID-19 pandemic.

### Hate speech & extremism

The Centre's datasets, particularly, the ExtremeBB collection, enable a systematic investigation into online hate, harassment, and extremism. The creation of this resource, a longitudinal collection of over 53.5 million posts from forums dedicated to the manosphere, white supremacy, hate and harassment, and other extremist ideologies [11], marks a deliberate expansion of the Centre's scope beyond cybercrime to specifically address radicalisation [4, 11].

This dataset facilitates deep inquiries into the nature of extremist ideologies. Research using ExtremeBB reveals significant intersections between far-right and misogynistic communities, identifying overlaps in their discourse on race and gender and highlighting shared radicalisation mechanisms [12]. However, the sheer volume of this data renders purely manual analysis infeasible and necessitates the development of automated methods. To this end, researchers use both the ExtremeBB and CrimeBB datasets to train and validate machine learning models for automated hate speech detection. This work has produced novel "span extraction" techniques capable of pinpointing specific toxic content within

lengthy posts, offering a practical tool for content moderators [38]. This approach builds on earlier efforts to classify "aggressive language" in hacking forums, which, while distinct from hate speech, first identified the presence of discriminatory content targeting women and ethnic minorities within these technical spaces [17]. This focus on methodological rigour is critical, as other work uncovered significant limitations in a widely used toxicity classifier that only processes the first 501 characters of a text – a crucial finding for the study of long-form forum content [30].

Finally, this body of research contextualises the scale of the challenge. By modeling user overlaps across extremist forums, one study estimates that the true population of active participants is 1.5 to 3.5 times larger than the number visible on monitored forums alone [9]. This finding demonstrates that the visible facets of online extremism represent only a fraction of a much larger, hidden ecosystem.

### Social and cultural dynamics of deviant online communities

Research leveraging large-scale datasets moves beyond purely technical analyses to illuminate the rich social and cultural dynamics governing deviant online communities. This body of work challenges the traditional image of the lone hacker, revealing that the modern cybercrime economy relies on a large, structured ecosystem with many participants performing routine administrative and support tasks. This fosters unique social structures and cultural norms, where psychological factors like boredom and burnout can play a surprisingly critical role in the stability of these ecosystems [39].

A central challenge within these anonymous marketplaces is the establishment of trust in the absence of legal recourse. Research has extensively examined the mechanisms developed to mitigate the inherent risk of scams, particularly the "cold start problem" faced by new users with no established reputation [40]. To overcome this, communities develop both market-level and individual-level trust signals. Market-level mechanisms include formal reputation systems and the public shaming of untrustworthy actors, or 'rippers', which serve as a vital community-enforced regulatory tool [40, 41]. At the individual level, Marjanov et al. [40] identify transparency –the choice to make contract details public – as a key trust-building strategy for newcomers.

This constant negotiation of trust is deeply intertwined with how participants perceive and manage risk. While actors engage in rational risk-benefit calculations, their assessments are often skewed by a distinct "optimistic bias", where experienced members are perceived as less likely to face negative consequences [29]. The perceived risk of arrest is often deemed low, contingent on the use of specific avoidance strategies like operational security and careful vendor selection [41]. This risk perception is also highly contextual; Bermudez Villalva [42] finds that cybercriminals on the dark web often exhibit a more cautious and sophisticated modus operandi than their surface web counterpart, while Wang et al. [43] reveal that fraudsters' decision-making is directly informed by whether they view the money itself as "easy", "difficult", or "risky".

Beyond market mechanics and risk, identity and gender are powerful forces shaping the culture of these subcultures. Research by Bada et al. [28] indicates that many forums operate within a "boy culture" that frames technical mastery as an intrinsically masculine pursuit. This has profound implications for behaviour, as misogyny often serves as a significant 'pathway' into cybercrime, with a relationship breakdown being a common trigger for individuals to seek tools for stalking or harassing former intimate partners [28, 44]. These cultural norms are reflected in communication patterns, where the prevalence of aggressive language can be seen as a manifestation of the underlying masculine and often misogynistic social environment [25, 28]. Collectively, this body of work demonstrates that

underground forums are not merely technical marketplaces but complex social systems, governed by intricate cultural norms, trust mechanisms, and identity dynamics that are essential for understanding offender behaviour.

### Threat measurement, cyber threat intelligence, and technical analysis

The technical analysis of cybercrime provides the foundation for effective threat measurement and intelligence gathering. The adversarial and noisy nature of underground forums necessitates the development of robust, large-scale data collection and analysis methodologies, with much of the foundational work leveraging the Centre's CrimeBB dataset. Given the impracticality of manually reviewing millions of posts, researchers have established a sophisticated toolkit of NLP and machine learning techniques to extract insights at scale. Foundational work by Caines et al. [20] established schemas and hybrid models to automatically classify forum posts by their function and intent, such as distinguishing offers from requests and tutorials. Building on this, Moreno-Vera et al. [45] applied more advanced models to categorise threads that cite Common Vulnerabilities and Exposures (CVEs) into the distinct stages of an exploit's lifecycle: Proof-of-Concept (PoC), Weaponization, and Exploitation. To bring structure to the vast lexicon of the domain, Mahaini et al. [46] proposed methods for building comprehensive cybersecurity taxonomies by combining automated NLP tools with human expert validation.

Beyond understanding the content of discussions, a significant research effort has focused on identifying the actors themselves, particularly those operating multiple accounts ("sockpuppets"). To this end, a variety of innovative, at-scale attribution techniques have been developed. Cabrero-Holgueras and Pastrana [47] pioneered a method that links accounts by extracting and correlating unique digital artifacts like email addresses and cryptocurrency wallets, aggregating them into a Multi-Feature Similarity (MultFS) score. Complementing this, the GeekMAN approach developed by Masud et al. [48] focuses on disambiguating "technogeek" usernames through novel "deslangification" and "chunkification" algorithms. Furthermore, Tereszkowski-Kaminski et al. [49] have successfully applied code stylometry to analyse the stylistic "fingerprints" within shared code snippets, enabling authorship attribution even from the small and incomplete fragments common in forum posts.

This methodological toolkit has enabled critical findings regarding the structure and dynamics of the cybercrime ecosystem. Collier and Clayton [50] use such technical analysis to deconstruct the pervasive but often misleading narrative of the "sophisticated attack", arguing the ecosystem is better understood as a division of labour. This model posits a small core of skilled hackers who develop exploits, which are then packaged by "tool builders" and deployed by a much larger base of less-skilled "entrepreneurs". This "Cybercrime-as-a-Service" (CaaS) model is observable across diverse illicit markets, including the trade of video game cheating injectors [51].

Consequently, underground forums have been proven to be an invaluable source of proactive Cyber Threat Intelligence (CTI). A longitudinal analysis by Paladini et al. [52] reveals that forum discussions of new threats, such as the DarkComet RAT, can precede official security reports by several years. Schröer et al. [53] quantify this, finding that while these "first-hand" sources contain significant noise, approximately 20% of the content is CTI-relevant, offering both strategic and technical intelligence. Through the use of ML explainability techniques, Moreno-Vera et al. [7] identify specific keywords like "fud" (fully undetectable) and "pm" (private message) as powerful indicators of active exploitation.

This theme of threat measurement is particularly prominent in the study of Distributed Denial of Service (DDoS) attacks. Research shows how malicious actors use search engines

like Shodan for passive reconnaissance to build botnets from vulnerable IoT devices [54]. However, accurately measuring the scale of DDoS activity remains a profound challenge. In his doctoral work, Nawrocki [55] demonstrates a significant visibility gap in common monitoring infrastructures, finding that large-scale honeypot networks observe less than 5% of baseline reflective amplification attacks compared to data from DDoS mitigation providers. This highlights that different vantage points, such as Internet Exchange Points (IXPs), capture largely disjoint sets of attack data. The analysis of attack campaigns during geopolitical events like the wars in Ukraine [16] and the Israel-Gaza conflict [56] further reveals that the engagement of low-level cybercrime actors, while initially intense with sharp spikes in defacements and DDoS attacks, is ultimately fleeting and diminishes rapidly. This body of work collectively underscores that a multi-source, correlated approach is essential for achieving a comprehensive understanding of the modern threat landscape.

### Evaluating interventions & policing

A critical area of cybercrime research involves the empirical evaluation of interventions designed to disrupt illicit online activities. This body of work moves beyond technical descriptions to measure the real-world effects of law enforcement and industry actions. A central theme is the shift away from traditional policing methods towards novel strategies tailored for the distributed, transnational nature of the internet. Collier et al. [32] provide a valuable framework for this analysis, conceptualising these emerging strategies as "infrastructural policing", which targets the technical backbone of cybercrime markets, and "influence policing", which uses targeted messaging to alter user behaviour.

Research consistently demonstrates that traditional law enforcement actions, such as arrests and sentencing, have a limited and often fleeting impact on the broader cybercrime market. In two related papers Collier et al. [22, 32] find that while high-profile arrests and sentencing decisions might cause short-term statistical dips in attack numbers, they fail to produce a lasting deterrent effect. The global and decentralised nature of these markets allows for rapid displacement, where new actors quickly fill the void left by those who are apprehended.

In contrast, infrastructural policing has proven to be a more potent, though still challenging, disruptive strategy. The initial "Booting the Booters" study by Collier et al. [22] provided the first large-scale evidence of this, showing that the closure of the "Server Stress Testing" section on Hackforums — a central marketplace for DDoS-for-hire services—led to a significant and sustained reduction in global DDoS attacks for 13 weeks. Similarly, a coordinated FBI takedown of 15 booter domains in 2018 suppressed attack numbers by a third for at least ten weeks and permanently altered the market's structure. Building on these findings, the most extensive evaluation to date by Vu et al. [14] assesses a major global takedown that began in late 2022. This intervention, which combined infrastructure seizures with deceptive domains, caused an immediate 20-40% reduction in global DDoS attacks. However, this study also highlights the remarkable resilience of the market, which managed to recover to pre-takedown levels within approximately six weeks, demonstrating the ecosystem's capacity for rapid adaptation.

The effectiveness of deplatforming extends beyond CaaS markets to other deviant online communities, though with similar complexities. The work of Vu et al. [57] on the concerted, industry-led effort to deplatform a hate and harassment forum reveals that even a multi-pronged campaign struggles to achieve a permanent shutdown. While the intervention successfully cut the forum's user base and activity by about half and displaced some activity to platforms like Telegram, it failed to eliminate the core community, which

demonstrated significant resilience and eventually returned. This suggests that without the ability to restrict key maintainers, deplatforming primarily affects casual users while the dedicated core often remains.

Finally, influence policing, particularly the use of targeted online advertising campaigns, has been a subject of evolving academic debate. The initial analysis by Collier et al. [22, 32] suggests that the UK National Crime Agency's (NCA) campaign, which targeted potential booter users with ads about the illegality of DDoS attacks, successfully flattened the growth of attacks in the UK. However, a subsequent cross-national, quasi-experimental study by Moneva and Leukfeldt [31] presents a more nuanced picture. Their analysis of seven similar campaigns across six European countries yields mixed effects, concluding that the evidence for the effectiveness of these ad campaigns is largely inconclusive. They suggest that the success observed in the initial UK case might have been influenced by concurrent interventions, such as police "knock-and-talk" visits, which were not accounted for in earlier models. This scholarly exchange underscores the methodological challenges in isolating the impact of specific interventions and highlights the necessity of rigorous, multi-faceted evaluation designs to accurately measure the effects of modern online policing strategies.

### 5.2.3 Enabling a new generation of researchers

While the breadth of published scholarship is significant, the Centre's most profound contribution lies in its role as a fundamental enabler of research. A thematic analysis of survey respondents' free-text descriptions of the datasets' influence reveals that the overwhelming majority of comments focus on this theme. As Figure 5.7 illustrates, the most common impact cited was that the data "Enabled New Research / Made Research Possible".

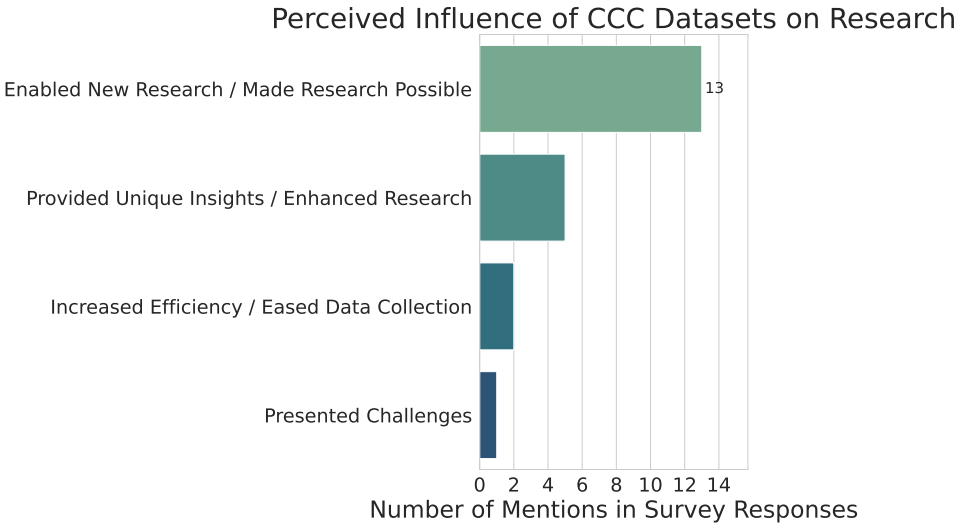


Figure 5.7: Primary Influence of the Centre's Datasets as Described by Survey Respondents.

This sense of enablement is the dominant theme in both survey and interview data. For many, the datasets provide the essential raw material for projects that would otherwise be infeasible. As one respondent stated, "Most of it would not have been possible without the CCC's data or been substantially more difficult, time consuming, and expensive". Another put it even more directly: "It has completely enabled my research as it is primarily based on using data from the shared datasets". For one researcher, the impact was career-

defining, stating the resources "[g]ave me a career... a huge training ground for learning to do complex mixed methods research".

Indeed, the interviews confirmed that the Centre's datasets often provide a solution when research teams are struggling to find appropriate data. As PhD student [P8] recounts, their team was actively "struggling with how to get data" for their project on offensive security software before learning about the Centre's resources.

Beyond making research possible, a second key contribution is the provision of unique data that enhances the quality and scope of projects. Survey respondents noted that the data "[p]rovided unique insights into that subculture" and allowed them to discover facets of their topic they would have otherwise missed. Another highlighted that the datasets "provided new linguistic diversity not previously seen in other hate speech datasets".

The impact on researcher development is concrete and measurable. As previously stated, our analysis of publications using the Centre's data identified six PhD theses, and the survey data confirms this is a primary use case, with one respondent calling the data the "Key to doctoral student's research". This is especially critical for postgraduate students, who often face the greatest hurdles in data acquisition. Therefore, the ability to provide students with large-scale, real-world data dramatically accelerates their research trajectory. As [P5], a psychology professor, highlights, this is a significant advantage for his current PhD students: "Because I already had the ethics for CrimeBB... basically on the first day of the PhD, we already had the data, which is an unusual situation to be in". Thus, the datasets allow students to move beyond logistical hurdles and focus on the core analytical work of their projects. As [P10] succinctly puts it, the data "really enabled me to do some fun research and to do research instead of just getting data".

Finally, it is important to acknowledge the single survey response coded under "Presented Challenges". The full comment reads: "Be able to identify and filter different languages, keywords, language patterns, and so on. Is a very interesting challenge". The phrasing here is ambiguous. It could be interpreted as a challenge presented by the dataset in the sense of a difficulty or limitation. However, it could equally be interpreted as a challenge enabled by the dataset (i.e., the data makes it possible to tackle interesting and difficult research questions). Given the overwhelmingly positive sentiment of all other responses, the latter interpretation is plausible. However, without further clarification, we present it as coded, while acknowledging that the challenges of working with large, multilingual, and complex real-world data are an inherent part of the research process that the Centre's resources make accessible.

## 5.3 Barriers and enablers of use

This section addresses our third research question: What challenges do researchers face in working with the Centre's datasets, and what infrastructures have helped to lower these barriers? The analysis of the user journey reveals two primary friction points: administrative delays in the access process and significant technical hurdles during data setup. To address the latter, the Centre developed PostCog, a web-based tool that has proven to be a critical enabler for many. However, its impact currently appears limited by an awareness gap among external researchers.

### Frictions in the user journey

The first barrier researchers encounter is the formal application process, which, according to the survey is viewed positively by a clear majority of users. The process is seen as

largely successful, particularly the 'Application Form' itself, which received the highest satisfaction rating, with 83.3% of respondents finding it satisfied or very satisfied.

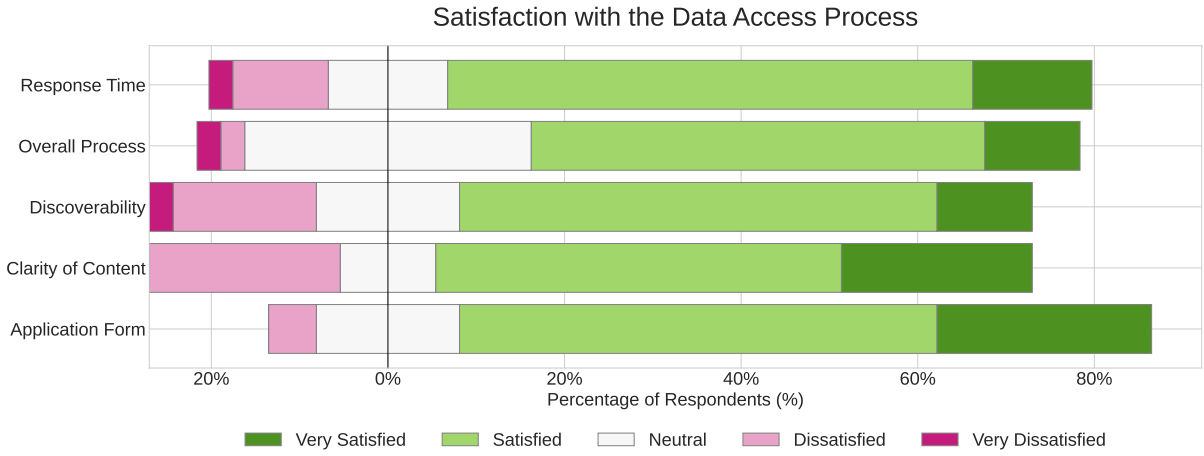


Figure 5.8: Satisfaction with the Data Access Process, as Reported by Survey Respondents.

*Note:* The chart displays the percentage of responses for each aspect of the data access process. Negative values represent dissatisfied responses, while positive values represent satisfied responses, centred around a neutral midpoint.

Figure 5.8 visualises the full distribution of user satisfaction. While the sentiment is overwhelmingly positive across all aspects, the chart allows us to identify specific areas with the greatest potential for improvement. For example, while a majority of respondents report a positive experience with the Centre's 'Response Time', this aspect also shows the highest concentration of negative sentiment, with 23.1% of respondents expressing dissatisfaction. This, along with 'Discoverability' (20%), indicates that finding the datasets and the subsequent wait time for a response are the primary administrative hurdles.

The qualitative interview data provides further insights into these findings. The issue with response time is linked to administrative bottlenecks. Internal researcher [P4] wished the Centre could "prioritise more... the processing time for the... licensing" as some applicants "need to wait for a very long time to hear back". This was confirmed by external users; PhD student [P8] noted that after applying, they felt their application was "forgotten" and they had to "send a reminder email". We note that delays are sometimes incurred when finalising the legal requirements with the data recipient's institution, and are therefore often outside the Centre's control.

Once access is granted, a second significant barrier emerges: the technical challenge of setting up and working with the raw data. A substantial 32% of survey respondents reported encountering difficulties during this phase. The most common challenges are not related to the initial download but to the subsequent setup phase, with "Compatibility issues with analysis tools", "Database version conflicts", and "Could not find the relevant documentation" being the most frequently cited problems (Figure 5.9).

The interviews demonstrated first-hand accounts of some of these struggles, particularly for those from non-computational backgrounds. [P5] described the experience of their team: "we were trying to download the SQL files, which if you're a psychologist with no technological background is very confusing and difficult". In some cases, the challenges emerge from institutional constraints: [P10] reported that their university's IT policy meant she "could not use SQL on my university server ... I had to convert everything into CSV files. So, just by doing that, you already lost a lot of interesting things about the data".

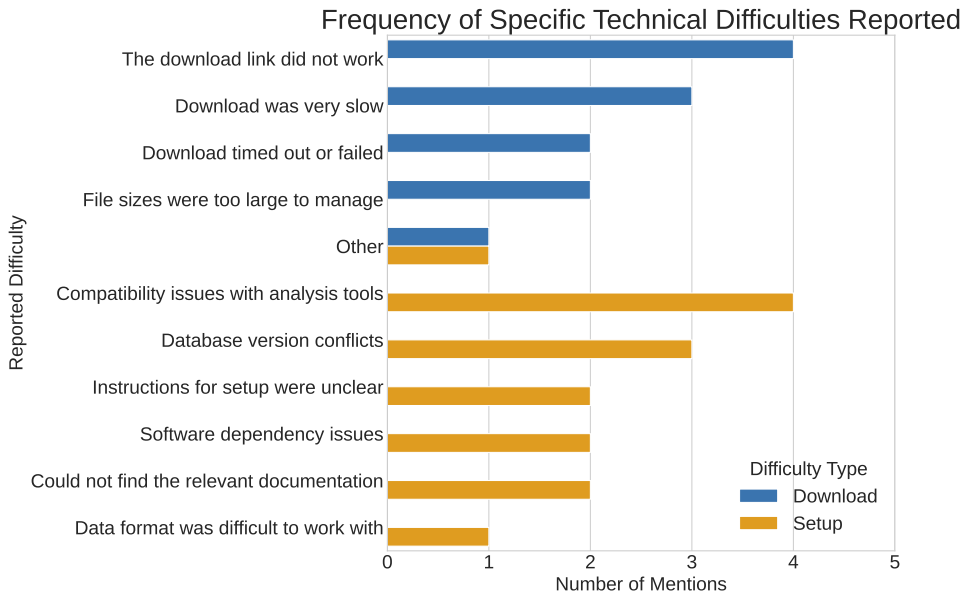


Figure 5.9: Frequency of Specific Technical Difficulties Encountered During Download and Setup.

*Note:* The chart shows the total number of times each specific difficulty was mentioned by survey respondents. Respondents could select multiple options.

### PostCog as the critical enabler

To address some of these technical barriers, the Centre developed PostCog, a web application for exploring the major datasets. Despite being only two years old, 50% of survey respondents reported using the tool. While we initially surmised that the tool would be disproportionately adopted by non-technical researchers, the data reveals a more nuanced story: PostCog is valued equally, if not more so, by technical users. The adoption rate among 'computational' researchers is slightly higher than among 'social science' researchers. However, as the overall number of respondents is on the low side, and more trained computer scientists than social science researchers replied to the question, these differences should not be overstated.

The interviews provide some level of explanation for this finding. For technical users, PostCog still functions as a powerful efficiency and exploration tool. They use it for reasons of convenience and to overcome the inherent challenges of scale. Before committing to the significant overhead of setting up a local database, some researchers use the tool for a "first glance" to quickly explore the data and validate their research questions. As [P4] explains, "I think it's particularly very useful when we just need to look at the dataset for a first glance". Similarly, [P8] initially tried to work with the raw data before switching: "at first, I used the SQL dumps but then I realised that it really is a lot of data and... then I found PostCog".

Ultimately, PostCog represents the path of least resistance for many common tasks, making it a preferred tool even for those with the skills to use the raw data dumps. Thus, it serves a dual purpose: it is an accessibility tool for non-technical users, making the data usable for them in the first place, and it is an efficiency tool for technical users, saving them significant time and effort. The Centre's leadership confirms this was a strategic decision – as the Director notes, providing data isn't enough, it needs to be made usable for the entire research community.

The enthusiasm with which the community has embraced this tool is evident in the survey feedback. Respondents described it as "really, really cool" and "very intuitive

and easy to use". It is highly valued by its non-technical audience, with one respondent stating, "PostCog is very helpful for researchers like myself, who do not have a technical background". At the same time, technical users praise it for its efficiency, with one noting, "I like PostCog much more that the old data sharing platform, I like that I can explore the data before I have to download a huge file, it's directly usable, it is much faster than if I ran the queries on the full data on my machine".

### The awareness gap and future directions

Despite its success, PostCog's impact is somewhat hampered by a lack of awareness of its existence, especially among external, early-career researchers who are among the ones who could benefit from it the most. In our interviews, all four external postgraduate students were initially unaware of PostCog's existence – two found out about it by chance after they had already started their data analysis, and the other two were unaware of the tool until it was mentioned by the interviewer.

This gap can be partially explained by the tool being relatively new. However, a further explanation may lie in the primary way researchers discover the datasets. As previously established, the main dissemination channel is through academic networks, with students often learning about the datasets from their superivsors. This creates a potential 'generational' information gap as supervisors who first accessed the datasets before PostCog's release may be passing on their knowledge of the raw data dumps only, being themselves unaware of the newer, more accessible tools.

The gap represents a missed opportunity to lower the barrier to entry. Upon learning of the tool, [P10] who had struggled to convert the SQL data, immediately recognised its value: "I think if I had known that this existed, especially during the exploratory phase of my study, it really would have helped... That would have been a helpful tool".

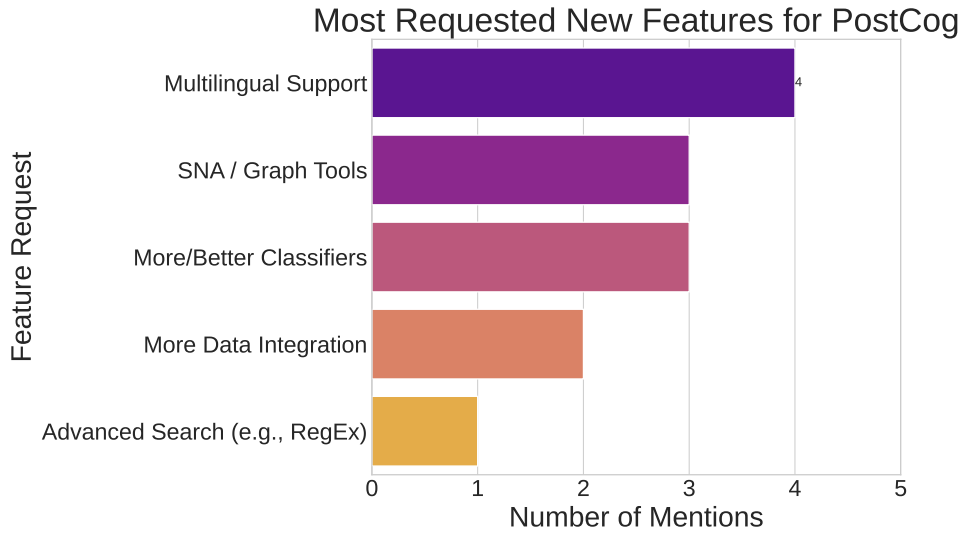


Figure 5.10: Most Requested New Features for PostCog.
*Note:* The chart shows the total number of times a feature category was mentioned in the free-text responses of survey participants who had used PostCog.

This suggests that updating the knowledge of existing members of the community who act as information hubs is at least as important as reaching new researchers. For those who do use the tool, there is a strong desire for ongoing development and new enhanced features. Figure 5.10 shows the most commonly requested additions with a demand for more advanced analytical tools like social network analysis and multilingual support.

### 5.3.1 Strategic lessons and future directions

This section addresses our fourth research question: What lessons can be drawn from the Centre's first decade, and how might its vision be carried forward? Drawing primarily on interviews with the Centre's founding and current Directors, and triangulating their perspectives with user data, we identify two key strategic lessons. First, the landscape of online harms is constantly evolving, demanding a flexible and forward-looking approach to data collection. Second, the Centre's core value lies not just in its datasets, but in the entire ecosystem it provides. Finally, we show that the Centre's future vision reflects the needs of the research community it serves.

#### From data provision to a usable ecosystem

An important strategic lesson from the Centre's first decade is the recognition that providing raw data, while necessary, is insufficient for maximising research impact. The initial success of the Centre's founding vision – providing timely and reproducible data – created its own challenges, primarily around technical barriers faced by a diverse user base. The leadership's key insight was that the Centre's value proposition had to evolve. As the Founding Director notes, even technically proficient users struggled with the raw data dumps: "we used to give people this SQL database and after a week they would come back to us and say 'how do you load this?'" [P1]. This barrier was even more acute for the interdisciplinary researchers the Centre sought to attract.

This realisation led to a strategic pivot towards prioritising usability and culminating in the development of PostCog. This was necessary to support the Centre's mission to provide more than just a data repository by building and maintaining an accessible research ecosystem. The Director of the Centre explains this strategic decision: just providing data isn't enough; it needs to be usable, which led to the development of PostCog.

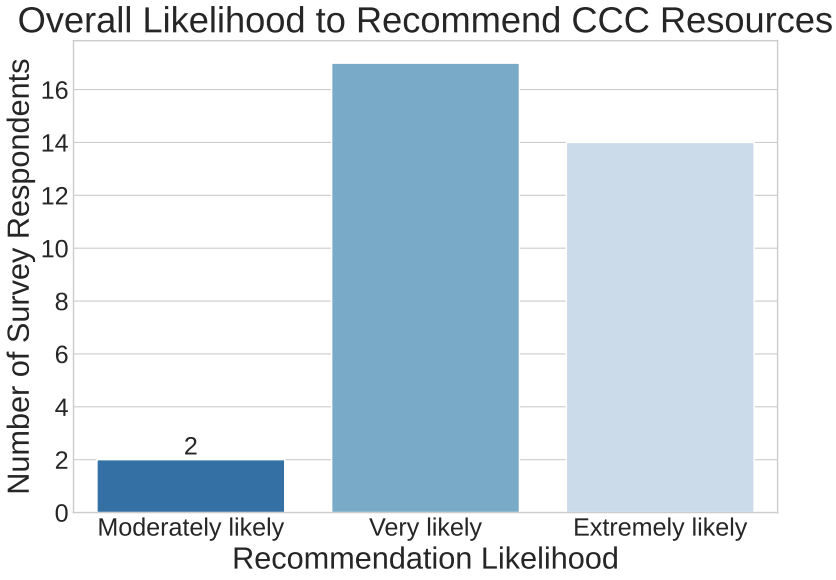


Figure 5.11: Overall Likelihood to Recommend the Centre's Resources.
*Note:* The chart displays all responses received for this question. No respondents selected 'Unlikely' or 'Very Unlikely'.

The success of this ecosystem approach is confirmed by the user community. All respondents who answered the question indicate they would be likely to recommend the

Centre's resources. As Figure 5.11 shows, the overall likelihood is overwhelmingly positive, with the vast majority of respondents selecting 'Very Likely' or 'Extremely Likely'. This high level of satisfaction is a direct endorsement of the value provided by the combination of data and accessible tooling.

### Future directions: adapting to an evolving landscape of harms

The Centre's vision for the future is shaped by the understanding that the landscape of online harms is constantly evolving. The Centre's leadership recognises that the forum-based model of cybercrime that dominated in 2015 is declining in relevance and therefore necessitates a strategic shift in the data being collected. As [P2] stated, the plan is to be "expanding out to look at and capture data on other types of harm" beyond traditional cybercrime, including "extremist datasets ... technology-facilitated domestic abuse, hate speech, hate and harassement". This requires the Centre to "shift how we collect data" [P2] towards new platforms, with [P1] specifically mentioning the need to better scrape "general cybercrime chat channels".

The user community seems to support this strategic vision: The popularity of the Centre's extremism-focused resources is already high; as established in 5.1.4, ExtremeBB is the second most requested dataset in the Centre's portfolio. In alignment with the leadership's plans, the research community also expressed a desire for expanded data collection from modern chat-based platforms. While the Centre already collects data from these sources, survey respondents expressed a clear demand for more extensive and targeted collection. When asked what other data the Centre should collect, the most frequent requests were for more Discord and Telegram channels, with some users specifying a need for data from servers focused on emerging harms like deepfakes. This indicates that the Centre's pivot to chat-based platforms is on the right track, and there is a growing appetite among researchers for even deeper and broader data from these modern channels.

### Future directions: pushing the methodological frontiers

A second direction for the Centre's future seeks to expand not only the type of data to be collected, but also pioneering new ways of how it can be analysed. The Centre's leadership identifies key challenges in current data processing capabilities and aims to leverage new technologies to add value to the datasets. The Founding Director highlights the current weakness in processing non-textual data, which is critical for understanding modern online discourse: "we don't do enough with pictures and so forth...when the message is the picture, right? And if we don't pick up the picture... then that message is lost". He also identifies the "very difficult problem" of "stringing conversations together on chat channels".

### Sustainability as an enduring challenge

Underpinning these future visions is the persistent and overarching challenge of sustainability. The success of the Centre's model – maintaining a fragile scraping infrastructure, storing petabytes of data, and developing user-facing tools – is resource-intensive and does not fit neatly into traditional academic funding models. The Centre's Director states "the biggest challenge at the moment is funding". This is echoed by internal researcher [P3], who notes that "one of the problems with the Centre is that you can't get funding for long-term projects, it's very challenging". Securing "maintenance fund" grants to refactor and sustain the existing infrastructure is therefore a constant and critical priority [P3].

# Chapter 6

# Discussion

This chapter discusses the findings presented in the preceding sections, placing them in the broader context of the Cambridge Cybercrime Centre's evolution and the wider field of cybercrime research. We begin by summarising the key findings before discussing their direct implications for the Centre, particularly in light of a foundational usablity study by Chua and Pete in 2019 [33]. We then consider the broader implications for the field and conclude by acknowledging the limitations of this evaluation.

## 6.1 Implications for the Cambridge Cybercrime Centre: then and now

Pete and Chua's study, "An Assessment of the Usability of Cybercrime Datasets" [33], provides a helpful baseline against which to measure the Centre's progress and identify persistent challenges. This current study, conducted several years later, reveals a story of both successful strategic adaptation and the emergence of new operational hurdles.

A key point of the Centre's evolution lies in the diversification of the user base. While Pete and Chua [33] identify a technically competent user base drawn primarily from computer science and criminology, our findings show that the community has since broadened to more consistently include researchers from fields like psychology, sociology, and governance. This wider interdisciplinary reach makes the problem of data setup (a persistent challenge identified in both studies) even more acute now than it was in 2019. The difficulties associated with handling raw SQL dumps, noted by Pete and Chua [33], are now compounded by institutional IT restrictions and the varied technical skill levels of a more diverse community, thereby emphasising the strategic necessity of accessible tools.

The development of PostCog is the Centre's most significant and successful strategic response to the usability issues identified in the 2019 study. Where the earlier work highlighted the steep learning curve of SQL as a barrier, our findings demonstrate that PostCog effectively mitigates this for both non-technical users (for whom it is an accessibility tool) and technical users (for whom it is a powerful efficiency tool).

However, our study also identifies new and emergent challenges. First, while the 2019 study focused on technical usability, our analysis highlights the administrative bottleneck of "Response Time" as a point of friction. Second, while PostCog was met with enthusiasm by researchers who are using it, several researchers were unaware of its existence, which means that more outreach might be necessary to ensure the research community has access to the tools it requires to make the most of the available data.

## 6.2 The Centre's contribution to criminology and cybercrime research

The findings of this review point to two fundamental contributions the Cambridge Cybercrime Centre has made to the field. First, by providing sustained, large-scale, and longitudinal data, it has enabled a shift towards more robust, empirical forms of inquiry that were previously infeasible. Second, by investing in accessible tooling and support, it has lowered the barrier to entry for a more diverse range of scholars, thereby broadening the types of questions being asked of cybercrime data.

The most significant contribution is the creation of a stable, longitudinal evidence base. Prior to the establishment of shared repositories like CrimeBB, much of the research on underground forums relied on fragmented, one-off data scrapes that produced valuable but disconnected "snapshots" of criminal activity. As the body of scholarship reviewed in this report demonstrates, the availability of consistent, decade-long datasets has enabled a more mature and scientifically rigorous mode of analysis. It allows researchers to move beyond static descriptions to empirically measure the multi-week effects of police takedowns, track the evolution of underground markets through distinct economic eras, and analyse the impact of geopolitical events on cybercrime. This makes a strong case that such data collection efforts should be viewed and funded as critical academic infrastructure as they enable research communities to answer questions that are otherwise intractable.

The Centre's second key contribution is its work in making this data usable for a broader interdisciplinary community. The diversification of the user base to include more researchers from non-computational fields like psychology and sociology is not merely an institutional statistic; it represents a direct contribution to the field by enabling new types of questions to be asked. The success of the PostCog tool among both technical and non-technical users provides a strong proof-of-concept for the entire field: investment in user-facing infrastructure dramatically increases the reach and impact of research resources. The persistent technical setup difficulties reported by survey respondents highlight a crucial lesson for anyone seeking to collect and make large-scale data accessible: It is a challenge that is as much about understanding the diverse needs and skills of the human researchers as it is about the engineering of the data infrastructure itself. By creating an ecosystem that supports a wider range of scholars, the Centre has helped to ensure that the study of cybercrime is not confined to a single discipline, but benefits from the multiple theoretical and methodological approaches necessary to understand this complex phenomenon.

## 6.3 Limitations

The findings of this report should be considered in light of several methodological limitations that affect the scope and generalisability of our conclusions.

- **Literature mapping:** The set of 203 publications represents a conservative estimate of the Centre's academic footprint. The collection process may have missed publications that did not explicitly name the Centre, as well as a significant body of unpublished work such as student theses. Furthermore, the thematic analysis of this literature used NotebookLM to summarise publications. While all summaries were reviewed, it is possible that the AI tool did not always capture the primary nuances of every paper. Finally, inconsistent dataset naming conventions in some

publications (e.g., ExtremeCC vs. Telegram) mean that while broad usage patterns are reliable, fine-grained distinctions should be interpreted with caution.

- **Survey data:** The quantitative findings from the survey are based on a final sample of 44 completed responses. This data is subject to self-selection bias, where users with particularly strong opinions may be more inclined to respond. Many survey invitations were undeliverable, likely due to early career researchers being at a precarious career stage, where movement across institutions is common. Therefore, responses may be skewed towards more senior academics. Furthermore, as all questions were optional, and not all respondents provided an answer for every item, leading to varying sample sizes for individual analyses. The findings should therefore be considered indicative of trends within this respondent pool rather than statistically generalisable to the entire user base of over 400 individuals. Finally, the survey was only disseminated to researchers who had successfully put in place a datasharing agreement. This sampling introduces a potential survivorship bias; we did not capture the experiences of researchers who may have applied but failed to complete the process. Such failures could be due to issues with the application itself or because an applicant's home institution was unwilling to sign the agreement.

- **Interview data:** The qualitative findings, while providing crucial depth and context, are based on a small sample of 12 individuals. The interviews were also conducted and analysed by a single researcher, which may limit interpretive triangulation. The purpose of this data is therefore to illustrate, explain, and add nuance to the quantitative findings, not to make standalone representative claims.

# Chapter 7

# Conclusion and recommendations

## 7.1 Summary of key findings

The evidence from our mixed-methods evaluation answers our four core research questions. First, the analysis of publications and user data confirms a consistent and growing uptake in the Centre's datasets, characterised also by an expanding international reach beyond affiliated researchers. This growth is rooted in the Centre's founding philosophy of providing timely, reproducible data and is anchored by the flagship CrimeBB and ExtremeBB datasets. Secondly, these resources serve a diverse, interdisciplinary community, enabling a wide range of scholarship on topics from underground economies to online extremism. The Centre's most significant contribution is the fundamental enablement of new research, especially for postgraduate and early-career researchers. Thirdly, the user journey is marked by barriers and enablers – while administrative delays and the technical complexity of raw data present significant frictions, the PostCog tool proves to be a critical enabler for both technical and non-technical users, although its impact is limited by researchers' awareness of its existence. Finally, the Centre's strategic vision for the future – expanding data collection to new forms of online harm and enhancing analytical tools – mirrors the expressed needs of its user community. However, the long-term sustainability of this work remains an enduring challenge.

## 7.2 Conclusion

The story of the Cambridge Cybercrime Centre's first decade is one of a successful evolution from establishing itself first as a provider of data into building a more holistic research ecosystem. The findings in this report demonstrate that the Centre's primary contribution to the field lies not in single datasets, but in the making available of a powerful combination of longitudinal, reproducible data and accessible tooling. It is this that successfully lowered the barrier to entry, enabling a diverse, interdisciplinary research community – particularly postgraduate and early-career researchers – to conduct empirical work that would have otherwise been infeasible due to technical or resource constraints. The development of PostCog is the clearest outcome of this lesson: it serves both as an accessibility tool for social scientists and an efficiency tool for computationally inclined researchers, highlighting that the Centre's value is in making its data truly usable. The enduring challenge for the next decade will be to sustain and adapt this successful model. The landscape of online harms is already shifting from the forums that defined the Centre's first decade to more ephemeral chat platforms, requiring a constant and flexible evolution in data collection. Furthermore, the success of the Centre's model – based on maintaining

fragile scrapers, storing petabytes of data, and developing user-facing tools – creates persistent financial challenges that do not always fit neatly with traditional academic funding cycles. Ultimately, the Centre's ability to continue enabling the next generation of cybercrime research will depend on its capacity to meet these twin challenges of adaptation and long-term operational viability.

## 7.3 Recommendations for the Centre's next decade

Based on the findings of this review, we offer five key recommendations to guide the Cambridge Cybercrime Centre's strategic priorities for the next phase of its work.

1. **Enhance dataset discoverability on the Centre's website.** The analysis reveals significant "initial user uncertainty," which leads to speculative data requests. This is supported by survey respondents and interviewees who requested clearer, public-facing summaries of each dataset's content and scope to better inform their applications.

2. **Streamline the data licensing process.** The findings identify "Response Time" as the single biggest administrative bottleneck and the aspect of the access process with the highest rate of user dissatisfaction (23.1%). This is corroborated by interview data from both internal and external researchers some of whom noted significant delays.

3. **Adapt data collection to address evolving online harms.** The interviews with the Centre's Directors and the data from the user survey are in agreement: Both identify a strategic need to pivot data collection away from declining forums and towards more contemporary platforms like Discord and Telegram to address "emergent harms". Continuing to adapt the data portfolio is essential to ensuring the Centre's ongoing relevance.

4. **Invest in the continued development and dissemination of Postcog.** The findings reveal an awareness gap regarding the Centre's primary accessibility tool, with all external postgraduate students interviewed being initially unaware of its existence. This suggests that there is value in a targeted outreach campaign & training. Furthermore, survey respondents who provided feedback expressed a clear desire for ongoing development. There was also a small number of specific feature requests such as social network analysis tools and multilingual support that could be investigated in the future.

5. **Secure long-term, dedicated funding for infrastructure maintenance.** The interviews with both leadership and internal researchers identify the precariousness of short-term academic funding as the single greatest threat to the long-term viability of the Centre's core data collection and tooling infrastructure. Securing dedicated maintenance funding is essential for the Centre's continued operation and impact.

# Acknowledgments

The Centre was made possible through the passion and dedication of Professor Ross Anderson, the Founding PI, whose memory continues to inspire us since his unexpected passing in 2024.

The Centre has been supported by a number of grants, including:

- EPSRC (EP/M020320/1): Interdisciplinary Centre for Finding, Understanding and Countering Crime in the Cloud
- EPSRC (EP/V026178/1): Tracking Covid Cybercrime and Abuse
- ESRC (ES/T008466/1): CybercrimeNLP (CC-NLP): A natural language processing toolkit for the interdisciplinary analysis of underground online forums
- European Research Council under the Horizon 2020 programme (grant agreement No 949127): iCrime: The Interdisciplinary Cybercrime Project
- EPSRC (EP/W032473/1): AP4L: Adaptive PETs to Protect & emPower People during Life Transitions

Our research has also been supported by donations from Meta and Google and gifts in kind from Digital Ocean and Solarflare Communications.

# Bibliography


[1] Anh V Vu, Jack Hughes, Ildiko Pete, Ben Collier, Yi Ting Chua, Ilia Shumailov, and Alice Hutchings. Turning up the dial: The evolution of a cybercrime market through set-up, stable, and COVID-19 eras. In *Proceedings of the ACM Internet Measurement Conference*, pages 551–566, 2020.

[2] Jack Hughes and Alice Hutchings. Digital drift and the evolution of a large cybercrime forum. In *Proceedings of the 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 183–193, 2023.

[3] Sergio Pastrana, Daniel R Thomas, Alice Hutchings, and Richard Clayton. CrimeBB: Enabling cybercrime research on underground forums at scale. In *Proceedings of the World Wide Web Conference*, pages 1845–1854, 2018.

[4] Jack Hughes, Yi Ting Chua, and Alice Hutchings. Too much data? Opportunities and challenges of large datasets and cybercrime. In *Researching Cybercrimes: Methodologies, Ethics, and Critical Approaches*, pages 191–212. Springer, 2021.

[5] Jack Hughes, Sergio Pastrana, Alice Hutchings, Sadia Afroz, Sagar Samtani, Weifeng Li, and Ericsson Santana Marin. The art of cybercrime community research. *ACM Computing Surveys*, 56(6):1–26, 2024.

[6] Ildiko Pete, Jack Hughes, Andrew Caines, Anh V Vu, Harshad Gupta, Alice Hutchings, Ross Anderson, and Paula Buttery. PostCog: A tool for interdisciplinary research into underground forums at scale. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 93–104, 2022.

[7] Felipe Moreno-Vera, Daniel Sadoc Menasché, and Cabral Lima. Beneath the cream: Unveiling relevant information points from CrimeBB with its ground truth labels. In *International Symposium on Cyber Security, Cryptology, and Machine Learning*, pages 280–290, 2024.

[8] Tina Marjanov, Taro Tsuchiya, Konstantinos Ioannidis, Jack Hughes, Nicolas Christin, and Alice Hutchings. Stayin' Alive: How global stolen data markets thrive on Telegram. In *under review*, 2025.

[9] Jonah Gibbon, Tina Marjanov, Alice Hutchings, and John Aston. Measuring the unmeasurable: Estimating true population of hidden online communities. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 56–66, 2024.

[10] Simon Butler. Cyber 9/11 will not take place: A user perspective of Bitcoin and cryptocurrencies from underground and dark net forums. In *Proceedings of the International Workshop on Socio-Technical Aspects in Security and Trust*, pages 135–153, 2020.

[11] Anh V Vu, Lydia Wilson, Yi Ting Chua, Ilia Shumailov, and Ross Anderson. ExtremeBB: A database for large-scale research into online hate, harassment, the manosphere and extremism. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[12] Yi Ting Chua and Lydia Wilson. Beyond black and white: The intersection of ideologies in online extremist communities. *European Journal on Criminal Policy and Research*, 29(3):337–354, 2023.

[13] Daniel R Thomas, Richard Clayton, and Alastair R Beresford. 1000 days of UDP amplification DDoS attacks. In *Proceedings of the IEEE APWG Symposium on Electronic Crime Research (eCrime)*, pages 79–84, 2017.

[14] Anh V Vu, Ben Collier, Daniel R Thomas, John Kristoff, Richard Clayton, and Alice Hutchings. Assessing the aftermath: The effects of a global takedown against DDoS-for-hire services. In *Proceedings of the USENIX Security Symposium (USENIX Security)*, pages 3595–3612, 2025.

[15] Luis A Saavedra, Hridoy S Dutta, Alastair R Beresford, and Alice Hutchings. ModZoo: A large-scale study of modded Android apps and their markets. In *Proceedings of the IEEE APWG Symposium on Electronic Crime Research (eCrime)*, pages 162–174, 2024.

[16] Anh V Vu, Daniel R Thomas, Ben Collier, Alice Hutchings, Richard Clayton, and Ross Anderson. Getting bored of cyberwar: Exploring the role of low-level cybercrime actors in the Russia-Ukraine conflict. In *Proceedings of the ACM Web Conference 2024*, pages 1596–1607, 2024.

[17] Andrew Caines, Sergio Pastrana, Alice Hutchings, and Paula Buttery. Aggressive language in an online hacking forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 66–74, 2018.

[18] Kieron Turk, Sergio Pastrana, and Ben Collier. A tight scrape: Methodological approaches to cybercrime research data collection in adversarial environments. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 428–437, 2020.

[19] Jack Hughes, Seth Aycock, Andrew Caines, Paula Buttery, and Alice Hutchings. Detecting trending terms in cybersecurity forum discussions. In *Proceedings of the ACL Workshop on Noisy User-Generated Text (W-NUT)*, 2020.

[20] Andrew Caines, Sergio Pastrana, Alice Hutchings, and Paula J Buttery. Automatically identifying the function and intent of posts in underground forums. *Crime Science*, 7(1):19, 2018.

[21] Gilberto Atondo Siu, Ben Collier, and Alice Hutchings. Follow the money: The relationship between currency exchange and illicit behaviour in an underground forum. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 191–201, 2021.

[22] Ben Collier, Daniel R Thomas, Richard Clayton, and Alice Hutchings. Booting the booters: Evaluating the effects of police interventions in the market for denial-of-service attacks. In *Proceedings of the Internet Measurement Conference*, pages 50–64, 2019.

[23] Anna Crowder, Allison Lu, Kevin Childs, Carson Stillman, Patrick Traynor, and Kevin RB Butler. Data to infinity and beyond: Examining data sharing and reuse practices in the computer security community. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 2678–2696, 2025.

[24] Jack Hughes. *Computational criminology: at-scale quantitative analysis of the evolution of cybercrime forums*. PhD thesis, University of Cambridge, 2023. URL `https://www.repository.cam.ac.uk/handle/1810/366133`.

[25] Sergio Pastrana, Alice Hutchings, Andrew Caines, and Paula Buttery. Characterizing Eve: Analysing cybercrime actors in a large underground forum. In *Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*, pages 207–227, 2018.

[26] Ugur Akyazi, Michel van Eeten, and Carlos H Gañán. Measuring Cybercrime as a Service (CaaS) offerings in a cybercrime forum. In *Workshop on the Economics of Information Security*, pages 1–15, 2021.

[27] Zhengyi Li and Xiaojing Liao. Understanding and analyzing appraisal systems in the underground marketplaces. In *Proceedings of the Network and Distributed System Security (NDSS) Symposium*, 2024.

[28] Maria Bada, Yi Ting Chua, Ben Collier, and Ildikó Pete. Exploring masculinities and perceptions of gender in online cybercrime subcultures. In Marleen Weulen Kranenbarg and Rutger Leukfeldt, editors, *Cybercrime in Context: The Human Factor in Victimization, Offending, and Policing*, pages 237–257. Springer, 2021.

[29] Maria Bada and Yi Ting Chua. Understanding risk and risk perceptions of cybercrime in underground forums. In *Proceedings of the IEEE APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–11, 2021.

[30] Anna Talas and Alice Hutchings. Hacker's Paradise: Analysing music in a cybercrime forum. In *Proceedings of the IEEE APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–14, 2023.

[31] Asier Moneva and E. Rutger Leukfeldt. The effect of online ad campaigns on DDoS-attacks: A cross-national difference-in-differences quasi-experiment. *Criminology & Public Policy*, 22:869–894, 2023.

[32] Ben Collier, Daniel R. Thomas, Richard Clayton, Alice Hutchings, and Yi Ting Chua. Influence, infrastructure, and recentering cybercrime policing: Evaluating emerging approaches to online law enforcement through a market for cybercrime services. *Policing and Society*, 32(1):103–124, 2021.

[33] Ildiko Pete and Yi Ting Chua. An assessment of the usability of cybercrime datasets. In *Proceedings of the 12th USENIX Workshop on Cyber Security Experimentation and Test (CSET)*, 2019.

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

[35] Alice Hutchings and Sergio Pastrana. Understanding eWhoring. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 201–214, 2019.

[36] Sergio Pastrana, Alice Hutchings, Daniel Thomas, and Juan Tapiador. Measuring eWhoring. In *Proceedings of the Internet Measurement Conference (IMC)*, pages 463–477, 2019.

[37] David Nevado-Catalán, Sergio Pastrana, Narseo Vallina-Rodriguez, and Juan Tapiador. An analysis of fake social media engagement services. *Computers and Security*, 124:103013, 2023.

[38] Linda Zhou, Andrew Caines, Ildiko Pete, and Alice Hutchings. Automated hate speech detection and span extraction in underground hacking and extremist forums. *Natural Language Engineering*, 29(10):1247–1274, 2023.

[39] Ben Collier, Richard Clayton, Alice Hutchings, and Daniel Thomas. Cybercrime is (often) boring: maintaining the infrastructure of cybercrime economies. In *Workshop on the Economics of Information Security (WEIS)*, 2020.

[40] Tina Marjanov, Konstantinos Ioannidis, Tom Hyndman, Nicolas Seyedzadeh, and Alice Hutchings. Breaking the Ice: Using Transparency to Overcome the Cold Start Problem in an Underground Market. In *Workshop on the Economics of Information Security (WEIS)*, 2025.

[41] Maria Bada and Yi Ting Chua. Examining risk and risk perception on LSD and MDMA in online marketplaces. *Journal of Crime and Justice*, 47(4):456–471, 2024.

[42] Dario Adriano Bermudez Villalva. *Understanding the difference in malicious activity between Surface Web and Dark Web*. Doctoral thesis (Ph.D.), University College London (UCL), May 2022.

[43] Fangzhou Wang, Timothy Dickinson, and Adam K. Ghazi-Tehrani. Not all money is the same: The meanings of money in online fraud. *Crime & Delinquency*, 0(0):1–29, 2025.

[44] Quincy Taylor, Anna Talas, and Alice Hutchings. Love Bytes Back: Cybercrime following relationship breakdown. In *Proceedings of the IEEE APWG Symposium on Electronic Crime Research (eCrime)*, pages 123–135, 2024.

[45] Felipe Moreno-Vera, Mateus Nogueira, Cainã Figueiredo, Daniel S Menasché, Miguel Bicudo, Ashton Woiwood, Enrico Lovat, Anton Kocheturov, and Leandro Pfleger de Aguiar. Cream skimming the underground: Identifying relevant information points from online forums. In *Proceedings of the IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 66–71, 2023.

[46] Mohamad Imad Mahaini, Shujun Li, and Rahime Belen Sağlam. Building taxonomies based on human-machine teaming: Cyber security as an example. In *Proceedings of the ACM 14th International Conference on Availability, Reliability and Security*, 2019.

[47] José Cabrero-Holgueras and Sergio Pastrana. A methodology for large-scale identification of related accounts in underground forums. *Computers & Security*, 111: 102489, 2021.

[48] Md Rayhanul Masud, Ben Treves, and Michalis Faloutsos. Disambiguating usernames across platforms: the GeekMAN approach. *Social Network Analysis and Mining*, 14 (1):177, 2024.

[49] Michal Tereszkowski-Kaminski, Sergio Pastrana, Jorge Blasco, Guillermo Suarez-Tangil, et al. Towards improving code stylometry analysis in underground forums. In *Proceedings on Privacy Enhancing Technologies (PETS)*, 2022.

[50] Ben Collier and Richard Clayton. A "sophisticated attack"? innovation, technical sophistication, and creativity in the cybercrime ecosystem. In *Workshop on the Economics of Information Security (WEIS)*, 2022.

[51] Panicos Karkallis, Jorge Blasco, Guillermo Suarez-Tangil, and Sergio Pastrana. Detecting video-game injectors exchanged in game cheating communities. In *European Symposium On Research In Computer Security*, pages 305–324, 2021.

[52] Tommaso Paladini, Lara Ferro, Mario Polino, Stefano Zanero, and Michele Carminati. You might have known it earlier: Analyzing the role of underground forums in threat intelligence. In *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*, pages 368–383, 2024.

[53] Saskia Laura Schröer, Noe Canevascini, Irdin Pekaric, Philine Widmer, and Pavel Laskov. The dark side of the web: Towards understanding various data sources in cyber threat intelligence. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 79–89, 2025.

[54] Maria Bada and Ildiko Pete. An exploration of the cybercrime ecosystem around Shodan. In *Proceedings of the IEEE 7th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, pages 1–8, 2020.

[55] Marcin Nawrocki. *DDoS Attacks: Coverage, Mitigation, and Prevention.* Doctoral Dissertation (Dr. rer. nat.), Freie Universität Berlin, Department of Mathematics and Computer Science, 2023.

[56] Anh V Vu, Alice Hutchings, and Ross Anderson. Yet another diminishing spark: Low-level cyberattacks in the Israel-Gaza conflict. In *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 126–131, 2025.

[57] Anh V. Vu, Alice Hutchings, and Ross Anderson. No Easy Way Out: The effectiveness of deplatforming an extremist forum to suppress hate and harassment. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, pages 717–734, 2024.

# Appendix A: survey instrument

This appendix contains the full list of questions and logic used in the online survey.

## Participant information and consent

**QID2: Welcome to the Cambridge Cybercrime Centre Survey**

Thank you for helping us assess the impact of the datasets and resources we provide. Your feedback is crucial for improving our services for the research community. The survey has four sections:

- Section 1: Your Background & Preferences
- Section 2: Experiences with Dataset Access & Usage
- Section 3: Research Goals & Outcomes
- Section 4: Final Feedback

---

*The survey included a detailed pop-up Participant Information Sheet covering the purpose of the research, what participation involves, data handling, and contact information for the research team.*

---

**QID3: Consent Form**

Please read the following statements. By ticking the box below, you confirm your agreement to all points:

- I have read and understood the Participant Information Sheet for this study.
- I have had the opportunity to ask questions and have had them answered to my satisfaction.
- I voluntarily agree to take part in this research project.
- I understand that my participation is voluntary and that I may stop at any point before submitting my responses, without needing to give a reason.
- I understand that my responses will be stored securely, analysed anonymously, and that any contact details I choose to provide will be stored separately and only used for the purposes stated.
- I understand that anonymised data from this project may be used in academic publications or future research and will not include any personally identifiable information.

- I understand that my data will be stored within the UK on secure University of Cambridge systems and will not be exported in identifiable form.

**Q52:**

_ I have read the statements above and agree to participate in this study.

# Section 1: User Background and Dataset Awareness

**QID4: What is your primary academic discipline/field of study?**
*(e.g. Computer Science, Criminology, Sociology, Law, ...)*
*(Open text response)*

**QID5: What is your current academic role?**

- PhD Student
- Master's Student
- Postdoc
- Professor/Lecturer
- Other

**QID6: Please specify:**
*(Open text response. Display Logic: Shown if QID5 is 'Other')*

**QID8: Have you used other cybercrime or cybersecurity datasets from sources outside the Cambridge Cybercrime Centre before?**

- Yes
- No

**QID9: Please briefly describe these datasets (e.g. type of data, approximate size, format) and the tools you used for their extraction and analysis.**
*(Open text response. Display Logic: Shown if QID8 is 'Yes')*

**QID10: Which data formats do you generally prefer for research datasets?**

_ CSV

_ TXT

_ SQL

_ JSON

_ ZIP

_ Parquet

_ Other

**QID11: Please specify:**
*(Open text response. Display Logic: Shown if 'Other' is selected in QID10)*

# Section 2: Experiences with Access and Process

**QID16: How did you first learn about the Cambridge Cybercrime Centre and its datasets?**

- ⎯ Academic publications
- ⎯ Conferences
- ⎯ Colleagues
- ⎯ CCC website
- ⎯ Internet search
- ⎯ Other

**Q53: Please specify:**
*(Open text response. Display Logic: Shown if 'Other' is selected in QID16)*

**Q55: Please rate your experience with the following aspects of the data access process:**

- Discoverability (Finding CCC datasets online)
- Clarity (Understanding dataset contents and application steps)
- Application Form (Ease of completing the required information)
- Response Time (Speed of communication from the Centre)
- Overall Process (How straightforward did it feel from start to finish)

*Scale: Very Difficult, Difficult, Neutral, Easy, Very Easy*

**QID19: Did you encounter any technical difficulties during the download of the datasets?**

- Yes
- No

**Q54: What types of download difficulties did you encounter?**
*(Select all that apply. Display Logic: Shown if QID19 is 'Yes')*

- ⎯ Download was very slow
- ⎯ Download timed out or failed
- ⎯ File sizes were too large to manage
- ⎯ The download link did not work
- ⎯ Other

**Q58: Did you encounter any technical difficulties during the setup of the datasets?**

- Yes
- No

**Q59: What types of setup difficulties did you encounter?**
*(Select all that apply. Display Logic: Shown if Q58 is 'Yes')*

- _ Instructions for setup were unclear
- _ Could not find the relevant documentation
- _ Database version conflicts
- _ Software dependency issues
- _ Compatibility issues with my analysis tools (e.g., Python, R, NVivo)
- _ Data format was difficult to work with
- _ Other

**QID24: Which specific CCC dataset(s) have you used for your research?**
*(Select all that apply)*

- _ DDos attack data
- _ CrimeBB
- _ CrimeCC
- _ CryptoBB
- _ Chat Channels (Telegram/Discord)
- _ ExtremeBB
- _ ExtremeCC
- _ Honeypots
- _ Hopscotch
- _ Investment-scams
- _ Malware
- _ Modded-Apps
- _ Scanning
- _ Spam datasets
- _ Whowas
- _ None
- _ I'm not sure

**Q51: Besides the datasets you selected above, were you granted access to any others that you have not yet used?**

- Yes
- No
- I'm not sure

**QID22: What were the main reasons for not using the other datasets?**
*(Select all that apply. Display Logic: Shown if Q51 is 'Yes')*

- Data not as expected
- Too complex
- Lack of time
- Required technical expertise
- Project focus changed
- Found alternative data
- Just received access/ Haven't started yet
- Other

**QID13: Have you used PostCog?**
*(web application designed for exploring CrimeBB, ExtremeBB, CrimeCC, ExtremeCC, and CryptoBB). Display Logic: Shown if specific datasets are selected in QID24.*

- Yes
- No

**QID14: Please provide feedback on your experience [with PostCog].**
*(Open text response. Display Logic: Shown if QID13 is 'Yes')*

**QID15: What additional features would you like to see integrated into PostCog?**
*(e.g. SNA tools, multilingual support for NLP classifiers, integration with other datasets). (Open text response. Display Logic: Shown if QID13 is 'Yes')*

**QID26: Are there other specific cybercrime channels or forums you wish the CCC collected?**
*(Open text response)*

# Section 3: Research Goals and Outcomes

**QID29: What were the primary research objectives or goals of your project(s) that utilised Cambridge Cybercrime Centre datasets?**
*(Select all that apply)*

- Key Actor Analysis/Detection
- Discourse Analysis
- Subcultural Analysis
- Crime Type Analysis
- Longitudinal/Evolutionary Analysis
- Economic Analysis
- Cyber Defence/Threat Intelligence
- General Measurement/Characterisation
- Intervention/Policy Impact
- Other

**QID32: What primary research methodologies did you employ in your analysis of the Cambridge Cybercrime Centre data?**
*(Select all that apply)*

- Quantitative Analysis
- Qualitative Analysis
- Mixed Methods
- Other

**QID34: What types of research outcomes have resulted from your and your team's use of CCC data?**
*(Select all that apply)*

- Published journal article(s)
- Published conference paper(s)
- Technical report(s)
- MPhil/PhD thesis or student project(s) (published)
- MPhil/PhD thesis or student project(s) (unpublished)
- Policy brief(s) or report(s) for non-academic audiences
- Applied impact/interventions (e.g. industry, law enforcement)
- Talks/seminars

- Grant applications
- Grant funding
- Other

**QID36: If published output(s): Please provide citation(s) or link(s) to your published work.**
*(Open text response)*

**QID37: Do you have any forthcoming publications or works currently in the pipeline that have utilised CCC datasets?**

- Yes
- No

**QID38: Please provide a working title or brief description of these forthcoming publications/works.**
*(Open text response. Display Logic: Shown if QID37 is 'Yes')*

# Section 4: Final Feedback

**QID39: How has the CCC's data influenced your research?**
*(e.g. enabled new work, provided unique insights, presented challenges,...). (Open text response)*

**QID41: Overall, how likely are you to recommend the CCC's datasets/PostCog to other researchers in your field?**

- Not at all likely
- Slightly likely
- Moderately likely
- Very likely
- Extremely likely

**QID42: Do you have any final comments or suggestions for the Cambridge Cybersecurity Centre?**
*(Open text response)*

**QID43: Would you be willing to participate in a follow-up interview to discuss your experiences in more detail?**

- Yes
- No

**QID44: Please provide your preferred email address.**
*(Open text response. Display Logic: Shown if QID43 is 'Yes')*

**QID45: Would you like to receive a summary of the findings when the project concludes?**

- Yes
- No

**QID46: Please provide your preferred email address.**
*(Open text response. Display Logic: Shown if QID45 is 'Yes')*

# Appendix B: Topic Guide for Semi-Structured Interviews

The following topic guide was used to provide a consistent structure for the semi-structured interviews. The exact wording and order of questions were adapted based on the interviewee's role (e.g., leadership, internal researcher, external user) and the natural flow of the conversation. Follow-up and probing questions were used throughout to explore emergent themes.

## 1. Researcher Background & Pathways

- Can you tell me a little bit about your current role and institution?
- What is your primary academic discipline or field of study?
- How did you first become interested in cybercrime or cybersecurity research?
- Would you consider your work to be interdisciplinary? If so, could you describe the challenges or benefits of working across disciplines?

## 2. Dataset Discovery and Access

- How did you first learn about the Cambridge Cybercrime Centre and its datasets?
- Did you approach the data with a specific research question in mind, or was your approach more exploratory?
- Can you describe your experience with the data sharing agreement process?
- Did your institution's ethics board have any specific concerns about using the data?

## 3. Dataset Usage and Tooling

- Which specific CCC datasets have you used?
- Can you describe the main technical or practical challenges you faced when working with the data (e.g., download, setup, format)?
- Have you used the PostCog web interface?
- *(If yes) What was your experience with it? What features were most useful, and what could be improved?*
- *(If no) Were you aware of its existence?*

## 4. Research Contributions & Impact

- What have been the main research outcomes from your use of the data (e.g., publications, theses, tool development)?
- How has access to these datasets influenced your research trajectory or enabled work that might not have been possible otherwise?

## 5. Strategic Vision (Prompts for Leadership Roles)

- What was the original, founding vision for the Centre?
- What do you see as the key lessons learned from the first decade?
- What is your vision for the Centre's future direction (e.g., new data types, new methods)?
- What do you see as the biggest challenges to sustaining the Centre's work?