The Art of Cybercrime Community Research

JACK HUGHES, University of Cambridge, United Kingdom
SERGIO PASTRANA, University Carlos III of Madrid, Spain
ALICE HUTCHINGS, University of Cambridge, United Kingdom
SADIA AFROZ, ICSI & Avast Software, USA
SAGAR SAMTANI, Indiana University, Bloomington, USA
WEIFENG LI, University of Georgia, USA
ERICSSON SANTANA MARIN, California State Polytechnic University, USA

In the last decade, cybercrime has risen considerably. One key factor is the proliferation of online cybercrime communities, where actors trade products and services, and also learn from each other. Accordingly, understanding the operation and behavior of these communities is of great interest, and they have been explored across multiple disciplines with different, often quite novel, approaches. This survey explores the challenges inherent to the field and the methodological approaches researchers used to understand this space. We note that, in many cases, cybercrime research is more of an art than a science. We highlight the good practices and propose a list of recommendations for future cybercrime community scholars, including taking steps to verify and validate results, establishing privacy and ethical research practices, and mitigating the challenge of ground truth data.

CCS Concepts: • **Security and privacy → Human and societal aspects of security and privacy**; • **Social and professional topics → Computer crime;**

Additional Key Words and Phrases: Cybercrime, communities, forums, marketplaces, data processing, ethics

ACM Reference Format:
Jack Hughes, Sergio Pastrana, Alice Hutchings, Sadia Afroz, Sagar Samtani, Weifeng Li, and Ericsson Santana Marin. 2024. The Art of Cybercrime Community Research. *ACM Comput. Surv.* 56, 6, Article 155 (February 2024), 26 pages. https://doi.org/10.1145/3639362

This work is supported by the European Research Council (ERC) under the European Union.s Horizon 2020 research and innovation programme (grant agreement No 949127) (for JH and AH), grant TED2021-132170A-I00 from the Spanish Ministry of Science and Innovation, funded by MCIN/AEI /10.13039/501100011033, and the European Union-NextGenerationEU/PRTR (for SP). The work is also supported by the National Science Foundation under the Secure and Trustworthy Cyberspace programme (grant No CNS-1936370) (for WL), under grants OAC-2319325 (for SS), DGE-1946537 (for SS), OAC-1917117 (for SS), CNS-2338479 (for SS), and award No 2246220 (for EM).
Authors' addresses: J. Hughes and A. Hutchings, University of Cambridge, JJ Thomson Avenue, Cambridge, United Kingdom; e-mails: {Jack.Hughes, Alice.Hutchings}@cl.cam.ac.uk; S. Pastrana, University Carlos III of Madrid, Avenida de la Universidad, Leganes, Spain; e-mail: Sergio.Pastrana@uc3m.es; S. Afroz, ICSI & Avast Software, San Fransisco, USA; e-mail: sadia@icsi.berkeley.edu; S. Samtani, Indiana University, Bloomington, 107 S Indiana Ave, Bloomington, IN 47405, Bloomington, USA; e-mail: ssamtani@iu.edu; W. Li, University of Georgia, C422 Benson Hall, 630 South Lumpkin Street, Athens, USA; e-mail: weifeng.li@uga.edu; E. Santana Marin, California State Polytechnic University, Pomona, 3801 West Temple Avenue, Pomona, CA, USA; e-mail: santanamarin@cpp.edu.

ACM 0360-0300/2024/02-ART155
https://doi.org/10.1145/3639362

1 INTRODUCTION

In 2007, the seminal article "An inquiry into the nature and causes of the wealth of internet miscreants" [50] was published. This study analyzed a cybercrime market, using seven months of data from an IRC channel. Now, over a decade later, there is an established body of work analyzing, understanding, and exploring these online platforms used as a place to share knowledge, tools, techniques, and socialize. Cybercrime communities have evolved from IRC channels to forums and cryptomarkets, and are increasingly moving to mobile chat platforms. Many articles have since been published, with authors from a variety of academic disciplines providing their own unique insights into how these platforms are used for cybercrime.

The main goal of this article is to provide researchers new to the field with an overview of what they would be building upon, and the scientific principles underpinning prior work. After analyzing 99 research articles on cybercrime communities, we conclude that in many cases, cybercrime community research represents more "art" than "science". Research is often exploratory, depending on the dataset and methods available, in contrast to scientific hypothesis testing. Cybercrime communities are difficult to study, with some of the main shortcomings including:

(1) Lack of available data: Collections of cybercrime community data have only recently become available, with early research constrained to those with the resources to collect their own datasets, or to small subsets of community posts. Data collection requires setup and ongoing maintenance, which can increase research time, and may not be comprehensive. More recently, shared datasets have become available to avoid this issue, however, they need to be kept up to date to not become stale. Indeed, whereas in other disciplines historic data might be actual and relevant, some types of cybercrime are rather volatile and unforeseen, e.g., COVID related scams, or cryptomining. Thus, it requires agile (and often not too concise) methods for data collection and analysis.
(2) Lack of validation and generalization: Due to the adversarial and hidden nature of cybercrime, ground truth can be particularly difficult to establish. Furthermore, findings from one cybercrime community, or for one type of crime, may not generalize to another.
(3) Lack of reproducibility: This falls into two concerns, namely providing sufficient detail to enable other researchers to reproduce the study faithfully, or through sharing datasets and code, but also that research findings are strengthened if they are reproduced by others. Often, researchers focus on novel approaches and application areas, and reproduction of prior work is not incentivized. Most research is exploratory, with very little prior work testing formulated hypotheses.
(4) Little consensus about the ethics of working with data, although some norms have emerged. The little consensus is particularly concerning with cybercrime data, since the processing of this data can cause additional harm to users if they are later prosecuted [69, 70].
(5) Privacy and legal issues when working with leaked and scraped private data. In this case, the data might contain relevant cybercrime information (e.g., actual purchases for hacking material) intermixed with data from licit users (e.g., personal messages). Thus, processing requires careful protections to keep data private.

We systematize the literature relating to cybercrime communities, from measuring the growth of multiple platforms over time, to estimating the proceeds from specific marketplace activities, for informing new researchers to the field of the possibilities, as well as common pitfalls, limitations, and issues that arise when doing research in this space. We start by providing a broad overview of the field (Section 3), describing what cybercrime communities are (Section 3.1), and goals of cybercrime community research (Section 3.3). We then explore the methodological approaches used in underground community research (Section 3.4), and explore ways to classify text, images and

attachments data, including tools for categorizing threads or understanding the social networks based on posting activities. Next, we discuss the main challenges and limitations (Section 4), both inherent to the field and weaknesses of prior research, including limited ground-truth datasets, and ethical considerations of research projects. Finally, we provide an overview of recommendations for future researchers and present future challenges in the field (Section 5). Also, we consider the different data sources being analyzed (Section 3.5), including how datasets are collected and shared.

2 SCOPE AND DEFINITIONS

In the simplest terms, cybercrime consists of any criminal activities (such as fraud, theft, or distribution of child pornography) committed using a computer especially to illegally access, transmit, or manipulate data.[1] To understand cybercrime communities, we consider platforms hosted on both the *dark* web (the part of the web inaccessible by standard browsers and not indexed by common search engines) and also the *surface* web, where discussions of cybercrime and advertisements for cybercrime products/services take place. We also include "cryptomarkets" that are primarily used as drug markets with a focus on cybercrime activity. Note that we *exclude* from this scope hate and harassment platforms, so as not to duplicate existing work [124].

To find relevant research articles on cybercrime communities, we start by using keywords related to cybercrime forum and marketplace communities, namely "cybercrime forums"; "cybercrime communities"; "cybercrime marketplaces"; and "underground forum analysis". We used keywords on Google Scholar to find articles, and for each article, we expand our search using both the reference list of articles, and who the articles were cited by. Using this snowballing approach, we filtered articles to only those that fit our scope of *cybercrime communities*, including cybercrime forums, marketplaces, and chat channels (e.g., IRC and chat platforms), resulting in a collection of 99 articles. We systematize the methodological approaches used in these articles, exploring what is being measured and how, from the micro to the macro.

3 OVERVIEW OF THE FIELD

Table 1 shows the results of our review and the articles that were included. First, we explore the sources of data used, whether these made use of leaked data sources, obtained through sharing agreements, scraped for the article, or reused. Next, we indicate if the authors used multiple sites for their analyses, or relied on a single cybercrime community. We further indicate if they used the entire site, or instead explore a subset, such as a bulletin board or posts within a limited time period. We explore whether the authors indicate they obtained approval from a **Research Ethics Board** (**REB**), such as an **Institutional Review Board** (**IRB**) in the US, or an ethics committee, or if they were considered exempt. We note that some articles may have REB approval, but did not disclose this in the article. We highlight if the authors used English language and non-English language communities. For validation and evaluation, we indicate how authors obtained or created ground truth data, if they removed any outliers in their method, and if they used external data sources for verification or enrichment.

The articles in the review have a diverse set of methods, with very few taking the same approach. At most, we have 5 articles in a grouping within the table. We noticed that the majority of the articles focus on English forums (96.0%), scrape data on their own (54.5%), and for articles that use ground truth, the majority of these rely on manual annotations (40.6%). The majority of the articles do not remove outliers (87.9%) or consider external verification (76.8%) to validate their results. Most concerning is the fact that most articles do not mention REB approval or exemption (82.8%).

[1] https://www.merriam-webster.com/dictionary/cybercrime

Table 1. Common Issues and Limitations

	Source of data				Extent of sample		Research ethics		Language	Ground truth			
Research articles	Leaked	◐Reused/ scraped ○Scraped for article ●Shared/ reused			●Multiple forums ○Single forum	●Full dataset ○Subset	●REB approval ○REB exemption		●Multiple ○English ◐Non-English	●Manual annotation ○Unsupervised ◐Other		Outliers removed	External verification/ enrichment
[7, 79, 80, 112, 113]		●			●	○			●	◐			✓
[6, 8, 78, 104]		◐			●	○			●	◐			✓
[17, 38, 54, 62]		○			●	○			●				
[2, 52, 61, 87]	✓				●	●			●				
[39, 132, 133]		○			○	○			○	●			
[49, 103, 126]		○			●	○			○				
[107, 108, 128]		○			●	○			●	○			
[1, 4, 26]		●			○	○			○	●			
[12, 57, 59]		●			○	○	●		○	●			
[74, 131]		○			○	○			○				
[50, 105]		○			○	○			○	●		✓	
[75, 109]		○			○	●			○	○			
[42, 72]		○			●	●			●	●			
[35, 122]		●			●	○			○	◐			✓
[114, 115]		●			●	○			○	○			
[43, 66]	✓				○	●			○	◐			
[127]		◐			○	○	●		○			✓	✓
[129]		◐			●	○			○	◐			✓
[121]		◐			●	●			○	○			
[67]		◐			●	●			○	●			✓
[56]		○			○	○			◐	○			
[5]		○			○	○			◐	●		✓	✓
[76]		○			○	○			○	◐			
[41]		○			○	○			●	●		✓	
[46]		○			○	○	○		○	●			
[60]		○			○	○	●		○				✓
[106]		○			○	●			○				
[135]		○			○	●			○	●			
[71]		○			○	●			○	●			✓
[55]		○			●	○			○	○			
[134]		○			●	○			○	●			
[88]		○			●	○			○	●		✓	
[130]		○			●	○			●	○		✓	
[73]		○			●	○			●	●			
[81]		○			●	○			●	●		✓	
[120]		○			●	○	○		●	●			✓
[45]		○			●	●			◐	○			
[91]		○			●	●			○				
[111]		○			●	●			○	●			
[51]		○			●	●			●				
[53]		○			●	●			●	◐			
[18]		○			●	●	○		●				
[14]		●			○	○			○				
[63]		●			○	○	●		○				
[89]		●			○	●			○			✓	
[25]		●			○	●			○	◐			✓
[92]		●			○	●	●		○	●			✓
[84]		●			●	○			◐				
[77]		●			●	○			○	○			
[102]		●			●	○			○	●			
[110]		●			●	○			○	●		✓	✓
[9]		●			●	○			●	○			✓
[15]		●			●	○	●		○				
[93]		●			●	○	●		○	●			✓
[97]		●			●	○	●		●				
[82]		●			●	●			○	◐		✓	
[101]		●			●	●			○	●			
[16]		●			●	●			●	◐			
[40]	✓	●			●	○			○	●			
[19]	✓	●			●	○	○		●	●		✓	
[37]	✓				○	○			○	○			
[36]	✓				○	○			○	●			
[65]	✓				○	●			○				
[3]	✓				●	○			●	◐			
[90]	✓				●	○	○		●	◐			
[44]	✓	○			●	○			○	◐			
[100]	✓	○			●	○			●	●			
[118]	✓	○			●	●	●		○	●			
[119]	✓	●			○	●	○		○	●			

A minority of articles have used data from leaked datasets (18.8%), raising privacy concerns. These datasets may contain personal data in private messages, and account data such as e-mail and IP addresses which could be used to identify users. In addition, data obtained from public sources including scraped or shared datasets, may contain personal data in public posts. This could include personal information shared on a forum because an individual has been doxxed. Although this data is public, it does not exempt researchers from potential privacy violations.

3.1 Cybercrime Communities

A cybercrime community is an online platform (e.g., forum or marketplace) where the members engage in cybercriminal activities. The research community has analyzed a variety of cybercriminal communities, with researchers often focusing on forums analyzed by prior work, which makes a limited number of forums overly analyzed and many new forums relatively unexplored. While some communities are invite-only to limit membership to trusted individuals, this necessarily restricts all users, not just researchers. Cybercrime communities have varying levels of criminality, with a mix of discussion on both crime and other topics [12], used to build community and trust.

Cybercrime communities are purpose-driven. Despite differences in online communities, there is some similarity in the way forums are commonly set up and structured [59, 94]. Forums tend to be broken down into boards or subforums, each relating to a topic or theme of conversation which members can contribute to. Boards contain threads: an ordered set of posts around a specific theme or question set by the first post. While longer threads may vary from the original topic ("off-topic"), moderators on the website may choose to close threads that do so. The first post in a thread is significant in comparison to the replies: the user has chosen to start a new conversation topic, to propose a new idea or ask a question, and replies may either contribute back to the first post, or to later replies. Each thread will also have a title.

Marketplaces are structured differently [30]. Used for advertising goods and services for sale rather than discussion, they are usually segmented by "department" for types of goods. Listings can be sorted and filtered. Each listing has a title, and also may have a description of the item, price, seller information (including their username and current rating), and a section for feedback or reviews. Compared to forums, the domain of text appearing in posts is typically smaller: posts advertising items will often provide a title, a description of the product, and a price. While extracting information from these is not straightforward, it can be a simpler task than extracting useful information from the free-form text discussions that appear in forums, which require processing natural language.

Marketplaces and forums thus have some structure, provided by categories and boards and other moderator-chosen types of structures, which is useful for analysis. This structure also differentiates them from other platforms such as Discord, Telegram, and IRC [38]. However, despite the structured nature of forums and marketplaces, extracting useful information from the threads and advertisements is not a trivial task [58]. We note that some communities might contain a mixture of the different types [127], e.g., a marketplace with a dedicated forum, or a forum with an embedded chat.

Also, forum datasets may contain additional information which are unique to the forum. One example is reputation voting data, where each user has a reputation score which may signal trust, and other users can send positive or negative reputation points to provide feedback [92].

3.2 Typical Methodology of Cybercrime Community Research

Research steps can fall into one of two categories: problem driven, when a researcher is acquiring data and analyzing for a specific use-case, and data driven, when a researcher has an existing

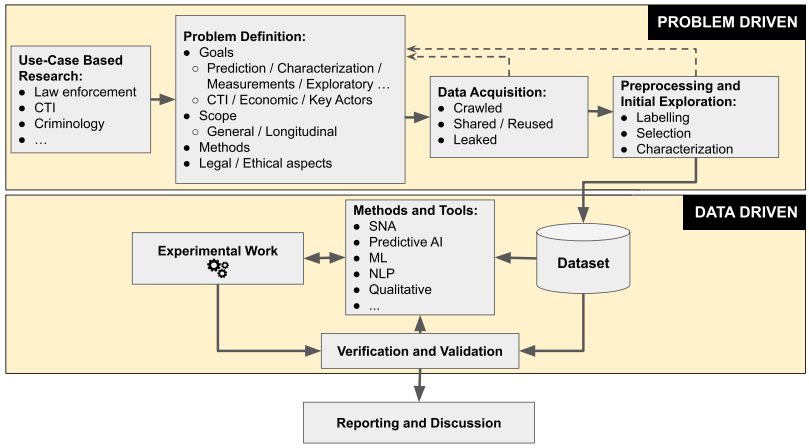

Fig. 1. Overarching methodology for cybercrime community research.

dataset to develop new analysis methods or advance prediction performance for a given task. Figure 1 shows a typical methodology used for cybercrime community research.

In the problem stage, use-cases typically vary depending on the purpose of the research, including law enforcement, **cyber threat intelligence** (**CTI**), or criminological. The use case is used to formulate a problem definition: setting out the goals, scope, methods, and legal and ethical aspects to the work. Data is then acquired by the researcher, pre-processed, and explored. This step can be iterative, as the problem definition may be further refined or changed. Then, researchers carry out the data driven part of the task.

In the data driven stage, various methods and tools are used. These are selected according to the goals, and may require experimental work to choose the optimal or a new approach. This is followed by verification and validation. Again, the details of this step depend on the method selected, but could include checking for outliers, checking for anomalous results, and using external data sources to verify results. Finally, researchers will report and discuss their results.

3.3 Goals of Cybercrime Community Research

We annotated each article included in Table 1 according to their goals. We first developed the categories **key actor detection**, **key actor analysis**, **longitudinal**, **economic**, and **cyber defence** from our prior knowledge of the area. These are selected to categorize the goals within the diverse corpus of articles, and we note that each article may have multiple goals. We then extended with three categories, namely **discourse** with topics and concepts discussed in communities, **subcultural** exploring aspects of community culture, and **crime type** for research focusing on specific activities. We note that each article may have multiple goals. These are displayed by year of publication in Figure 2, which shows that the rate of publication greatly increased around 2016.

This analysis illustrated how prior research in the field has focused on different tasks: forum-wide measurements [94], measurements of forums over time (**longitudinal**) [3], analyzing specific groups of members (**key actor analysis**) [16, 92, 120], classifying types of cybercrime (**crime type**) [12], and classifying posts, threads, and products [26]. General measurement studies have been used with forum datasets, for introducing new datasets [18, 38, 94], exploring the evolution of different platforms [117], or movement across forums [49]. Measurement studies [2, 4, 5, 50–52, 71, 87, 91, 127, 130, 131] include various tasks, such as counting products, counting people, analyzing the economy of a marketplace (**economic**), or looking at social demographics

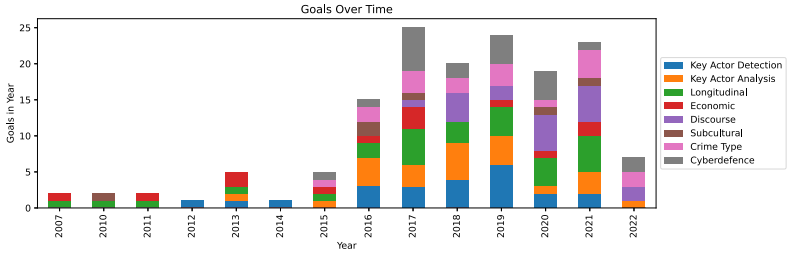

Fig. 2. Goals of Articles Over Time.

at scale. Researchers have looked at forums across a range of languages including English [94], Russian [3], and Arabic [53].

As in every meritocratic system, online hacker forums have participants with different levels of knowledge and influence. An emerging area of research is to identify underground communities to identify key cybercriminals (**key actor detection**) [46, 59, 65, 66, 92, 132–135]. Researchers predict if actors belong to a subset of members, such as "key actors", "proficient cybercriminals", "expert hackers" or "key-hackers" [1, 45, 77, 82, 106], where individuals organically stand out for their high reputation when compared to the vast majority of forum members.

Rather than looking back over prior user behavior, some researchers aim at detecting what new topics are the focus of discussion (**discourse**) [57, 121]. The use case for such research is primarily to detect emerging threats [35, 37, 75, 101, 102, 109, 111]. Others aim at uncovering the type of jargon used within underground communities [73, 114, 115, 129]. Stylometry is also used to identify users who may be operating multiple accounts within and across forums. Afroz et al. [3] use lexical, syntactic, and jargon features to detect multiple accounts, and analyze their feature set by identifying those with the greatest information gain.

Analyzing cybercrime subcultures is another key research objective. The **subcultural** category includes how trust is gained and lost within communities [39, 54], performances of masculinity and perceptions of gender [14], and the use of aggressive language on forums [25]. Other researchers explore how underground markets are governed and organized [89]. A wide variety of malicious activities thrive in cybercrime communities, which are explored in the **crime type** category. Researchers have analyzed currency exchange activities [12], eWhoring (a type of fraud where intimate images–often stolen–are used to simulate sexual encounters for financial gain) [63, 93], money laundering [84], malware [53, 61, 74, 77], and credit card fraud [126]. Others have categorized illicit products and services across forums and marketplaces [12, 40, 43, 44, 67, 81, 100, 122] and identified potential supply chains [19].

Most organizations do not have resources to patch the increasing amount of vulnerabilities disclosed every month, and many of those software flaws are never exploited in the wild [6, 8]. The idea of predicting software vulnerability exploitation is to prioritize the remediation of vulnerabilities that are likely to be targeted by malicious actors (**cyber defense**). Cybersecurity researchers explore the opportunities for early exploit detection by mining information about software vulnerabilities shared in underground forums together with security advisories published on white hack communities [6, 88, 105, 110].

A body of research attempts to predict future attacks. This includes correlating malicious activities in underground forums and marketplaces with cyber-incidents collected from the logs of real-world enterprises [7, 79, 80, 112, 113]. Another approach studies adoption behavior among community members to predict their future activities [76, 78]. Values, ideas and techniques are transmitted from one person to another, and this behavior is also observed for malicious actors [47, 76, 85, 99, 116].

3.4 Methods for Cybercrime Community Analysis

This section categorizes methods used for analysis of cybercrime communities. We differentiate between two overarching methodological approaches: quantitative analysis methods, typically involving large-scale measurements and statistical models, and qualitative methods involving human reviewers reading and analyzing a subset of forum data. We further break down quantitative approaches into **Social Network Analysis** (**SNA**); **Natural Language Processing** (**NLP**) approaches including Text Analysis & Analytics, and Topic Modeling; and **Machine Learning** (**ML**). For qualitative approaches, we explore content analysis, including its use for crime script analysis. Figure 3 shows the methods used by articles over year of publication (note that some articles may use several methods).

Social Network Analysis. Understanding the social interactions of cybercriminals is of great interest, since cybercrime is often fueled by a rich and active supply chain of products and services [19, 119]. SNA techniques are used to investigate, describe, and predict the overall relational cybercrime community network structures, and identify key-hackers. [6, 29, 56, 78, 79, 82, 98, 104, 112]. Nodes, clusters, and relations can represent an approximation of the communication structure and position of individuals within communities.

Text Analysis and Analytics. Text analysis and analytics uses techniques from ML, NLP, and linguistics to extract measurements and insights from text data such as stylometry analysis [3] and hacker terms identification [81, 82]. This includes using embedding models (e.g., word2vec or BERT), and text graph representations. In cybercrime communities, using these techniques is particularly challenging due to the high use of changing jargon [115, 129].

Forum datasets may also include attachments, including tools for carrying out cybercrime, e.g., snippets of code [122]. ML and NLP approaches have been used to classify these attachments for potential threats [108] and into known categories [128].

Topic Modeling. Topic models are capable of discovering the semantic themes (i.e., topics) within discussions and advertisements, by capturing associations among keywords (e.g., slang, new terminology, or product names) from an unlabeled dataset. These methods are useful to summarize large datasets containing slang, without needing to build a training dataset, and will continue to work with new unseen slang. Topic models require validation of results to ensure these are not meaningless. This includes choosing a suitable number of topics, which can be estimated using the coherence scores to identify cohesion, and can be manually checked using domain knowledge of words in each topic, combined with reading sample posts in each topic. Also, researchers may need to clean text data, such as removing URLs.

One commonly used topic model for analyzing cybercrime community datasets is **Latent Dirichlet Allocation** (**LDA**) [20]. LDA is used with a corpus of documents (e.g., posts in forums) to discover topics represented as distributions over words, and can additionally identify the topic distribution of each document. LDA has been applied to attachments and tutorials to discover topics in hacker assets and threats [37, 107], and to identify topics discussed by key actors [72, 92].

Machine learning approaches. Existing research using ML techniques have primarily focused on detection and categorization tasks. Initially, categorization tasks were used on hacker communities, leveraging off-the-shelf unsupervised learning techniques, including topic modeling and clustering. [36, 37, 72, 81, 107, 108]. These techniques enable further understanding of content and trends on these forums.

Due to links of cybercrime communities with real-world attacks and frauds [69, 92], researchers moved toward predictive techniques in an attempt to anticipate potential incidents, and also

focus on the analysis of actors of interest (community members of interest). Applications include automatically classifying malicious hacker content into pre-defined exploit labels [9, 128], predict software vulnerability exploitation and enterprise cyber-attacks [6, 79, 80, 112], and developing adversarial learning and cross-lingual knowledge transfer techniques for detecting cyber threats [41, 42]. ML predictive approaches typically require feature engineering. Features used can be split into either metadata or text-based. Metadata-based features are based upon data obtained directly from the dataset, such as times of posting activity or members posting in the same threat to detect communities. For example, these have been used to predict where members may send a private message by training a model on sent public messages [90, 119], for later use on datasets which may not contain private messages. These use a combination of features about the user's posting activity in addition to text-based features. Text-based features are commonly used for classification models, such as for detecting certain types of activities across a forum dataset.

ML approaches have been combined with other techniques, combining with NLP tags and SNA features to analyze key actors on forums [92], and combining topic models with SNA metrics to rank forum members [55].

Qualitative approaches. Qualitative research approaches are useful for exploratory research into problems where there is little already known about the topic [60, 63], such as cybercrime communities, which tend to include hidden and small populations. Many researchers also use a mixed methods approach, in which they combine the rich insights obtained from qualitative approaches with quantitative measurements [14, 118]. Qualitative research on cybercrime forums tends to be passive, content analysis, instead of interviews or ethnographic research. Qualitative research allows researchers to develop models, typologies, and theories to describe and explain issues. Theory that is built upon qualitative data in such an inductive approach is called grounded theory. Other research may take a deductive approach, where existing theory (for example, criminological theory about how criminal behavior is learned [61]) is applied and tested.

An example of a deductive approach is crime script analysis. Crime script analyses are built upon the idea of "schemata" from cognitive science, namely that people have basic understanding of how to interact in various social settings [33]. Applying this understanding to criminal activities allows us to map out how specific types of crimes are carried out, to gain insight into quite complicated crime types. In relation to cybercrime communities, crime script analysis has been applied to understand stolen data markets [62], credit card fraud [126], travel fraud [60], and eWhoring [63]. Qualitative research has also been used for examining the ecosystem around the Internet of Things [15] and social behavior inside marketplaces [103].

Researchers need to carefully consider how to present research findings, and skillful writing is required. Furthermore, qualitative research can be quite time consuming and resource intensive, which is often under-estimated (and poorly understood by many researchers who are not familiar with the process).

3.5 Data Sources

A key issue in research on cybercrime communities is the acquisition of data required for the analysis. Researchers either rely on their own collection technologies, i.e., web crawlers and scrapers [125], or use existing datasets available for research. Some researchers gain access to datasets collected by law enforcement or security companies and made available under a NDA (e.g., [61]), which may limit reproducibility. In this section, we describe current methods and goals, systematize the steps needed for data collection, and provide an overview of datasets of cybercrime communities that are available for academic research. We also discuss data enrichment by means of external sources.

3.5.1 Collection Methods. Depending on the research needs and the available resources, there are different methods for data collection. Our taxonomy of existing collection methods for forums and markets is based on the method used and the desired scope.

Method. Crawling large communities at scale typically requires the use of automated tools. However, these are not always needed or available. Three possible collection methods when sampling content from forums include:

- *Manual collection* requires investigators to manually visit the sites online, selecting and storing the required content locally (typically, in text format). The method is valid in cases where a small sample is needed, and the effort required for the development and use of automated tools is not worthwhile. Indeed, this is the only collection method available to researchers lacking technical skills to develop or use automated techniques.
- *Bulk crawling* uses automated tools to fetch and store raw files for offline processing. Links (URLs) are obtained using regular expressions, and are visited indiscriminately. Crawlers can use allow- and deny-lists, e.g., to limit crawling to a single community, or prevent link-traps that automatically close sessions or remove accounts [125]. Bulk crawlers require low engineering effort to be scalable, but can obtain useless data and are sub-optimal, since they demand high resources (i.e., storage and network bandwidth).
- *Targeted Crawling* use custom scraping technology to adapt to the particular site being monitored. This way, while the content is automatically crawled, its information is being processed online (scraped). The collection can therefore be focused to the desired pages, at the cost of requiring custom adaptations for each site being monitored.

Scope. The research goals of a project determine both the spatial and temporal dimension of the crawling, and also the type of content collected.

- *Liveliness.* A crawler provides the state of the online community at a given time t [34]. However, cybercrime communities are dynamic and volatile (e.g., new market products might be removed once they are sold). To understand the dynamics of these communities, and get the most current content, data collection must also be dynamic. For this, two options are available. First, to perform *incremental* crawling: after an initial snapshot, fetching and storing new content. Second, to collect complete snapshots at different times, which can be compared to understand not only the new content added, but also content removed or modified (e.g., price changes or post views). The use of incremental crawling is faster and optimal and the preferred option for large and highly volatile communities, e.g., forums. The collection of multiple snapshots allow for a proper analysis of items being removed or modified, by means of diffing tools, but they require re-visiting and storing the site entirely, and thus are more suitable for small sites.
- *Comprehensiveness.* Cybercrime communities contain miscellaneous content, e.g., boards for discussing politics as well as trading cyber-weapons, or markets selling both drugs and virtual items. Depending on the research needs the crawling can be for specific sections or the entire site. Collecting all data requires more effort than focusing on particular areas. However, having a complete picture allows for a broader analysis, e.g., to understand pathways into crime. Most cybercrime forums contain many licit content, where community building and trust building take place. The relevancy of these sections depends on the research goal, and researchers with limited resources may choose to not collect such data.
- *Content.* Collection can be restricted to textual content only, or to also download media content and other artifacts (e.g., binaries or documents). The former is simpler, and still can include the URLs to the linked media or attachments for its collection afterward (at

Table 2. Summary of Goals and Ethical Issues Considered in Previous Works on the Data Collection Process (✓=documented, -=not documented,not Specified or Not Applicable, C=Complete, P=Partial)

Related work	Year 20XX	Ethics	Method	Scope		Challenges					Content
				Completeness	Incremental	Accessibility	Connectivity	Anti-bot	Anti-crawling	Maintenance	
[51]	10	-	Targeted	C	-	✓	✓	✓	✓	✓	Text; attachments
[103]	10	-	Manual	P	-	-	-	✓	-	-	Text
[30]	14	✓	Targeted	C	✓	✓	✓	✓	✓	-	Text (markets)
[75]	15	-	Bulk	C	-	-	-	-	-	-	Text
[117]	15	✓	Targeted	C	✓	✓	✓	✓	✓	-	Text (markets)
[132]	15	-	Targeted	P	-	✓	-	-	-	-	Text
[62]	15	-	Manual	P	-	-	-	✓	-	-	Text
[88]	16	-	Targeted	P	✓	✓	✓	-	✓	-	Text
[64]	16	-	Bulk	C	-	✓	-	-	-	-	Text
[39]	16	-	Targeted	P	-	-	-	-	-	-	Text
[54]	16	-	Manual	P	-	✓	-	✓	-	-	Text
[126]	16	-	Manual	P	-	-	-	✓	-	-	Text
[100][40]	17	-	Targeted	P	-	-	-	-	-	-	Text
[5]	17	✓	Manual	P	-	✓	-	✓	✓	-	Text
[53]	17	-	Targeted	P	-	-	✓	-	-	-	Text; attachments
[108]	17	-	Targeted	C	-	✓	✓	✓	✓	-	Text; attachments
[94]	18	✓	Targeted	C	✓	✓	✓	✓	✓	✓	Text
[128]	18	-	Targeted	C	✓	✓	✓	✓	✓	✓	Text;Attachments
[38]	18	-	Targeted	-	-	-	-	✓	✓	-	Text
[60]	18	✓	Manual	P	-	-	-	✓	-	-	Text
[28]	19	-	Targeted	C	-	-	✓	✓	✓	✓	Text
[44]	19	-	Targeted	P	-	-	-	-	-	-	Text
[130]	19	-	Targeted	P	-	✓	-	-	-	-	Text
[127]	20	✓	Targeted	P	-	✓	✓	-	-	-	Text (Trades)
[67]	21	-	Targeted	P	-	-	-	-	-	-	Attachments
[55]	21	-	Targeted	P	-	-	-	-	✓	-	Text
[118]	21	✓	Targeted	P	-	-	-	-	-	-	Text

the uncertainty of these link expiring). Downloading non-textual content allows for a more complete analysis, e.g., image banners advertising products or services. However, they also put researchers at risk due to legal concerns (e.g., for downloading illegal images).

Table 2 summarizes the collection methods implemented and the challenges discussed in Section 4 for the works that explicitly describe the data collection process.

3.5.2 Steps for Data Collection. Data collection is an engineering project, composed by a set of steps, including:

(1) Investigating and selecting the communities for the research, i.e., the sites to be crawled. Analyzing the access requirements and studying the Terms of Service. Consulting an IRB or equivalent for legal and ethical advice (see Section 4.1.1).

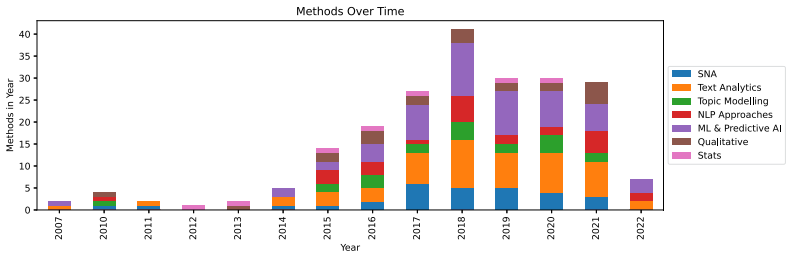

Fig. 3. Methods of Articles Over Time.

(2) Gaining initial access. If needed, registering accounts using disposable e-mails, or creating custom accounts.
(3) Conducting preliminary observation, including manual navigation and content inspection. Collecting and storing a small sample subset to design and test the crawler offline.
(4) Defining the scope, such as collecting a single snapshot, or periodic re-visits, classifying the content that needs to be extracted, selecting those areas of higher interest or priority.
(5) Crawler design and implementation. Defining the database, scheduling incremental crawling (if needed), designing custom scrapers, managing accessibility (e.g., using session cookies), implementing anti-crawling bypass techniques, and so on.
(6) Crawler deployment. Configuring the infrastructure, e.g., deploying database, installing necessary software and preparing connections through VPNs, proxies, or Tor Circuits.
(7) Production. Tracking logs for status monitoring, maintenance, and error management.
(8) Post-crawling. If needed, conducting incremental crawling, re-designing scrapers for site updates, adding further anti-crawling bypass methods, and so on.

3.5.3 Datasets. To avoid the tedious task of web crawling and scraping, several works use existing datasets of forum and market data. These datasets come from leaked databases and public data repositories (containing either scraped, leaked or both). Data repositories allow researchers to quickly get started with data analysis without needing to set up infrastructure and collect data themselves. In addition, data repositories can hold data collected over a longer period of time, and support reproducible results. There are three main repositories for cybercrime forums and markets.

The **DarkNet Market Archives (DMA)** is a repository originally collated by Branwen, with later contributions by others [22]. It contains scrapes from 89 DarkWeb markets and more than 37 forums, totaling 1.6 TB of data. The data is now quite dated (covering 2011–2015, with partial scrapes from 2017), and therefore not representative of the current cybercrime landscape. Still, these datasets serve as a good benchmark and are still being used in cybercrime research. This dataset has been used in more than 70 studies [22].

AZSecure is a repository providing selected hacker community content [13]. It contains various datasets publicly available for researchers, obtained by scraping forums, markets, IRCs, and carding shops. It contains data from 51 forums, totaling more than 32 m posts, and 12 markets, totaling 249 k listings. Also, it includes a dataset of hacking assets, i.e., attachments and source code. It spans more than a decade, with the latest dataset being from 2019.

The **CrimeBB** dataset of cybercrime forums is available to academic researchers under a legal agreement with the Cambridge Cybercrime Centre to prevent misuse [94]. It contains data from over 100 million posts scraped from 34 forums, dating back more than 20 years, and it is kept updated through regular incremental crawlings. The same repository also contains other assets, like attachments obtained from Game Cheating communities [67], or contract data of actual tradings occurring in cybercrime marketplaces [127]. The dataset has been used for at least 65 studies.

The main strengths of this dataset is that it is of easy access, and provides frequent updates. Until recently, it was only provided as SQL binary dumps, which presented technical challenges for non-technical researchers [95]. To address these challenges, a search interface was developed for interdisciplinary researchers [96].

Finally, some researchers also provide the datasets used for their studies to allow for reproducibility and foster new research. For example, Portnoff et al. [100] and Durrett et al. [40] provide the leaked and partially scraped datasets used for the automated analysis of cybercriminal markets. Also, Yuan et al. [129] provide the processed data (originally obtained from the DMA repository) used for the understanding and analysis of jargon in forums.

3.5.4 Data Enrichment. To address data quality and validation issues, some researchers are able to verify data using external sources ("data enrichment"), with 21.7% of articles in the review using this type of data. For example, Vu et al. [127] verified marketplace transactions using Bitcoin on the blockchain, and Pastrana et al. [92] used public reports of forum actors who had been arrested or prosecuted for cybercrime offenses to as ground truth for actors of interest to law enforcement ("key actors"). Almukaynizi et al. [6, 8] used anti-virus or **intrusion detection systems** (**IDS**) attack signatures provided by security companies such as Symantec to validate the software vulnerabilities to be exploited in future. Similarly, to validate the results produced by cyber-attack prediction models, previous works have used historical records of cyber incidents recorded from the logs of real-world enterprises [7, 79, 80]. Other data sources used for data enrichment include X (formerly known as Twitter) [111], CVEs [5, 105], ExploitDB, and OSINT reports [5, 111].

4 COMMON ISSUES AND LIMITATIONS

Research of cybercrime forums and marketplaces can be affected by data issues and method limitations. We systematize some of the approaches taken in existing studies, and types of limitations they face. For this section, we rely heavily on our review of the articles as detailed in Table 1. Common issues and limitations can be split into two parts: challenges that are inherent to the field, and limitations of prior work to be addressed by future research.

4.1 Challenges Inherent to Research on Cybercrime Communities

Data collection challenges. Crawling cybercrime communities poses several challenges, which increases the complexity of the process [125]. Whether these challenges need to be addressed or not depends on the particular communities being monitored. Most common challenges are

- Accessibility. Gaining initial access to communities might not be straightforward [5]. Some sites are open to everyone. Others require registration, often by means of a valid e-mail address for account verification. Registration is often free of charge, but in some cases a fee is payable (see Section 4.1.1).
- Connectivity. To remain active, crawlers need to connect through various sources, typically using web proxies or VPNs. This provides robustness against IP blocking, and also preserves the anonymity of researchers. If the site is hosted on a hidden service only reachable using Tor, the crawler must provide proper circuit management, offering various exit nodes in case some of the circuits fail [94].
- Bot-detection methods. Online sites might attempt to detect and ban bots. This relies on techniques such as anomalous detection of networking patterns, or monitoring accesses to old content (which probably no one visits). To prevent this, crawlers must mimic human behavior, for example randomizing accesses to content, establishing delays, or following human navigation and connectivity patterns. These techniques degrade the crawling operation [28].

— Anti-crawling methods. In addition to bot-detection, crawlers face anti-crawling methods [27]. The most common is the use of *Captcha* challenges. These can be technically bypassed using automatic solvers, but these services are questionable from an ethical (and legal) view [86]. A second option is to solve the captcha manually, keeping and re-using session cookies for following connections. Also, sites can implement traps to deter bulk-crawling methods, e.g., linking pages that lead to loops, or providing links that automatically delete the account.
— Maintenance. Online communities are dynamic. Despite changes in the content, which require implementing incremental crawling techniques (see above), the underlying HTML structure can change as well, i.e., as part of an update. Thus, crawling must be robust and modular enough to easily adapt to changes with low engineering effort. Also, it is desirable to put in place a logging system to facilitate status monitoring and error management [51].

Unstructured data. One limitation inherent to research of cybercrime communities is the lack of structured data, which thwarts its analysis at scale. Often, researchers rely on text analysis methods, or limit their research to just metadata. Other researchers focus on the social network structures found in forums and communities, but as the network is implicit, researchers rely on assumptions when building a network from post data. Language used on cybercrime communities contains jargon and specific language, which can be due to obfuscation of illegal activities, or just due to the way they talk [129]. Research methods using text analysis with "off-the-shelf" NLP models trained on standard language may not perform effectively in this field, and require further fine-tuning with manual annotations [26].

Use of leaked data. Leaked data provides researchers privileged access to non-public information, e.g., private messages or login details. These are useful for studies that aim at correlating public and private interactions [90] or analyze actual trading from private messages [87]. However, using this data has drawbacks. First, there are ethical concerns (see Section 4.1.1). Second, it suffers from the "observer effect": once actors know about the database leakage, they may change their patterns or move to another forum. Third, these datasets provide a single snapshot, and are rapidly outdated.

Incomplete and volatile data. Collected datasets, through scraping, leaking, or data sharing, may appear to be immediately useful. However, if researchers assume that the data is correct and useful without further inspection, this can lead to incorrect results [34]. Further to the lack of structured data available, datasets collected may not necessarily be complete collections, and authors should take this into account when measuring the whole forum. This can be due to platforms preventing researchers from scraping them, by using bot detection techniques. Other adversarial techniques used by forums may try to slow down scrapers by using rate limiting, or use adversarial ML techniques to make text analysis and classification difficult [125]. In addition, a dataset may not be complete as older posts or boards can be deleted by users or moderators. In some cases, entire sections of a forum may be deleted by administrators, as happened in October 2016, when HackForums removed the "Server Stress Testing" section and banned booter services (which offer denial of service attacks for a fee) from advertising [31]. In some of these cases, repositories have been able to reconstruct missing threads from existing datasets [94]. Some researchers choose to focus on a subset of communities, either due to limited resources (such as time) or to avoid working with large datasets that analysis tools may not handle. Still, different from other communities, cybercrime community research requires agile and often non-validated analysis methods to understand abrupt activities that occur.

Lack of ground truth. Usually, analysis is built on top of imperfect or missing labels. This presents issues with evaluation. Ideally, researchers have access to "ground truth" labels, but these are

typically manual annotations, and for large datasets this task is often outsourced to non-domain experts. This is highly time consuming, and is likely to be constrained by the availability of resources available to the researcher. This has implications for performance evaluations, as models may not necessarily be predicting what they were designed to predict. Also, data from cybercrime communities are particularly challenging to label because they often include non-conventional vocabulary, and users refer to their activities in different terms as one would expect in other communities (e.g., using jargon, or using coded sentences to disguise illegal behavior).

Also, researchers have used user-generated forum data to validate results, such as reputation voting data [82]. Researchers may need to consider if these are a trustworthy or untrustworthy signal, as votes can either be sent in good faith or be attempts to game the reputation system. Ground truth for validating results from key actor identification or community detection requires researchers to either manually review results which can lead to subjective results, or use a metric to validate, which may not reflect the definition of key actor used by the researchers. External verification and validation is one of the biggest hurdles faced by researchers, particularly given the adversarial nature of the cybercrime population, where there may be incentives to present a different version of the truth, and also the potential complexities to access external data from companies.

Ground truth for social network reconstruction is not possible to obtain with forum post data, as social connections between users are not explicitly available in comparison to social networks, e.g., using a friend or follower feature. It is, however, possible to build a proxy social graph using signals (e.g., replies in a thread), but these may not be accurate. Care must be taken to ensure results do not over or underestimate the scale of social networks. Alternatives to ground truth include unsupervised ML, which can overcome some of the difficulties with manual labeling. However, this requires researchers to interpret the results of algorithms, potentially leading to bias. In some cases, researchers may use annotated data to evaluate unsupervised approaches [57].

Limitations of ML vs. other heuristic approaches. While ML can be used to automate some cybercrime community analysis tasks, the use of simple heuristics for classification can outperform the time taken by a complex ML model, provided that the heuristics are a suitable alternative. Heuristics can be chosen using domain knowledge from reading posts on forums, instead of using ML techniques [93]. Also, a hybrid approach could be used, such as using ML for classification of products and replies, with heuristics to build up supply chains [19].

Due to the use of jargon and frequent spelling errors, use of off-the-shelf language modeling tools is infeasible [26]. Creating new classification models for posts, however, requires a large amount of training data for a well-performing model. It might be needed to manually annotate the intent and function of text with a group of annotators for agreement [26]. Creating these gold-standard datasets can be resource intensive for both time and the number of domain-experts needed. However, these are important for validating and evaluating models correctly, and evaluation datasets and metrics need to ensure they can be used to check the ML models are predicting what it is *expected* to predict on.

Models trained on one dataset may not transfer well to other datasets, and training a model on a new dataset can be time consuming to both build the training set for and to train. Therefore, researchers may choose to repurpose a model trained on one dataset for use on a different cybercrime community. Durrett et al. [40] looked at the problem of domain adaptation within the scope of identifying products on four cybercrime communities, using both named entity recognition and slot-filling. They found models trained on some forums have better generalizability than models trained on different types of forums. This could also depend on which parts of the forum were annotated. They suggest improvements are needed, including for out-of-vocabulary words. Out-of-vocabulary words could be jargon specific to one forum or words not present in the training data.

4.1.1 Ethical Considerations. While the research community is yet to come to a consensus about the need for an ethical review of passive cybercrime community data analysis, some norms have emerged. As shown in Table 1, some research is considered by IRBs (or equivalent), while others consider such research to be exempt. A minority of articles discussed ethics, with 13.0% of articles receiving an approval, and 8.7% receiving an exemption. There are also discipline-specific guidelines. For computer science, the Menlo Report [68] is an important guide for ethical security research, while criminology has established guidelines developed by academic societies. In this section, we outline some of the common ethical issues considered in prior research. These primarily arise in relation to data collection, analysis, and reporting findings. Ethical considerations are particularly important for cybercrime community research, as there is usually no informed consent, coupled with the potential to create direct harm for research subjects, such as arrest and prosecution.

Data collection. There may be differing privacy and ethical considerations depending on how forum data are collected. Some issues that researchers may need to consider are making payments to obtain access, obtaining invitations to closed communities (sometimes by misrepresentation), and potentially breaking Terms of Service. Researchers will need to weigh the costs, such as payments fueling the criminal endeavor, against possible risks. There are also legal considerations. We note there is relevant case law from the US that has ruled that web scraping from public sites does not violate the Computer Fraud and Abuse Act. Researchers also need to comply with relevant privacy provisions if dealing with personal information that may be contained within forums and markets.

Leaked data may be particularly sensitive due to personal information (including victim data) being shared using private messages, and may also contain users' IP and e-mail addresses. Many forums have a mechanism for users to send private messages to each other, such as for replying to product advertisements. While a web scraper will not have access to private messages, they may be included in leaked forum datasets. This not only creates opportunities for interesting research (e.g., [90, 119]), but as the authors of these posts did not intend for them to be public, it introduces new ethical issues for researchers to consider [123].

When collecting data, researchers may wish to consider the resources being consumed during the process. Researchers can avoid the replication of data collection activities by using data repositories. Martin and Christin [83] suggest one way to compensate for consuming significant resources over the Tor network is by operating Tor nodes.

Some types of data may introduce additional risk to researchers, e.g., malware infection. Another concern is that some types of content raise legal issues, such as downloading child sexual abuse images. For these reasons, some researchers rely on text-data only, without downloading attachments [63]. Pastrana et al. [93] downloaded packs of nude images used for eWhoring. They outline how they worked with their REB and the INHOPE hotline operator in their jurisdiction to implement guidelines to detect, report, and delete child exploitation material. While they had not anticipated finding such material, due to the precautions taken they were able to ascertain that such images were being shared on the forums.

Data analysis. One common ethical issue is informed consent. Often, the data are publicly available, in that there are no restrictions in who can register and open an account. Researchers note it is not feasible to obtain informed consent from forum users for passive studies, where data are scraped or leaked. Prior researchers have pointed to the British Society of Criminology's Statement of Ethics [12, 15, 57, 59, 63, 92, 97, 127]. This justifies not obtaining informed consent if the data are collected from publicly available communities, and is used for research on collective behavior without aiming at identifying individuals [23]. The Menlo Report similarly contains provisions for

when obtaining informed consent is impractical [93]. In this situation, it is advised that researchers should seek a waiver of informed consent from their REB [68].

Actors in cybercrime communities tend to use pseudonyms, rather than identifying themselves using their real name. Pastrana et al. [93] note that usernames would be difficult to remove, as they are used in the text of posts, and doing so would not reduce any risk to forum users. In some situations, the data being analyzed may be particularly distressing. In these situations, the researchers may be able to take steps to reduce exposure to such content. For the work with nude images referred to above, Pastrana et al. [93] automated the analysis of images, to avoid manual analysis and review of pornographic content.

Reporting findings. Further precautions taken by researchers include not reporting findings that could potentially identify individuals (including not publishing usernames), and presenting results objectively. While some researchers decide to obfuscate the name of the forum they analyze, not all do. Vu et al. [127] justify this by pointing out forum characteristics can make obfuscation infeasible. Another justification on scientific principles is replicability.

4.2 Limitations and Weaknesses of Previous Research

Previously, we discussed inherent issues that researchers encounter in this field, which requires mitigation and a careful evaluation of assumptions. Next, we highlight limitations of prior work which need to be addressed by future research.

Lack of generalizability. Some of the prior work in cybercrime communities has focused on prediction and analysis. Techniques have been developed for these tasks on datasets of single communities, however, these often do not work on other datasets. This limitation can also apply to other techniques and methods in cybercrime community research, as communities are not homogeneous. Results which are of cybercrime forums found on the "surface" web can only reflect cybercrime communities with open membership, and generalizations cannot be made to closed communities.

It is important for researchers to rigorously validate the predictions and analysis made by models, to check they are predicting what researchers expect them to. This can be achieved by using ground truth annotation data, or by combining their dataset with other sources. Using a test/train split of the dataset (with "ground truth annotation data") could be used to check accuracy and F1 scores. Also, it is important to note that models trained on one forum may not generalize to others depending on the task. Annotations could be created for a new forum to perform a validation step for checking if a model can generalize.

This idea is also common in general cybercrime literature: measurements of attacks on the internet are likely to pick up those occurring en masse, rather than specific complex and targeted attacks, and findings have limitations when generalizing to other populations or scenarios [97]. Also, while cybercrime forums were used to discuss configuration files for the Zeus banking trojan, it is likely such discussions exclude criminals that have well developed skills and who operate good operational security. Instead, the data may be more reflective of recent entrants to the field who may be more willing to post questions and share their experiences (although this does not make them any less interesting to research) [61].

Removal of outliers and validation of measurements. During validation, and checking of results from analysis and measurements, it is important to consider the removal of outliers. Some measurement studies have focused on the proceeds from crime. For example, previous research used prices on marketplace adverts to estimate this and identify top earners [117]. However, this can have impacts on findings and conclusions, as measuring income based on advertised marketplace prices may not provide a useful result, only providing a rough estimation of proceeds, and may

be affected by outliers in the dataset. Good practices, such as identifying and manually checking outliers are important for these tasks if using non-robust metrics, e.g., mean. Otherwise, a single outlier can affect summary results, so care needs to be taken during measurements to provide useful and accurate results [48]. We found only a minority of articles in the review stated that they checked for outliers (15.9%). However, while other articles may have also taken this step, they did not include this in the article.

Such prices in advertisements may be higher than actual trade amounts, where members may negotiate, and may not actually be sold at the price, where members may increase an item's price beyond its value to keep the advert online when out of stock ("holding price"), to keep the adverts online [117]. Including these holding prices in summary statistics of trading activity would considerably change results. For many articles, distributions are not explored, and in some cases it is not specified that outliers were removed, in which case, we assume they were not. Pastrana et al. relied on screenshots of PayPal and Amazon dashboards posted by eWhoring scammers aimed as "proof of earnings" [93] to attempt to mitigate this issue, however, this information is not complete and indeed could be modified.

Trust mechanisms on marketplaces have typically relied on reviews of merchants ("feedback"). Feedback on adverts may not specify quantity purchased leading to the assumption of one item only per feedback, or where feedback may not have precise timestamps, researchers may need to assume the advertised price was the same at which the feedback was left [117]. Vu et al. [127] carry out measurements of a contract system, where only some contracts have public transaction information. To estimate trading activity, they assume that private transactions have at least the same value as public transactions. Cuevas et al. [34] explore this limitation further, by looking at the problem of carrying out measurements by proxy, where the ground truth dataset cannot be directly obtained. They build a model to measure the accuracy of measurements from scraped data, finding marketplace measurements provide a lower bound using this method, and recommend a high frequency of scraping to avoid missing data points.

Some marketplaces have introduced new mechanisms for trust, including using Bitcoin transactions on the Blockchain to log transactions that have occurred between members, or contract systems to show that transactions have taken place. However, care must be taken with these "proof" systems, as vendors may also trade with themselves under dummy accounts to gain reputation [24]. Vu et al. [127] manually check 163 high-value transactions, including looking up bitcoin addresses to match transactions with contract timestamps, for validating these. While there may be "ground truth" with transaction logs, or where there are only marketplace advert prices to measure, it is important for researchers to note marketplaces are an adversarial environment where reputation matters, and vendors are incentivized to over-report. Overblown claims are also seen across security research in general [10, 11].

With text-based forums, measurements may exclude "lurkers", members who only read posts, and may focus on members who create a high number of posts. However, these may be low-quality contributions, such as spam or plagiarized posts. Also, while it is straightforward to measure metadata of forum datasets, measurements of topics and other derived datasets require classification and topic models. These are not perfect, and may add bias to results. Researchers should state their assumptions, including if they aim for an upper or lower bound estimate for a measure of activity.

Focus on English language forums. There is a considerable language bias in cybercrime community research. Most works in the field has focused on English language forums, with 96.0% of the articles in the review focusing on this. There are considerably fewer studies of platforms using

other languages (such as Spanish, Russian, and Arabic), which could potentially skew our perception of cybercrime as a result of where we are looking.

Stale data. Stale data can be an issue in the field. It is time-consuming for researchers to collect datasets themselves, so they may wish to use shared datasets. However, these can quickly become old if scrapers are not maintained and datasets are not regularly updated.

Lack of ethical review. We find the majority of articles do not discuss the ethics of research on cybercrime communities. Of the articles we reviewed, 11 had approval from a REB, and 6 were exempt. While we are unable to tell if researchers have considered research ethics during their work, this review highlights that researchers often neglect to discuss ethical considerations in their publications.

5 RECOMMENDATIONS AND FUTURE WORK

There is a body of interesting, novel, and useful research on cybercrime communities. There are many useful insights such research can provide. Researchers have had to battle many methodological issues, which we have outlined in this article. From our findings of common issues and limitations in cybercrime community research, we recommend researchers to

(1) Note assumptions made by creating a structure with unstructured datasets of cybercrime communities;
(2) explore suitable approaches for using ground truth data, including use of domain experts in cybercrime research if available and external data sources, and state limitations;
(3) acknowledge that collected datasets are unlikely to be complete, due to the evolving nature of forums and marketplaces and may contain volatile data;
(4) note that datasets can become stale, and where they are unable to keep this updated, state the limitations of using this for analysis;
(5) be aware of underlying bias in datasets, such as a skewed perception of cybercrime by only analyzing English-language forums;
(6) consider limitations of the generalizability of models to other cybercrime communities, including where NLP and ML models may overfit to bias;
(7) take care to verify and validate results, including removing outliers from analysis and checking for anomalous results with measurements of cybercrime communities;
(8) consider and set out the ethical case across all cybercrime community research; and
(9) take care when reporting the methods used to enable other researchers to replicate research, which may also involve sharing both datasets, which may vary depending on when the data was collected, and code.

Further challenges that researchers are likely to face are explored below. These are likely to arise due to community fragmentation following disruption and displacement, further lack of structure with the move to "micro" communities, and adversarial attacks.

Displacement. Online forums and marketplaces have experienced displacement following law enforcement action. For example, Silk Road was once most widely used cryptomarket. A police investigation resulted in the arrest and ultimate prosecution of the operator. Soska and Christin [117] analyzed what happened after the takedown. Within a month, SilkRoad 2.0 was set up, operated by former administrators and vendors of the original Silk Road. Within a few months, numerous marketplaces followed the same model. They varied in levels of sophistication, durability, and specialization. Some marketplaces disappeared due to law enforcement action. Some disappeared voluntarily, including "exit scams", where they ran off with what had been sent to the site

administrators in escrow. They concluded that the Silk Road takedown resulted in significant evolution of the marketplace ecosystem, compared to when Silk Road was a monopoly.

A similar phenomenon occurred after the "stresser" subforum on Hackforums was removed. As a result, the community became increasingly decentralized, spread across multiple Discord servers and Telegram chats [32]. Some of these are also used by booter operators to provide support for their customers. The shift from a centralized community to many smaller communities has introduced further issues when trying to capture an overall view across the field, as researchers cannot be certain they have identified all popular communities. These "micro" sources require more effort to scrape, as researchers need to know which are of interest, check if they are still active, and look out for displacement of members to other servers and chat channels.

Further loss of structure. In addition, data from chat channels are less structured than those from forums and marketplaces. At most, they may be broken down into channels, similar to boards found on cybercrime forums. Within these, conversation threads are mixed together. Users may talk about different topics simultaneously, one user may switch to a different topic, and conversations may gradually switch topics, where on cybercrime forums, a new thread would have been created instead. Thread disambiguation is not a trivial task, and adds to the complexity of analyzing these data sources.

Attacks on NLP and ML models. In the future, researchers may also need to pay attention to a new class of poisoning attacks against NLP models. Such attacks may require researchers to sanitize data to minimize the likelihood that forum users can successfully use imperceptible text-encoding attacks to disrupt ML models used for analysis [21].

6 CONCLUSION

This article explored the goals and methodological approaches used in cybercrime community research. Prior research has been useful in addressing cybercrime problems, often with novel methodologies. We highlight both challenges inherent to research of these communities, and limitations and weaknesses of prior research. We proposed a list of recommendations, which includes researchers setting out and ethical case across all cybercrime community research, raising awareness of method limitations for predictive tasks, and consideration of steps needed to validate and verify results in studies. Future work may be affected by a changing landscape, including displacement of communities to smaller platforms, a further loss of structure to datasets, and new attacks on NLP and ML models in use.

REFERENCES

[1] A. Abbasi, W. Li, V. Benjamin, S. Hu, and H. Chen. 2014. Descriptive analytics: Examining expert hackers in web forums. In *Proceeding of the 2014 IEEE Joint Intelligence and Security Informatics Conference*. 56–63.

[2] Sadia Afroz, Vaibhav Garg, Damon McCoy, and Rachel Greenstadt. 2013. Honor among thieves: A common's analysis of cybercrime economies. In *Proceedings of the APWG eCrime Researchers Summit*.

[3] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. Doppelgänger Finder: Taking stylometry to the underground. In *Proceedings of the 2014 IEEE Symposium on Security and Privacy*. 212–226. DOI: https://doi.org/10.1109/SP.2014.21

[4] Ugur Akyazi, Michel van Eeten, and Carlos H Gañán. 2021. Measuring cybercrime as a service (caas) offerings in a cybercrime forum. In *Proceedings of the Workshop on the Economics of Information Security*.

[5] Luca Allodi. 2017. Economic factors of vulnerability trade and exploitation. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1483–1499.

[6] Mohammed Almukaynizi, Alexander Grimm, Eric Nunes, Jana Shakarian, and Paulo Shakarian. 2017. Predicting cyber threats through hacker social networks in darkweb and deepweb forums. In *Proceedings of the ACM International Conference of The Computational Social Science Society of the Americas*.

[7] Mohammed Almukaynizi, Ericsson Marin, Eric Nunes, Paulo Shakarian, Gerardo I. Simari, Dipsy Kapoor, and Timothy Siedlecki. 2018. DARKMENTION: A deployed system to predict enterprise-targeted external cyberattacks. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. 31–36.

[8] Mohammed Almukaynizi, Eric Nunes, Krishna Dharaiya, Manoj Senguttuvan, Jana Shakarian, and Paulo Shakarian. 2017. Proactive identification of exploits in the wild through vulnerability mentions online. In *Proceedings of the 2017 International Conference on Cyber Conflict*. 82–88. DOI: https://doi.org/10.1109/CYCONUS.2017.8167501

[9] Benjamin Ampel, Sagar Samtani, Hongyi Zhu, Steven Ullman, and Hsinchun Chen. 2020. Labeling hacker exploits for proactive cyber threat intelligence: A deep transfer learning approach. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. 1–6. DOI: https://doi.org/10.1109/ISI49825.2020.9280548

[10] Ross Anderson, Chris Barton, Rainer Boehme, Richard Clayton, Carlos Ganan, Tom Grasso, Michael Levi, Tyler Moore, and Marie Vasek. 2019. Measuring the changing cost of cybercrime. In *Proceedings of the 18th Annual Workshop on the Economics of Information Security*. DOI: https://doi.org/10.17863/CAM.41598

[11] Ross Anderson, Chris Barton, Rainer Böhme, Richard Clayton, Michel J. G. van Eeten, Michael Levi, Tyler Moore, and Stefan Savage. 2013. Measuring the cost of cybercrime. In *Proceedings of the Economics of Information Security and Privacy*. DOI: https://doi.org/10.1007/978-3-642-39498-0_12

[12] Gilberto Atondo Siu, Ben Collier, and Alice Hutchings. 2021. Follow the money: The relationship between currency exchange and illicit behaviour in an underground forum. In *Proceedings of the 2021 IEEE European Symposium on Security and Privacy Workshops*. 191–201. DOI: https://doi.org/10.1109/EuroSPW54576.2021.00027

[13] AZSecure-data.org. AZSecure Data Portal. (n.d.). Retrieved February 2022 from https://www.azsecure-data.org/dark-web-forums.html

[14] Maria Bada, Yi Ting Chua, Ben Collier, and Ildiko Pete. 2021. Exploring masculinities and perceptions of gender in online cybercrime subcultures. In *Proceedings of the Cybercrime in Context*. Springer, 237–257.

[15] Maria Bada and Ildiko Pete. 2020. An exploration of the cybercrime ecosystem around shodan. In *Proceedings of the International Conference on Internet of Things: Systems, Management and Security*.

[16] Victor Benjamin and Hsinchun Chen. 2012. Securing cyberspace: Identifying key actors in hacker communities. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. 24–29. DOI: https://doi.org/10.1109/ISI.2012.6283296

[17] Victor Benjamin, Weifeng Li, Thomas Holt, and Hsinchun Chen. 2015. Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. 85–90. DOI: https://doi.org/10.1109/ISI.2015.7165944

[18] Victor Benjamin, Joseph S. Valacich, and Hsinchun Chen. 2019. DICE-E: A framework for conducting darknet identification, collection, evaluation with ethics. *MIS Quarterly: Management Information Systems* 43, 1 (2019), 1–22. DOI: https://doi.org/10.25300/MISQ/2019/13808

[19] Rasika Bhalerao, Maxwell Aliapoulios, Ilia Shumailov, Sadia Afroz, and Damon McCoy. 2019. Mapping the underground: Supervised discovery of cybercrime supply chains. In *Proceedings of the APWG Symposium on Electronic Crime Research (eCrime)*. 1–16. DOI: https://doi.org/10.1109/eCrime47957.2019.9037582

[20] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[21] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad characters: Imperceptible NLP attacks. In *Proceedings of the 2022 IEEE Symposium on Security and Privacy*. 1987–2004. DOI: https://doi.org/10.1109/SP46214.2022.9833641

[22] Gwern Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Delyan Kratunov Sohhlz, Vince Cakic, Van Buskirk, Whom, Michael McKenna, and Sigi Goode. 2015. Dark Net Market archives, 2011-2015. (2015). https://www.gwern.net/DNM-archives

[23] British Society of Criminology. 2015. Statement of ethics. (2015). Retrieved February 2022 from http://www.britsoccrim.org/ethics/

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

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

[26] Andrew Caines, Sergio Pastrana, Alice Hutchings, and Paula J Buttery. 2018. Automatically identifying the function and intent of posts in underground forums. *Crime Science* 7, 1 (2018), 1–14.

[27] Michele Campobasso and Luca Allodi. 2022. THREAT/crawl: A trainable, highly-reusable, and extensible automated method and tool to crawl criminal underground forums. In *Proceedings of the 17th APWG Symposium on Electronic Crime Research*.

[28] Michele Campobasso, Pavlo Burda, and Luca Allodi. 2019. Caronte: Crawling adversarial resources over non-trusted, high-profile environments. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshops*.

[29] Hsinchun Chen. 2012. Dark web: Exploring and mining the dark side of the web. In *Proceedings of the Formal Concept Analysis*. Florent Domenach, Dmitry I. Ignatov, and Jonas Poelmans (Eds.), Springer, Berlin, 1–1.

[30] Nicolas Christin. 2013. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd International World Wide Web Conference*. 213–224.

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

[32] Ben Collier, Daniel R. Thomas, Richard Clayton, Alice Hutchings, and Yi Ting Chua. 2022. Influence, infrastructure, and recentering cybercrime policing: Evaluating emerging approaches to online law enforcement through a market for cybercrime services. *Policing and Society* 32, 1 (2022), 1–22.

[33] Derek B. Cornish. 1994. The procedural analysis of offending and its relevance for situational prevention. *Crime Prevention Studies* 3, 1 (1994), 151–196.

[34] Alejandro Cuevas, Fieke Miedema, Nicolas Christin, Kyle Soska, and Rolf van Wegberg. 2022. Measurement by proxy: On the accuracy of online marketplace measurements. In *Proceedings of the 31st USENIX Security Symposium*.

[35] Ashok Deb, Kristina Lerman, and Emilio Ferrara. 2018. Predicting cyber-events by leveraging hacker sentiment. *Information* 9, 11 (2018), 2078–2489.

[36] Isuf Deliu, Carl Leichter, and Katrin Franke. 2017. Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks. In *Proceedings of the IEEE International Conference on Big Data*. 3648–3656. DOI: https://doi.org/10.1109/BigData.2017.8258359

[37] Isuf Deliu, Carl Leichter, and Katrin Franke. 2018. Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent dirichlet allocation. In *Proceedings of the IEEE International Conference on Big Data*. 5008–5013. DOI: https://doi.org/10.1109/BigData.2018.8622469

[38] Po-Yi Du, Ning Zhang, Mohammedreza Ebrahimi, Sagar Samtani, Ben Lazarine, Nolan Arnold, Rachael Dunn, Sandeep Suntwal, Guadalupe Angeles, Robert Schweitzer, and others. 2018. Identifying, collecting, and presenting hacker community data: Forums, IRC, carding shops, and DNMs. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. 70–75.

[39] Benoît Dupont, Anne-Marie Côté, Claire Savine, and David Décary-Hétu. 2016. The ecology of trust among hackers. *Global Crime* 17, 2 (2016), 129–151.

[40] Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Rebecca Portnoff, Sadia Afroz, Damon McCoy, Kirill Levchenko, and Vern Paxson. 2017. Identifying products in online cybercrime marketplaces: A dataset for fine-grained domain adaptation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, 2598–2607.

[41] Mohammadreza Ebrahimi, Yidong Chai, Sagar Samtani, and Hsinchun Chen. 2022. Cross-lingual cybersecurity analytics in the international dark web with adversarial deep representation learning. *MIS Quarterly* 46, 2 (2022), 1209–1226.

[42] Mohammadreza Ebrahimi, Sagar Samtani, Yidong Chai, and Hsinchun Chen. 2020. Detecting cyber threats in non-english hacker forums: An adversarial cross-lingual knowledge transfer approach. In *Proceedings of the 2020 IEEE Security and Privacy Workshops*. 20–26. DOI: https://doi.org/10.1109/SPW50608.2020.00021

[43] Yujie Fan, Yanfang Ye, Qian Peng, Jianfei Zhang, Yiming Zhang, Xusheng Xiao, Chuan Shi, Qi Xiong, Fudong Shao, and Liang Zhao. 2020. Metagraph aggregated heterogeneous graph neural network for illicit traded product identification in underground market. In *Proceedings of the IEEE International Conference on Data Mining*.

[44] Yong Fang, Yusong Guo, Cheng Huang, and Liang Liu. 2019. Analyzing and identifying data breaches in underground forums. *IEEE Access* 7 (2019), 48770–48777. DOI: https://doi.org/10.1109/ACCESS.2019.2910229

[45] Z. Fang, X. Zhao, Q. Wei, G. Chen, Y. Zhang, C. Xing, W. Li, and H. Chen. 2016. Exploring key hackers and cybersecurity threats in chinese hacker communities. In *Proceeding of the IEEE ISI*.

[46] Shehroze Farooqi, Guillaume Jourjon, Muhammad Ikram, Mohamed Ali Kaafar, Emiliano De Cristofaro, Zubair Shafiq, Arik Friedman, and Fareed Zaffar. 2017. Characterizing key stakeholders in an online black-hat marketplace. In *Proceedings of the APWG Symposium on Electronic Crime Research*. 17–27. DOI: https://doi.org/10.1109/ECRIME.2017.7945050

[47] Nathan Fisk. 2006. *Social Learning Theory as a Model for Illegitimate Peer-to-peer use and the Effects of Implementing a Legal Music Downloading Service on Peer-to-peer Music Piracy*. Ph.D. Dissertation. Rochester Institute of Technology.

[48] Dinei Florêncio and Cormac Herley. 2013. Sex, lies and cyber-crime surveys. In *Proceedings of the Economics of Information Security and Privacy III*. Bruce Schneier (Ed.), Springer New York, New York, NY, 35–53.

[49] Richard Frank, Myfanwy Thomson, Alexander Mikhaylov, and Andrew J. Park. 2018. Putting all eggs in a single basket: A cross-community analysis of 12 hacking forums. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. 136–141.

[50] Jason Franklin, Adrian Perrig, Vern Paxson, and Stefan Savage. 2007. An Inquiry into the nature and causes of the wealth of internet miscreants. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*. Association for Computing Machinery, New York, NY, 375–388. DOI: https://doi.org/10.1145/1315245.1315292
[51] Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen. 2010. A focused crawler for dark web forums. *Journal of the American Society for Information Science and Technology* 61, 6 (2010), 1213–1231.
[52] Vaibhav Garg, Sadia Afroz, Rebekah Overdorf, and Rachel Greenstadt. 2015. Computer-supported cooperative crime. In *Proceedings of the Financial Cryptography and Data Security*.
[53] John Grisham, Sagar Samtani, Mark Patton, and Hsinchun Chen. 2017. Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. DOI: https://doi.org/10.1109/ISI.2017.8004867
[54] Thomas J. Holt, Olga Smirnova, and Alice Hutchings. 2016. Examining signals of trust in criminal markets online. *Journal of Cybersecurity* 2, 2 (2016), 137–145.
[55] Cheng Huang, Yongyan Guo, Wenbo Guo, and Ying Li. 2021. HackerRank: Identifying key hackers in underground forums. *International Journal of Distributed Sensor Networks* 17, 5 (2021), 15501477211015145.
[56] Shin-Ying Huang and Hsinchun Chen. 2016. Exploring the online underground marketplaces through topic-based social network and clustering. In *Proceedings of the IEEE Conference on Intelligence and Security Informatics*. 145–150.
[57] Jack Hughes, Seth Aycock, Andrew Caines, Paula Buttery, and Alice Hutchings. 2020. Detecting trending terms in cybersecurity forum discussions. In *Proceedings of the Workshop on Noisy User-generated Text*. DOI: https://doi.org/10.18653/v1/2020.wnut-1.15
[58] Jack Hughes, Yi Ting Chua, and Alice Hutchings. 2021. Too much data? Opportunities and challenges of large datasets and cybercrime. *Researching Cybercrimes* (2021), 191–212.
[59] Jack Hughes, Ben Collier, and Alice Hutchings. 2019. From playing games to committing crimes: A multi-technique approach to predicting key actors on an online gaming forum. In *Proceedings of the APWG Symposium on Electronic Crime Research*.
[60] Alice Hutchings. 2018. Leaving on a jet plane: The trade in fraudulently obtained airline tickets. *Crime, Law and Social Change* 70, 4 (2018), 461–487.
[61] Alice Hutchings and Richard Clayton. 2017. Configuring zeus: A case study of online crime target selection and knowledge transmission. In *Proceedings of the IEEE APWG Symposium on Electronic Crime Research*. 33–40.
[62] Alice Hutchings and Thomas J Holt. 2015. A crime script analysis of the online stolen data market. *British Journal of Criminology* 55, 3 (2015), 596–614.
[63] Alice Hutchings and Sergio Pastrana. 2019. Understanding eWhoring. In *Proceedings of the IEEE European Symposium on Security and Privacy*. 201–214.
[64] Christos Iliou, George Kalpakis, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2016. Hybrid focused crawling for homemade explosives discovery on surface and dark web. In *Proceedings of the IEEE 11th International Conference on Availability, Reliability and Security*. 229–234.
[65] Jan William Johnsen and Katrin Franke. 2018. Identifying central individuals in organised criminal groups and underground marketplaces. In *Proceedings of the International Conference on Computational Science*.
[66] Jan William Johnsen and Katrin Franke. 2020. Identifying proficient cybercriminals through text and network analysis. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*.
[67] Panicos Karkallis, Jorge Blasco, Guillermo Suarez-Tangil, and Sergio Pastrana. 2021. Detecting video-game injectors exchanged in game cheating communities. In *Proceedings of the European Symposium on Research in Computer Security*.
[68] Erin Kenneally and David Dittrich. 2012. The menlo report: Ethical principles guiding information and communication technology research. *Tech. Report., U.S. Department of Homeland Security*. (2012).
[69] Brian Krebs. 2017. Who is Anna-Senpai, the Mirai worm author? (January 2017). Retrieved August 2022 from https://krebsonsecurity.com/2017/01/who-is-anna-senpai-the-mirai-worm-author/
[70] Brian Krebs. 2017. Who is Marcus Hutchins? (September 2017). Retrieved August 2022 from https://krebsonsecurity.com/2017/09/who-is-marcus-hutchins/
[71] M. Kiran Kumar and Dr K. Bhargavi. 2020. An effective study on data science approach to cybercrime underground economy data. *Journal of Engineering, Computing and Architecture* 10, 1 (2020), 148–158.
[72] Weifeng Li, Hsinchun Chen, and Jay F. Nunamaker Jr. 2016. Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. *Journal of Management Information Systems* 33, 4 (2016), 1059–1086.
[73] Ying Li, Jiaxing Cheng, Cheng Huang, Zhouguo Chen, and Weina Niu. 2021. NEDetector: Automatically extracting cybersecurity neologisms from hacker forums. *Journal of Information Security and Applications* 58 (2021), 102784.
[74] Mitch Macdonald and Richard Frank. 2017. The network structure of malware development, deployment and distribution. *Global Crime* 18, 1 (2017), 49–69.

[75] Mitch Macdonald, Richard Frank, Joseph Mei, and Bryan Monk. 2015. Identifying digital threats in a hacker web forum. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 926–933.

[76] Ericsson Marin, Mohammed Almukaynizi, Eric Nunes, Jana Shakarian, and Paulo Shakarian. 2018. Predicting hacker adoption on darkweb forums using sequential rule mining. In *Proceedings of the IEEE International Conference on Parallel Distributed Processing with Applications, Ubiquitous Computing Communications, Big Data Cloud Computing, Social Computing Networking, Sustainable Computing Communications*.

[77] E. Marin, M. Almukaynizi, E. Nunes, and P. Shakarian. 2018. Community finding of malware and exploit vendors on darkweb marketplaces. In *Proceedings of the International Conference on Data Intelligence and Security*.

[78] Ericsson Marin, Mohammed Almukaynizi, Soumajyoti Sarkar, Eric Nunes, Jana Shakarian, Paulo Shakarian, and Edward G. Amoroso. 2021. *Exploring Malicious Hacker Communities: Toward Proactive Cyber-Defense*. Cambridge University Press. DOI: https://doi.org/10.1017/9781108869003

[79] Ericsson Marin, Mohammed Almukaynizi, and Paulo Shakarian. 2019. Reasoning about future cyber-attacks through socio-technical hacking information. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence*. DOI: https://doi.org/10.1109/ICTAI.2019.00030

[80] Ericsson Marin, Mohammed Almukaynizi, and Paulo Shakarian. 2020. Inductive and deductive reasoning to assist in cyber-attack prediction. In *Proceedings of the 10th Annual Computing and Communication Workshop and Conference*. DOI: https://doi.org/10.1109/CCWC47524.2020.9031154

[81] E. Marin, A. Diab, and P. Shakarian. 2016. Product offerings in malicious hacker markets. In *Proceedings of the IEEE Conference on Intelligence and Security Informatics*. 187–189.

[82] E. Marin, J. Shakarian, and P. Shakarian. 2018. Mining hey-hackers on darkweb forums. In *Proceedings of the International Conference on Data Intelligence and Security*. 73–80. DOI: https://doi.org/10.1109/ICDIS.2018.00018

[83] James Martin and Nicolas Christin. 2016. Ethics in cryptomarket research. *International Journal of Drug Policy* 35 (2016), 84–91.

[84] Alexander Mikhaylov and Richard Frank. 2016. Cards, money and two hacking forums: An analysis of online money laundering schemes. In *Proceedings of the European Intelligence and Security Informatics Conference*.

[85] Robert G. Morris and Ashley G. Blackburn. 2009. Cracking the code: An empirical exploration of social learning theory and computer crime. *Journal of Crime and Justice* 32, 1 (2009), 1–34.

[86] Marti Motoyama, Kirill Levchenko, Chris Kanich, Damon McCoy, Geoffrey M. Voelker, and Stefan Savage. 2010. Re:CAPTCHAs-understanding CAPTCHA-solving services in an economic context. In *Proceedings of the USENIX Security Symposium*. 3.

[87] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M. Voelker. 2011. An analysis of underground forums. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*. 71–80.

[88] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart, and P. Shakarian. 2016. Darknet and deepnet mining for proactive cybersecurity threat intelligence. In *Proceeding of the ISI*. IEEE, 7–12.

[89] Meltem Odabaş, Thomas J. Holt, and Ronald L. Breiger. 2017. Markets as governance environments for organizations at the edge of illegality: Insights from social network analysis. *American Behavioral Scientist* 61, 11 (2017), 1267–1288.

[90] Rebekah Overdorf, Carmela Troncoso, Rachel Greenstadt, and Damon McCoy. 2018. Under the underground: Predicting private interactions in underground forums. arXiv:1805.04494. Retrieved from https://arxiv.org/abs/1805.04494

[91] Andrew J. Park, Richard Frank, Alexander Mikhaylov, and Myf Thomson. 2018. Hackers hedging bets: A cross-community analysis of three online hacking forums. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

[92] Sergio Pastrana, Alice Hutchings, Andrew Caines, and Paula Buttery. 2018. Characterizing eve: Analysing cybercrime actors in a large underground forum. In *Proceedings of the Research in Attacks, Intrusions, and Defenses*.

[93] Sergio Pastrana, Alice Hutchings, Daniel Thomas, and Juan Tapiador. 2019. Measuring ewhoring. In *Proceedings of the Internet Measurement Conference*.

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

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

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

[97] Ildiko Pete, Jack Hughes, Yi Ting Chua, and Maria Bada. 2020. A social network analysis and comparison of six dark web forums. In *Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops*. IEEE, 484–493.

[98] Elizabeth Phillips, Jason R. C. Nurse, Michael Goldsmith, and Sadie Creese. 2015. Extracting social structure from darkWeb forums. In *5th International Conference on Social Media Technologies, Communication, and Informatics (SOTICS'15)*.

[99] Alexander Pons and Eugene Pons. 2015. Social learning theory and ethical hacking: Student perspectives on a hacking curriculum. In *Proceedings of the Information Systems Education Conference*. Foundation for IT Education, New York, NY, 289–299.

[100] Rebecca S. Portnoff, Sadia Afroz, Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Damon McCoy, Kirill Levchenko, and Vern Paxson. 2017. Tools for automated analysis of cybercriminal markets. In *Proceedings of the ACM International Conference on World Wide Web*.

[101] Andrei Lima Queiroz, Brian Keegan, and Susan Mckeever. 2020. Moving targets: Addressing concept drift in supervised models for hacker communication detection. In *Proceedings of the International Conference on Cyber Security and Protection of Digital Services*.

[102] Andrei Lima Queiroz, Susan Mckeever, and Brian Keegan. 2019. Detecting hacker threats: Performance of word and sentence embedding models in identifying hacker communications. In *Proceedings of the International Conference on Artificial Intelligence and Computer Science*.

[103] J. Radianti. 2010. A study of a social behavior inside the online black markets. In *Proceedings of the 4th International Conference on Emerging Security Information, Systems and Technologies*. 189–194.

[104] J. Robertson, A. Diab, E. Marin, E. Nunes, V. Paliath, J. Shakarian, and P. Shakarian. 2017. *Darkweb Cyber Threat Intelligence Mining*. Cambridge University Press.

[105] Sagar Samtani, Yidong Chai, and Hsinchun Chen. 2022. Linking exploits from the dark web to known vulnerabilities for proactive cyber threat intelligence: An attention-based deep structured semantic model. *MIS Quarterly* 46, 2 (2022), 911–946.

[106] S. Samtani and H. Chen. 2016. Using social network analysis to identify key hackers for keylogging tools in hacker forums. In *Proceeding of the ISI*.

[107] Sagar Samtani, Ryan Chinn, and Hsinchun Chen. 2015. Exploring hacker assets in underground forums. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. 31–36. DOI: https://doi.org/10.1109/ISI.2015.7165935

[108] Sagar Samtani, Ryan Chinn, Hsinchun Chen, and Jay F. Nunamaker Jr. 2017. Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. *Journal of Management Information Systems* 34, 4 (2017), 1023–1053. DOI: https://doi.org/10.1080/07421222.2017.1394049

[109] Sagar Samtani, Hongyi Zhu, and Hsinchun Chen. 2020. Proactively identifying emerging hacker threats from the dark web: A diachronic graph embedding framework (D-GEF). *ACM Transactions on Privacy and Security* 23, 4 (2020), 1–33.

[110] Anna Sapienza, Alessandro Bessi, Saranya Damodaran, Paulo Shakarian, Kristina Lerman, and Emilio Ferrara. 2017. Early warnings of cyber threats in online discussions. In *Proceedings of the IEEE International Conference on Data Mining Workshops*.

[111] Anna Sapienza, Sindhu Kiranmai Ernala, Alessandro Bessi, Kristina Lerman, and Emilio Ferrara. 2018. DISCOVER: Mining online chatter for emerging cyber threats. In *Companion Proceedings of the The Web Conference*.

[112] Soumajyoti Sarkar, Mohammad Almukaynizi, Jana Shakarian, and Paulo Shakarian. 2019. Mining user interaction patterns in the darkweb to predict enterprise cyber incidents. *Social Network Analysis and Mining* 9, 1 (2019). DOI: https://doi.org/10.1007/s13278-019-0603-9

[113] Soumajyoti Sarkar, Mohammad Almukaynizi, Jana Shakarian, and Paulo Shakarian. 2019. Predicting enterprise cyber incidents using social network analysis on dark web hacker forums. *The Cyber Defense Review* (2019), 87–102.

[114] Dominic Seyler, Wei Liu, XiaoFeng Wang, and ChengXiang Zhai. 2021. Towards dark jargon interpretation in underground forums. In *Proceedings of the European Conference on Information Retrieval*.

[115] Dominic Seyler, Wei Liu, Yunan Zhang, XiaoFeng Wang, and ChengXiang Zhai. 2021. DarkJargon.net: A platform for understanding underground conversation with latent meaning. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

[116] William F. Skinner and Anne M. Fream. 1997. A social learning theory analysis of computer crime among college students. *Journal of Research in Crime and Delinquency* 34, 4 (1997), 495–518.

[117] Kyle Soska and Nicolas Christin. 2015. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *Proceedings of the USENIX Security Symposium*.

[118] Zhibo Sun, Adam Oest, Penghui Zhang, Carlos Rubio-Medrano, Tiffany Bao, Ruoyu Wang, Ziming Zhao, Yan Shoshitaishvili, Adam Doupé, Gail-Joon Ahn, and others. 2021. Having your cake and eating it: An analysis of concession-abuse-as-a-service. In *Proceedings of the 30th USENIX Security Symposium*.

[119] Zhibo Sun, Carlos E. Rubio-Medrano, Ziming Zhao, Tiffany Bao, Adam Doupé, and Gail-Joon Ahn. 2019. Understanding and predicting private interactions in underground forums. In *Proceedings of the ACM Conference on Data and Application Security and Privacy*. 303–314.

[120] Srikanth Sundaresan, Damon McCoy, Sadia Afroz, and Vern Paxson. 2016. Profiling underground merchants based on network behavior. In *Proceedings of the IEEE APWG Symposium on Electronic Crime Research*. 1–9.

[121] Jakapun Tachaiya, Joobin Gharibshah, Evangelos E. Papalexakis, and Michalis Faloutsos. 2020. RThread: A thread-centric analysis of security forums. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

[122] Michal Tereszkowski-Kaminski, Sergio Pastrana, Jorge Blasco, and Guillermo Suarez-Tangil. 2022. Towards improving code stylometry analysis in underground forums. *Proceedings on Privacy Enhancing Technologies* 2022, 1 (2022), 126–147.

[123] Daniel R. Thomas, Sergio Pastrana, Alice Hutchings, Richard Clayton, and Alastair R. Beresford. 2017. Ethical issues in research using datasets of illicit origin. In *Proceedings of the Internet Measurement Conference*. 445–462.

[124] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. SoK: Hate, harassment, and the changing landscape of online abuse. In *Proceedings of the IEEE Symposium on Security and Privacy*. 247–267.

[125] Kieron Turk, Sergio Pastrana, and Ben Collier. 2020. 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*. IEEE, 428–437.

[126] Gert Jan Van Hardeveld, Craig Webber, and Kieron O'Hara. 2016. Discovering credit card fraud methods in online tutorials. In *Proceedings of the International Workshop on Online Safety, Trust and Fraud Prevention*.

[127] Anh V Vu, Jack Hughes, Ildiko Pete, Ben Collier, Yi Ting Chua, Ilia Shumailov, and Alice Hutchings. 2020. 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*.

[128] Ryan Williams, Sagar Samtani, Mark Patton, and Hsinchun Chen. 2018. Incremental hacker forum exploit collection and classification for proactive cyber threat intelligence: An exploratory study. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. DOI: https://doi.org/10.1109/ISI.2018.8587336

[129] Kan Yuan, Haoran Lu, Xiaojing Liao, and XiaoFeng Wang. 2018. Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces. In *Proceedings of the USENIX Security Symposium*.

[130] Wei T Yue, Qiu-Hong Wang, and Kai-Lung Hui. 2019. See no evil, hear no evil? Dissecting the impact of online hacker forums. *Mis Quarterly* 43, 1 (2019), 73.

[131] Xiong Zhang and Chenwei Li. 2013. Survival analysis on hacker forums. In *Proceedings of the SIGBPS Workshop on Business Processes and Service*. 106–110.

[132] X. Zhang, A. Tsang, W. Yue, and M. Chau. 2015. The classification of hackers by knowledge exchange behaviors. *Inform. Systems Frontiers* 17, 6 (2015), 1239–1251.

[133] Yiming Zhang, Yujie Fan, Shifu Hou, Jian Liu, Yanfang Ye, and Thirimachos Bourlai. 2018. iDetector: Automate underground forum analysis based on heterogeneous information network. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

[134] Yiming Zhang, Yujie Fan, Yanfang Ye, Liang Zhao, and Chuan Shi. 2019. Key player identification in underground forums over attributed heterogeneous information network embedding framework. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

[135] Yiming Zhang, Yujie Fan, Yanfang Ye, Liang Zhao, Jiabin Wang, Qi Xiong, and Fudong Shao. 2018. KADetector: Automatic identification of key actors in online hack forums based on structured heterogeneous information network. In *Proceedings of the IEEE International Conference on Big Knowledge*.

Received 3 April 2023; accepted 15 December 2023