Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Crime VIP: A Closed-Access Underground Hacking Forum

Mariella Mischinger[1,2], **Sergio Pastrana**[2], **Guillermo Suarez-Tangil**[1],
[1]IMDEA Networks Institute
Madrid, Spain
[2] Universidad Carlos III de Madrid
Leganés, Spain

Abstract

We present XIN, the dataset of a closed-access Russian-English underground hacking forum. It contains a collection of ≈1.3M posts from over 20 years (Feb 2005-Aug 2025), and — to the best of our knowledge — is the largest collection of a closed-access underground hacking forum available for research. While there is a wide range of underground forum datasets available, there is a lack of non-English forums, and especially closed-access forums. Those are particularly challenging to crawl as the access is gated. Hence, a limited few-shot opportunity requires extreme care when crawling to avoid detection, which leads to account banning. Our stealthy data collection spanned 5 years (2020-2025).

The statistical analysis of our data mirrors how cybercrime gradually shifted from technical, hands-on hacking to an industry where different building blocks can be assembled and applied in the absence of a broad or profound technical understanding. This dataset will contribute to a more complete evaluation of the cybercriminal landscape, shedding light on the activity that happens in closed-door, non-English communities.

Dataset —
https://www.cambridgecybercrime.uk/datasets.html

Introduction

Underground forums are a central infrastructure of the cybercriminal ecosystem and a substantial driver of severe cyberattacks that threaten society (Stone-Gross et al. 2011). On those platforms, participants exchange (technical) knowledge, provide advice, and trade a wide range of crime-related services, including spam or pay-per-install, as well as hacking-related tools and even ready-to-use malware like Trojans and Keyloggers (Li and Liao 2024; Mischinger, Pastrana, and Suarez-Tangil 2024). The product trading resembles that of the clear-web: vendors give away free samples to trusted members, who in return provide a detailed review that influences the purchase decision of other users. As such, discussions in underground forums often relate to early stages of cyberattacks, when programs are being developed and tested, or products are first offered for sale. Notably, recent research shows that underground forums frequently share knowledge prior to its appearance in cyber threat intelligence (CTI) reports (Paladini et al. 2024; Mischinger, Pastrana, and Suarez-Tangil 2024). Hence, underground forums are a valuable source for anticipating emerging threats and underscore their importance for early detection and prevention of cyberattacks. For instance, their analysis can reveal criminals' tactics (Siu, Collier, and Hutchings 2021; de la Cruz and Pastrana 2024), the risks they pose for society (Caballero et al. 2011), and actionable CTI for active threat prevention (Gharibshah, Papalexakis, and Faloutsos 2018; Mischinger, Pastrana, and Suarez-Tangil 2024).

At a broader level, these forums are structured around large underground hubs that connect a wide range of hacking communities. One natural demarcation between these communities is the language spoken by their members. Consequently, the forums catering to them are available in multiple languages. Another clear way to distinguish these communities is by their level of exposure, spanning from widely visible open forums to secretive ones. On the one hand, open forums allow users to browse freely without an account or to create one for free without any form of vetting. On the other hand, closed-access forums restrict entry, granting access only through invitations or the payment of an entrance fee. This access hurdle results in a more selective audience, as either a prior connection to the cybercriminal environment is required or the willingness to pay the entrance fee demonstrates real commitment, thereby filtering out those with casual curiosity. While large open-access communities often observe many entry-level cybercriminals, serious actors gather in more exclusive and even closed-access forums (Dupont et al. 2017).

Unfortunately, prior work on underground forum analysis is limited to open-access forums (Pastrana et al. 2018b), which English-speaking actors predominantly use. This oversight in the study of multilingual closed-access forums introduces bias into the findings, as it does not capture the full cybercriminal landscape. One reason could be the greater difficulty of accessing these forums, combined with the limited number of accounts the crawler can use to access the various sites. Consequently, the few-shot opportunity makes crawling these forums particularly difficult, as crawling-related behavior, such as increased website requests, may result in the account being banned. To contribute a step in that direction, we release the dataset of

a multilingual closed-access Russian-English underground hacking forum, called XIN. We have been crawling the forum from 2020 to 2025, and the dataset spans more than 20 years of forum data with $\approx$ 1.3M posts.

Background

Online underground forums serve as platforms where cybercriminals exchange knowledge, illicit goods, and services. The following section provides an overview of how such exchanges have contributed to the broader commoditization of cybercrime, thereby underscoring the importance of collecting data from these forums and analyzing their posts.

Commoditization of Cybercrime. While modern cyberattacks may still rely on individual expertise, cybercrime has become an industry in which the different building blocks of a cyberattack are offered by different actors (Stone-Gross et al. 2011; Caballero et al. 2011). Consequently, to conduct a successful cyberattack, a cybercriminal does not require technical expertise. The components required for a cyberattack can be obtained within the underground ecosystem, either by assembling services and ready-to-use tools or by engaging with affiliate programs and partners. As a result, although the technical learning curve is gradually decreasing, we see an increase in cybercrime activity and its impact.

Underground Forums. Forums are usually organized in so-called "sub-forums" or "boards," a first division of discussions into high-level topics. Through their title, these provide a first high-level description of the topics being discussed, e.g., "System Access": where conversation evolves around how to break into systems. These forums provide not only knowledge, but also access to key actors who can transform that knowledge into actionable products and services (Pastrana et al. 2018a). The topics, products, and services described therein have been organized into a taxonomy constructed by the so-called *profit centers* and *support centers* (Huang et al. 2015). From a financial perspective, funds are channeled into the underground through *profit centers*, i.e., the actual scams targeting victims. Those profit centers rely on underlying *support centers* that provide the infrastructure behind those scams, such as hosting services or malware distribution services, resulting in a (financial) dependency among the different services. This division of roles is reflected in the wide scope of topics discussed in underground forums and the diversity of their boards, which mirror the cybercrime supply chain (Pastrana et al. 2018b): Those boards cover a range of topics, including hardware and software, programming, leaks, money-related aspects, legal advice, and purchasing products and services. In addition to these, other non-hacking-related boards, such as forum organization or entertainment, are also present. Some forums are created for a very specific subject, e.g., forums dedicated to spam-related activity of a spam affiliate program (Stone-Gross et al. 2011), or forums for the exchange of stolen data (Marjanov and Hutchings 2025).

Closed-Access Forums. Many forums are open access, meaning that anyone can create an account and participate in the discussions. Some forums are restricted, with an administrator vetting new users. This vetting is typically enforced either through invitation-only access or by requiring the payment of a substantial fee. This has implications for the community. First, naturally, this leads to a more selective audience. The invitation requires prior access to the hacking environment; therefore, new members are likely already familiar with crime-related activities and are expected to possess a higher skill set. Members who gain entry by paying a fee often have a strong incentive to engage in cybercrime in order to recoup their initial investment. Either way, due to the entrance barriers, members might feel safer in sharing crime-related content. Furthermore, it may affect members' behavior (e.g., by prompting them to avoid disputes) because a possible account ban carries greater weight. Due to these attributes, closed-access forums are worth studying, besides the increased difficulty in crawling.

Languages. These forums are available in multiple languages, including English, Russian, and Spanish, and support multilingual content. The language barrier shapes who can participate in the community. While English forums address a more global audience, non-English forums foster regionally focused communities in which participants share a cultural background and community-specific attack vectors.

Crawling Methodology and Data Collection

In this section, we provide an overview of the data collection process and the data format. First, we discuss the challenges of crawling a closed-access forum and the implications for crawling methods and duration. Furthermore, we describe how we handle the data present in the forum. Underground forum posts may contain non-textual content such as images, video frames, or attachments. As we only download and store text, we explain how we contain other content present in a post. Finally, we present the data structure for organizing and storing forum content.

Crawling Challenges. The forum we crawl is closed-access, which presents particular challenges. First, we have access to only one account, provided by an industry partner who wishes to remain anonymous. As such, the crawling process cannot be parallelized like reference crawlers (Pastrana et al. 2018b). Second, in contrast to crawling other open-registration underground forums, stealth is crucial, as detection could result in our account being banned.

Crawling Methodology. Accordingly, to minimize the risk of detection, our crawling mimics human navigation in terms of connection patterns and fetch rates, an approach previously described by CARONTE (Campobasso, Burda, and Allodi 2019). While CARONTE focuses only on the 'interesting' areas of the forum, we encounter greater difficulties in crawling the forum in its entirety. For example, accessing historical data at a rapid pace (e.g., sequentially viewing posts written a decade ago) might raise suspicion if this activity is being monitored. These challenges increased the time and effort required to crawl the entire forum (Turk, Pastrana, and Collier 2020), motivating us to make the data available to other researchers, thereby reducing their scientific workload.

Crawl Duration. We crawled the forum for more than five

years, with the first request made on July 20, 2020, and the last on August 27, 2025. The prolonged crawling time necessitated increased maintenance due to expiring cookies or website changes. Furthermore, crawling the entire forum required multiple iterations, as by the time the first iteration finished, a substantial number of new posts had appeared. We later provide overall statistics of the dataset.

Content Type. Due to the nature of the forum we crawl, there is a risk of downloading illegal or malicious content. However, as our crawler gathers only raw text data from the forum, we do not expect any artifacts or apparent negative effects in these data. To display non-textual content, such as images or videos, we use delimiters (placed in front of keywords) to indicate their presence in the post. Furthermore, we indicate whether a post contains quotations or citations from a previous post, and we annotate the source code accordingly. Table 1 lists all crawler annotations. This would allow other researchers to download and conduct additional analyses with these artifacts (Karkallis et al. 2021; Tereszkowski-Kaminski et al. 2022), provided they have a proper legal and ethical framework for downloading these assets.

Annotation	Description
IMG	URL to an image.
CITING	IdPost of cited post.
CITING_MISSED_POST	Plain text of cited post in case IdPost could not be retrieved.
IFRAME	URL to an external iframe, usually used for embedding videos.
LINK	Other external URL.
CODE	Source code written explicitly. Note that in some cases, the source code is written as regular text so it won't be annotated.

Table 1: Crawler annotation types and their descriptions.

Data Structure. The data is available as a SQL database. The database has three tables (described in Table 2): *Forum* (Sub-forums), *Thread*, and *Post*.

- **Sub-forums.** The forum is divided into sub-forums (or boards), each with a headline and a unique identifier. Most headlines concern hacking, and the threads thematically align with the sub-forum's topic.
- **Thread.** Each sub-forum contains various threads, i.e., conversations. A thread comprises at least one post that discusses a particular topic. The discussion is initiated by the *Member*, who creates the thread and writes a headline and the first post that describe the subject. Subsequent posts are understood as replies to the previous content.
- **Post.** Posts contain the actual forum discussion text. A post can link to parts of other previously existing posts, indicating a targeted reply to that particular text.

We note that the account we use to crawl the forum does not have first-level access to restricted forums. The next section offers more details on the sections of the forum that are restricted.

Exploratory Analysis

The dataset spans over 20 years. The first post is from February 2005, and the last post is from August 2025 (when our crawling ended).

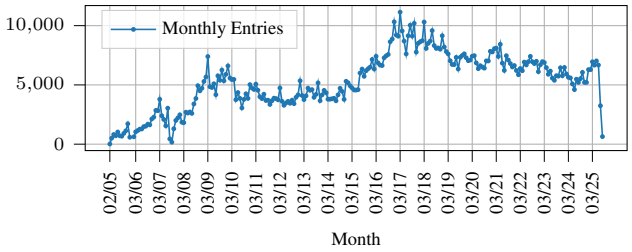

Figure 1: Number of posts (y-axis) per month in the forum.

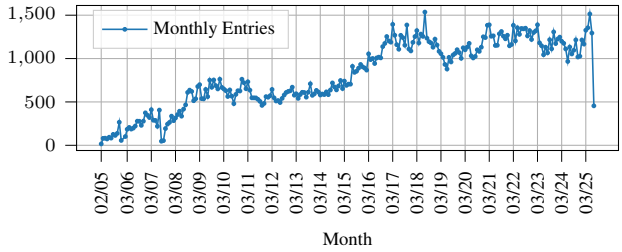

Figure 2: Number of threads (y-axis) per month in the forum.

Posting Activity. Our dataset contains 1,304,542 posts across 200,889 threads. On average, 177 posts and 28 threads appear in the forum each day. We investigate the distribution of posts in Figure 1, and threads in Figure 2 over time. We observe a steady increase in posts per month from the forum's founding in 2005 through March 2017, when a peak of 11,139 new posts was reached. Over the following two years, the number of posts per month decreased slightly and has remained stable since, fluctuating between 5,000 and 7,500 per month. We observe a similar pattern in newly appearing threads per month; however, after a plateau around March 2017, the number of newly appearing threads remained stable at approximately 1,250 per month, and the thread count exceeded 1,500 twice: in July 2018 and May 2025. This led us to investigate the number of posts per thread. On average, there are 6.5 posts per thread; the median is 3. The maximum is 6,967 posts (from a thread related to hash cracking, where participants typically post a cryptographic hash and ask the community to recover the corresponding message, a practice often used in password-cracking attacks). The minimum is 1, i.e., a post without replies.

Contained Artifacts. First, we report the annotations of our crawler, as described in Table 1, and find a number of 1,990 posts containing an ***IFRAME***, and 40,341 posts containing an ***IMG***. Moreover, we count how many times artifacts are mentioned in the posts with `iocsearcher` (Caballero et al. 2023). Three posts could

Table	Attribute	PSQL type	Description
Forum	IdForum	integer	Unique identifier of the forum in the site
	NumThreads	integer	Number of threads in DB for this forum
	Title	character varying(512)	Name of the forum
	URL	character varying(512)	URL of this forum
Thread	IdThread	integer	Unique identifier of the thread in the site
	Author	integer	ID of the member initiating this thread
	AuthorName	character varying(90)	Name of the member initiating this thread
	Forum	integer	Forum ID to which this thread belongs to
	Heading	character varying(512)	Name given to the thread by the initiator
	NumPosts	integer	Number of posts in database for this thread
	NumViews	integer	Number of views of the thread
	URL	character varying(512)	URL of this thread
Post	IdPost	integer	Unique identifier of post in the site
	Author	integer	ID of the member writing this post
	AuthorName	character varying	Name of the member writing this post
	Thread	integer	Thread ID to which this post belongs to
	Timestamp	timestamp with time zone	Indicates where the post was written
	Content	text	Actual content of the post
	AuthorNumPosts	integer	Number of posts of the author (as seen in the post)
	AuthorReputation	integer	Reputation of the author (as seen in the post)

Table 2: Description of the attributes of the Database.

not be processed with `iocsearcher` and are excluded. In the remaining posts we count 2,931 Bitcoin addresses, 517,645 URLs, 214,889 email addresses, 108,107 IPs, as well as 84,815 MD5, 16,070 Sha1, and 12,451 Sha256 hashes. These artifacts have been studied to predict posts containing Indicators of Compromise (Mischinger, Pastrana, and Suarez-Tangil 2024).

Forum Language. To examine communication patterns, we analyze the distribution of languages across all forum posts. Language identification was performed using langdetect.[1] The results indicate that Russian dominated the forum, accounting for 72.4% of all posts, followed by English (14.4%) and Bulgarian (3.9%). Less frequently used languages, such as Macedonian and French, collectively accounted for less than 9.2%. This distribution suggests that, although the forum primarily caters to a Russian-speaking audience, notable linguistic diversity exists, reflecting its international user base and highlighting opportunities for multilingual engagement. In this regard, a comparison of the topics discussed in the English language and Russian language of this forum revealed topics only discussed in one language (Mischinger et al. 2026).

Weekly Activity. We investigate the posting activity throughout the week in Figure 3. The least posting activity is between 2:00 and 3:00 UTC. Afterwards, the activity steadily increases, peaks around 12:00 to 15:00 UTC, and then steadily decreases again. These time patterns correspond to daytime hours in Europe and Western Asia. This is unsurprising, given that the main forum language is Russian and many countries where Russian is spoken are in this region. Russia spans 11 time zones from UTC+02:00 to UTC+12:00. The peak activity would occur in the late evening and night in the eastern parts of Russia. Furthermore, we observe a weekday pattern: most posts are made

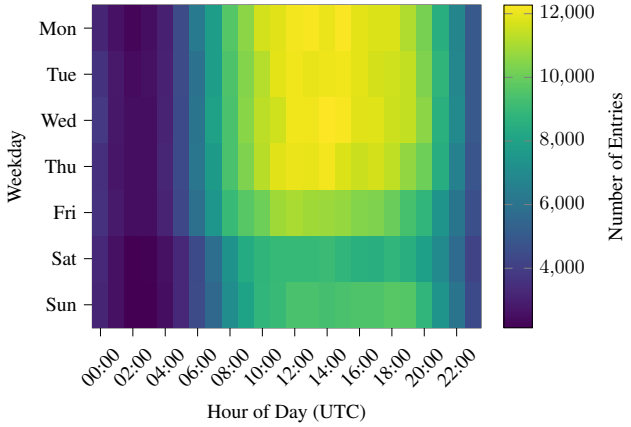

Figure 3: The activity over hours of the week.

during the week, whereas activity is lower on weekends, particularly Saturdays. This activity patterns would fit labor time in Europe and Western Asia.

Sub-Forum Activity. We collect 52 sub-forums. In summary, the sub-forums are either technical, entertainment-related, product-buying and selling, or organizational. A further description about sub-forum titles and their categories is available within the dataset. Table 3 shows the top-most active sub-forums over all years, the most recent ones (2025, 2024, 2023), and earlier years (2020, 2015, 2010). For clarity and ease of reading, the sub-forum names listed in the table are aliases that closely resemble the original titles and their categories. In the early years of the forum up until 2013 the sub-forum for *Other* topics was most used. Afterwards, the discussions shifted to the forums for buying and selling, where we observe a high level of activity to this day. This aligns with the transition of cybercrime into a well-established industry, in which cybercriminals operate

[1]https://pypi.org/project/langdetect/

through a constellation of building blocks to carry out their activities. We also observe that the sub-forum *Arbitration* has grown in popularity over the past three years, reflecting the forum's efforts to prevent it from turning into a lemon market.

ID	Forum Category	Number of Posts
Top sub-forums all years		
23	Other topics	112,124
85	Buying and Selling *Software*	102,539
88	Buying and Selling *Other*	72,646
84	Buying and Selling *Spam*	71,753
89	Buying and Selling *Jobs*	65,025
77	Auctions	59,514
86	Buying and Selling *Finance*	56,792
3	Security and Hacking	54,310
81	Buying and Selling *Traffic*	50,517
61	Malware	46,818
Top sub-forums 2025		
77	Auctions	4,489
88	Buying and Selling *Other*	3,392
75	Arbitration	3,382
84	Buying and Selling *Spam*	3,332
86	Buying and Selling *Financial*	2,816
Top sub-forums 2024		
77	Auctions	8,388
84	Buying and Selling *Spam*	5,010
88	Buying and Selling *Other*	4,521
75	Arbitration	4,254
89	Buying and Selling *Jobs*	3,934
Top sub-forums 2023		
77	Auctions	9,180
84	Buying and Selling *Spam*	5,621
88	Buying and Selling *Other*	4,535
75	Arbitration	4,359
85	Buying and Selling *Software*	4,205
Top sub-forums 2020		
85	Buying and Selling *Software*	7,100
77	Auctions	7,099
86	Buying and Selling *Finance*	5,716
89	Buying and Selling *Jobs*	5,622
84	Buying and Selling *Spam*	5,453
Top sub-forums 2015		
85	Buying and Selling *Software*	7,813
23	Other topics	5,440
81	Buying and selling *Traffic*	4,170
82	Buying and selling *Access*	3,635
84	Buying and selling *Spam*	3,521
Top sub-forums 2010		
23	Other topics	8,615
88	Buying and Selling *Other*	4,511
61	Malware	4,356
5	Money	3,912
3	Security and Hacking	2,743

Table 3: Showing Sub-forums and their IDs with the most posts for all years together and years individually.

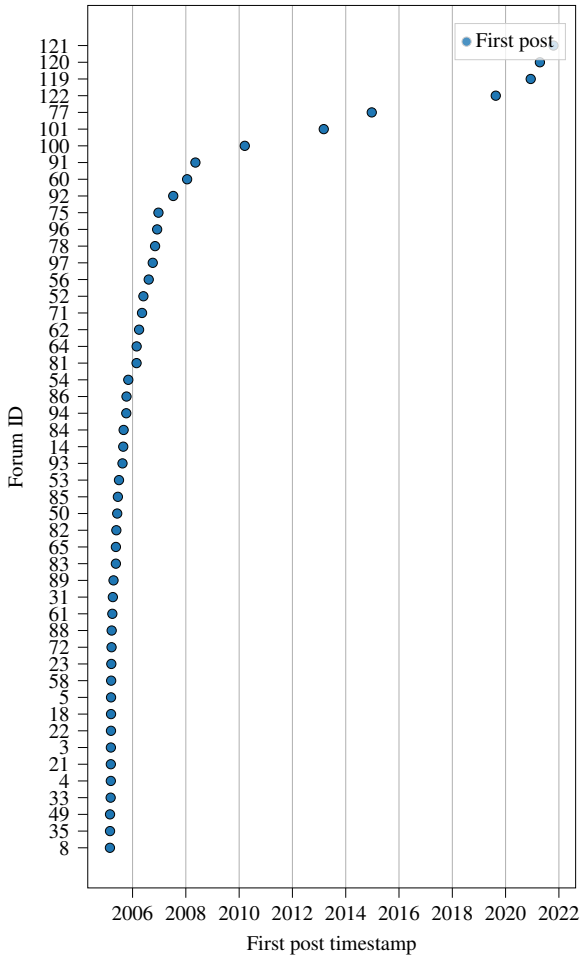

Figure 4: Timestamp of the first post that appeared in a sub-forum.

Viewed differently, Figure 4 depicts the time when each sub-forum got its first post. We observe that most sub-forums emerged within the first three years. However, we see other forums emerging opportunistically with each new technological development. For instance, in 2013, we see the emergence of a cryptocurrency sub-forum (101), which aligns with the period when Bitcoin gained popularity among cybercriminals (Kethineni and Cao 2020). Likewise, in 2019, a sub-forum was created to discuss various traffic-related content in the context of online advertising (122). Furthermore, we also see the integration of evolving market dynamics into the forum's hierarchy. For instance, we observe the emergence of a forum for auctions in 2014 and a forum dedicated to finding investors or partners in 2020; both illustrate the increase in affiliate cybercrime programs over the last decade. We also observe the creation of a sub-forum devoted to discussing instant messaging services in 2010, as well as a disposable notes service and a file-sharing service in 2021. This demonstrates how forums contribute to the infrastructure required for cybercrime, extending beyond mere text-based information exchange.

Restricted Access. Apart from the 52 sub-forums, there is a restricted-access sub-forum that is only accessible if other members vouch for and vote for you. Besides this, we note that a further restricted-access area appears to be transversal to the different sub-forums. These restricted posts may appear in any sub-forum and are hidden from users who have not created a specified number of posts in the forum. We note that the number of restricted posts accounts for a very small fraction of our dataset ($\approx 0.5\%$). Having restricted areas for different skill levels has been observed in previous closed-access forums (Dupont et al. 2017). As interacting with the forum is against our Ethics protocol, we were unable to access or collect data from this sub-forum.

Takeaway. Until around 2017, the forum experienced a steady increase in activity. Afterwards, the number of newly created posts declined slightly, whereas the number of new threads remained steady. This pattern suggests that cybercriminals engage in fewer dialogues per thread, while the overall activity remains high. Forum activity peaks during European and Western Asian working hours. We furthermore observe a shift in the discussion from technical topics toward the buying and selling of diverse cybercrime-related components. This development mirrors the transformation of cybercrime from hands-on technical engineering and expertise to a mature industry in which components can be purchased and assembled without requiring expertise across all domains.

Dataset Applications

Underground forums have been subject to intensive studies for a few years (Motoyama et al. 2011; Hughes et al. 2024). Cybercriminals frequently resort to underground forums to exchange knowledge in preparation for attacks or to distribute newly developed tools before deployment. The topics discussed in these forums often emerge well before they are reflected in formal security reports (Paladini et al. 2024), underscoring that effective CTI should include active monitoring of these forums. Furthermore, systematic analysis of these forums provides investigators with actionable insights, enabling them to anticipate emerging cyberattacks at an early stage and strengthen proactive threat intelligence efforts (Mischinger, Pastrana, and Suarez-Tangil 2024). In this regard, we note that new and more profound research is enabled by rapid advances in artificial intelligence. For instance, statistical ML can be applied to evaluate account behavior or event probability, while Natural Language Processing (NLP) can be used to extract information from the forum text. In any case, the design, implementation, and validation of these techniques depend heavily on the availability of real-world datasets for offline experimental settings. Taken together, we identify the following specific areas that could be interesting for research:

Extraction of CTI. Underground forums are a rich source of CTI that can be used to effectively fight cybercrime (Li and Liao 2024; Gharibshah, Papalexakis, and Faloutsos 2018). Research has investigated underground forums to explore cybercriminals' popular means of payment (Siu, Collier, and Hutchings 2021) , supply chains (Bhalerao et al. 2018) the distribution of malware (Grisham et al. 2017), and the relation of active cybercrime with malware distributed in those forums (Pastrana and Suarez-Tangil 2019), as well as trends and shifts in the cybercriminal infrastructure and strategies (Sood and Enbody 2013).

Social Network analysis. The relevance of individual members, as well as connections among all users in the forum, can reveal important information about upcoming threats. Unlike traditional CTI, which usually focuses on technical aspects, analyzing the social structures behind cybercrime enables the identification of influential actors and collaboration patterns. Reporting on high-level social structures and observations, while ensuring the anonymity of each individual user, aligns with ethical research principles for underground forums. Existing work investigates the posting behavior of well-connected users in underground forums (Pete et al. 2020). Other studies have identified key actors in underground forums (Benjamin and Chen 2012; Pastrana et al. 2018a), and authors active in multiple forums (Frank et al. 2018), revealing how cybercriminal online connections foster cybercrime and how forums facilitate cooperation among users through trust-building mechanisms such as reputation systems.

Social Studies on Underground Communities. Beyond its direct impact on cybersecurity, a social study can reveal interesting insights into these particular underground communities. These could, for example, include a comparison with regular forum communities like Reddit, Facebook, or X, regarding common and distinct aspects or the connection patterns among users. Furthermore, it could be studied if known social phenomena are visible in those forums, like echo chambers, sexism, and hate towards minorities or religion. Also, the level of political streams and involvement can be useful for anticipating potential targets.

Discovery of Dark Jargon. To obfuscate their activities, cybercriminals use benign-looking keywords with a hidden meaning, so-called dark jargon. Automatically detecting and understanding these dark keywords has been studied in literature (Seyler et al. 2021a,b; Li et al. 2021; Yang et al. 2017; Huang et al. 2023). This dark keyword detection has to be done in each language individually (Ke, Chen, and Wang 2022), however, recent work found that Large Language Models (LLMs) show capability correctly understanding such keywords, as well as translation as possible bridge to existing methods designed for English language (Mischinger et al. 2025). Our dataset fuels the detection of Russian dark jargon.

Study of non-English forums. Many studies of underground forums are based on English-language data only (Caines et al. 2018; Pastrana et al. 2018a; Siu, Collier, and Hutchings 2021). The reason for that is, on the one hand, the uncertainty about how to process multilingual data, which has been addressed in recent work (Mischinger et al. 2025), and, on the other hand, the limited availability of multilingual forum datasets. However, region-specific cybercrime activities can only be uncovered by including non-English datasets. Unfortunately, CrimeBB (Pastrana et al.

2018b), the largest publicly available collection of underground forums, is dominated by English forums, with only a small share representing other non-English communities, such as Russian, all of which are open forums. Our dataset is a further step toward supporting the research on a wide range of criminal activities.

Takeaway. The research community has extracted CTI and conducted social network analyses using publicly available datasets, mostly in English and in open hacking communities due to their accessibility and ease of processing. This paradigm, however, leaves a knowledge gap regarding non-English closed-door communities. Russian-speaking hacking communities are a significant hub for eCrime, and many serious activities take place in closed-access spaces. Our dataset is a cornerstone in bridging this gap by offering a multilingual perspective on a closed-access forum.

Related Work

There is a variety of underground economy datasets available for research. CrimeBB (Pastrana et al. 2018b) is the biggest, well-maintained collection of underground forums. As of 2025, CrimeBB contains only open-access forums; our dataset will therefore be a valuable contribution. Our work aims at complementing CrimeBB with our dataset. Both CrimeBB and our dataset, `XIN`, are available through data sharing agreements with the Cambridge Cybercrime Centre. Another dataset of underground forums is AZSecure Hacker Assets Portal (Samtani et al. 2016). Unlike our dataset, which collects forum data, AZSecure also contains hacker source code and attachment assets. Unfortunately, it has been discontinued, and the data is no longer available to the researcher community at the time of writing. DUTA (Al Nabki et al. 2017) is a collection and categorization of Darknet content, in which authors crawled onion domains and manually labeled them into 26 categories. While the dataset is not primarily focused on underground hacking forums, the work describes how such a forum was sorted into the "Hacking" category. In follow-up work, the dataset was extended to a collection of 10,367 onion domains, called DUTA-10K, of which 84% are English and 6% are Russian (Al-Nabki et al. 2019) — all sites accessible without registration. Likewise, another collection of mainly English dark web webpages is CODA (Jin et al. 2022). They categorize the pages into one out of 10 possible categories. These include hacking, porn, drugs, or financial issues. This dataset is not focused on underground hacking forums either, but contains a "hacking" category as well. Finally, Darknet Market Archives contains a collection of 89 darknet markets and more than 37 underground forum datasets. The dataset was collected between 2013 and 2015 and spans $\approx$ 1.6TB of data (Branwen et al. 2015). While restricted underground forums are often accessible through the Darknet, they differ from open Darknet sites in both purpose and structure. Forums are typically closed communities that require vetting or invitations to join, and they function primarily as spaces for knowledge exchange, technical discussions, and community building among hackers or cybercriminal groups. In contrast, open darknet sites are usually accessible with fewer barriers and are oriented toward commerce, serving as marketplaces or advertising platforms for illicit goods and services rather than fostering sustained interaction or knowledge sharing.

Data Availability

With regards to data availability, this repository adheres to the FAIR principles (FORCE11 2020) as follows:

Findability. Our dataset is associated with the unique Digital Object Identifier (DOI): 10.5281/zenodo.17130286

Accessibility. Our dataset, `XIN`, is available through data sharing agreements with the Cambridge Cybercrime Centre.[2]

Interoperability. The dataset is provided in a PostgreSQL file format, allowing researchers to store it offline and operate over it efficiently.

Reusability. A data-usage agreement allows use only by academic researchers.[3] Our dataset is released without further modification, so future work can reuse it as if it were directly crawled from the website. It is provided in a similar format as in other related datasets (Pastrana et al. 2018b).

Ethical Issues

To evaluate the ethical implications of our research, we consulted with the Institutional Review Board (IRB) of the former institution of the senior (last) author and the principal investigator of the project in which data collection was initiated. The data are accessed through a gatekeeper; however, we did not disclose our collection process to this gatekeeper, as such disclosure would have resulted in the termination of our account and rendered the study infeasible. We identified three principal risks associated with data collection. First, there is a potential risk to the privacy of forum participants. Although this risk is mitigated by the fact that users generally have strong incentives to remain anonymous and rarely disclose their legal identities, we adopt additional safeguards. Specifically, we do not attempt to deanonymize users if it were technically possible, and we demand a similar commitment from those who get access to this data. Second, there is a risk of disrupting the forum's normal operation through automated data collection. To address this, our crawler was deliberately designed to emulate human browsing behavior, thereby minimizing server load and preventing interference with the platform's functionality. Third, there is a risk that dissemination of our findings could inadvertently facilitate engagement in cybercrime by drawing attention to illicit pathways. To mitigate this concern, we omit all references to specific individuals or groups observed during data collection and refrain from disclosing the name of the forum studied.

We obtained ethics approval to collect and analyze forum data under application number LRS-19/20-17377. We note

[2] https://www.cambridgecybercrime.uk/datasets.html

[3] https://www.cambridgecybercrime.uk/data.html

that data recipients would require additional approval from their IRB, since each research project involves different ethical implications. We identify three additional risks: (1) Risk to forum authors and the gatekeeper, (2) Risk to the authors of this paper, and (3) Risks of data-sharing agreement violation. We discuss these risks next.

First, we respect the communities privacy by keeping the author names and the forum name anonymous. Second, the authors and the gatekeepers of the dataset acknowledge there is non-trivial risk that any bad actors involved with this forum may seek some form of retribution. We mitigate this risk by not publicly disclosing the actual name of the forum, and we prohibit future researchers from disclosing the forum name. Third, we note that access to the dataset is strictly controlled: researchers must apply for access and sign a data-sharing agreement prohibiting attempts to deanonymize users or engage in any activity that could cause harm. The XIN dataset is published through the Cambridge Cybercrime Centre, where also CrimeBB is hosted and maintained, a widely used research data hub for sharing underground forums. The data-sharing agreement, which is similar to the one used for CrimeBB, strictly permits access only to academic scholars for research purposes alone. We acknowledge that scholars may reshare this dataset, and that, in doing so, the dataset could fall into the "wrong hands". This would violate the agreement and could allow restricted access to be bypassed. However, we regard scholars as reliable and authoritative actors, and we are not aware of any violations of the agreements with CrimeBB since its release eight years ago. We also note that bad actors are already members of these communities or find alternative means to access the knowledge shared in such forums. This creates an asymmetry with benign actors, such as scholars, who struggle getting into closed-access communities. By releasing this dataset to the research community, we aim to address this asymmetry, promote research on mitigating the harm these communities cause, and ultimately benefit society. The tradeoff between the likelihood of scholars violating the agreement and the benefit was factored in our risk assessment. One of the authors of this paper was a co-author of the original development of CrimeBB (Pastrana et al. 2018b) and guided the process of mitigating ethical issues.

Conclusion

In this work, we describe XIN, a dataset from a closed-access Russian-English underground hacking forum. Our dataset contains ≈1.3M posts across ≈200k threads from 52 sub-forums. The collection spans more than 20 years, from February 2005 to August 2025. The data crawling lasted over five years, reflecting the inherent challenges of collecting data from a closed-access forum. The statistical forum analysis highlights how shifts in discussions and activity mirror the broader evolution of cybercrime—from reliance on individual technical expertise to an industry built on readily available, combinable building blocks. We believe our work is an important contribution to the study of the cybercriminal threat landscape. Due to the lack of comparable resources, research so far has largely relied on English-speaking, open-access forums. By making this dataset available for the research community, we enable a more comprehensive study of cybercrime that goes beyond the traditionally studied communities.

Acknowledgments

This work was supported by the project PID2022--143304OB-I00 funded by MICIU/AEI/10.13039/-501100011033/ and by the ERDF, EU. Guillermo Suárez Tangil is a 2020 RyC fellow RYC2020-029401-I funded by MCIU/AEI/10.13039/501100011033 and the ESF Investing in your future. The same grant has funded Mariella Mischinger's work. AI tools were used to support data visualization, as well as to improve the writing style and clarity, grammar, and syntax checks.

References

Al Nabki, M. W.; Fidalgo, E.; Alegre, E.; and De Paz, I. 2017. Classifying illegal activities on tor network based on web textual contents. In *Proceedings of the 15th Conference of the European Chapter of the Associatcn for Computational Linguistics: Volume 1, Long Papers*, 35–43.

Al-Nabki, M. W.; Fidalgo, E.; Alegre, E.; and Fernández-Robles, L. 2019. Torank: Identifying the most influential suspicious domains in the tor network. *Expert Systems with Applications*, 123: 212–226.

Benjamin, V.; and Chen, H. 2012. Securing cyberspace: Identifying key actors in hacker communities. In *2012 IEEE international conference on intelligence and security informatics*, 24–29. IEEE.

Bhalerao, R.; Aliapoulios, M.; Shumailov, I.; Afroz, S.; and McCoy, D. 2018. Towards automatic discovery of cybercrime supply chains. *arXiv preprint arXiv:1812.00381*.

Branwen, G.; Christin, N.; Décary-Hétu, D.; Andersen, R. M.; StExo; Presidente, E.; Anonymous; Lau, D.; Sohhlz, D. K.; Cakic, V.; Buskirk, V.; Whom; McKenna, M.; and Goode, S. 2015. Dark Net Market archives, 2011–2015. https://gwern.net/dnm-archive. Accessed: 2025-09-12.

Caballero, J.; Gomez, G.; Matic, S.; Sánchez, G.; Sebastián, S.; and Villacañas, A. 2023. The Rise of GoodFATR: A Novel Accuracy Comparison Methodology for Indicator Extraction Tools. *Future Generation Computer Systems*, 144: 74–89.

Caballero, J.; Grier, C.; Kreibich, C.; and Paxson, V. 2011. Measuring {Pay-per-Install}: The commoditization of malware distribution. In *20th USENIX Security Symposium (USENIX Security 11)*.

Caines, A.; Pastrana, S.; Hutchings, A.; and Buttery, P. J. 2018. Automatically identifying the function and intent of posts in underground forums. *Crime Science*, 7(1): 1–14.

Campobasso, M.; Burda, P.; and Allodi, L. 2019. Caronte: crawling adversarial resources over non-trusted, high-profile environments. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 433–442. IEEE.

de la Cruz, A.; and Pastrana, S. 2024. Understanding crypter-as-a-service in a popular underground marketplace. *arXiv preprint arXiv:2405.11876*.

Dupont, B.; Côté, A.-M.; Boutin, J.-I.; and Fernandez, J. 2017. Darkode: Recruitment patterns and transactional features of "the most dangerous cybercrime forum in the world". *American Behavioral Scientist*, 61(11): 1219–1243.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Frank, R.; Thomson, M.; Mikhaylov, A.; and Park, A. J. 2018. Putting all eggs in a single basket: A cross-community analysis of 12 hacking forums. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 136–141.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gharibshah, J.; Papalexakis, E. E.; and Faloutsos, M. 2018. RIPEx: Extracting malicious ip addresses from security forums using cross-forum learning. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*, 517–529. Springer.

Grisham, J.; Samtani, S.; Patton, M.; and Chen, H. 2017. Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. In *2017 IEEE international conference on intelligence and security informatics (ISI)*, 13–18. IEEE.

Huang, K.; GrierD, D. W. E. B. C.; Holt, T. J.; Kruegel, C.; McCoy, D.; Savage, S.; and Vigna, G. 2015. Framing dependencies introduced by underground commoditization. In *Workshop on the Economics of Information Security*.

Huang, L.; Wang, S.; Liu, C.; Cao, X.; Han, Y.; Liu, S.; and Chen, Z. 2023. Low-Frequency Aware Unsupervised Detection of Dark Jargon Phrases on Social Platforms. In *Pacific Rim International Conference on Artificial Intelligence*, 198–209. Springer.

Hughes, J.; Pastrana, S.; Hutchings, A.; Afroz, S.; Samtani, S.; Li, W.; and Santana Marin, E. 2024. The art of cybercrime community research. *ACM Computing Surveys*, 56(6): 1–26.

Jin, Y.; Jang, E.; Lee, Y.; Shin, S.; and Chung, J.-W. 2022. Shedding new light on the language of the dark web. *arXiv preprint arXiv:2204.06885*.

Karkallis, P.; Blasco, J.; Suarez-Tangil, G.; and Pastrana, S. 2021. Detecting video-game injectors exchanged in game cheating communities. In *European Symposium On Research In Computer Security*, 305–324. Springer.

Ke, L.; Chen, X.; and Wang, H. 2022. An unsupervised detection framework for Chinese jargons in the darknet. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 458–466.

Kethineni, S.; and Cao, Y. 2020. The rise in popularity of cryptocurrency and associated criminal activity. *International Criminal Justice Review*, 30(3): 325–344.

Li, Y.; Cheng, J.; Huang, C.; Chen, Z.; and Niu, W. 2021. NEDetector: Automatically extracting cybersecurity neologisms from hacker forums. *Journal of Information Security and Applications*, 58: 102784.

Li, Z.; and Liao, X. 2024. Understanding and Analyzing Appraisal Systems in the Underground Marketplaces. In *NDSS*.

Marjanov, T.; and Hutchings, A. 2025. SoK: Digging into the Digital Underworld of Stolen Data Markets. In *2025 IEEE Symposium on Security and Privacy (SP)*, 1–18. IEEE.

Mischinger, M.; Ghafouri, V.; Pastrana, S.; and Suarez-Tangil, G. 2026. Investigating and Comparing Discussion Topics in Multilingual Underground Forums. *arXiv preprint arXiv:2603.21849*.

Mischinger, M.; Hughes, J.; Vitiugin, F.; Pastrana, S.; Hutchings, A.; and Suarez-Tangil, G. 2025. Lost in Translation: Analyzing Non-English Cybercrime Forums. In *2025 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE.

Mischinger, M.; Pastrana, S.; and Suarez-Tangil, G. 2024. Ioc stalker: Early detection of indicators of compromise. In *2024 Annual Computer Security Applications Conference (ACSAC)*, i–xvii. IEEE.

Motoyama, M.; McCoy, D.; Levchenko, K.; Savage, S.; and Voelker, G. M. 2011. An analysis of underground forums. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 71–80.

Paladini, T.; Ferro, L.; Polino, M.; Zanero, S.; and Carminati, M. 2024. You might have known it earlier: Analyzing the role of underground forums in threat intelligence. In *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*, 368–383.

Pastrana, S.; Hutchings, A.; Caines, A.; and Buttery, P. 2018a. Characterizing eve: Analysing cybercrime actors in a large underground forum. In *International symposium on research in attacks, intrusions, and defenses*, 207–227. Springer.

Pastrana, S.; and Suarez-Tangil, G. 2019. A First Look at the Crypto-Mining Malware Ecosystem: A Decade of Unrestricted Wealth. In *Proceedings of the Internet Measurement Conference*, IMC '19, 73–86. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369480.

Pastrana, S.; Thomas, D. R.; Hutchings, A.; and Clayton, R. 2018b. CrimeBB: Enabling cybercrime research on underground forums at scale. In *Proceedings of the 2018 World Wide Web Conference*, 1845–1854.

Pete, I.; Hughes, J.; Chua, Y. T.; and Bada, M. 2020. A social network analysis and comparison of six dark web forums. In *2020 IEEE European symposium on security and privacy workshops (EuroS&PW)*, 484–493. IEEE.

Samtani, S.; Chinn, K.; Larson, C.; and Chen, H. 2016. Azsecure hacker assets portal: Cyber threat intelligence and malware analysis. In *2016 IEEE conference on intelligence and security informatics (ISI)*, 19–24. Ieee.

Seyler, D.; Liu, W.; Wang, X.; and Zhai, C. 2021a. Towards dark jargon interpretation in underground forums. In *European Conference on Information Retrieval*, 393–400. Springer.

Seyler, D.; Liu, W.; Zhang, Y.; Wang, X.; and Zhai, C. 2021b. Darkjargon. net: A platform for understanding underground conversation with latent meaning. In *Proceedings of the 44th International ACM SIGIR Conference on*

Research and Development in Information Retrieval, 2526–2530.

Siu, G. A.; Collier, B.; and Hutchings, A. 2021. Follow the money: The relationship between currency exchange and illicit behaviour in an underground forum. In *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 191–201. IEEE.

Sood, A. K.; and Enbody, R. J. 2013. Crimeware-as-a-service—a survey of commoditized crimeware in the underground market. *International journal of critical infrastructure protection*, 6(1): 28–38.

Soska, K.; and Christin, N. 2015. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th USENIX security symposium (USENIX security 15)*, 33–48.

Stone-Gross, B.; Holz, T.; Stringhini, G.; and Vigna, G. 2011. The Underground Economy of Spam: A Botmaster's Perspective of Coordinating {Large-Scale} Spam Campaigns. In *4th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 11)*.

Tereszkowski-Kaminski, M.; Pastrana, S.; Blasco, J.; Suarez-Tangil, G.; et al. 2022. Towards improving code stylometry analysis in underground forums. In *Proceedings on Privacy Enhancing Technologies (PETS)*.

Thomas, D. R.; Pastrana, S.; Hutchings, A.; Clayton, R.; and Beresford, A. R. 2017. Ethical issues in research using datasets of illicit origin. In *Proceedings of the 2017 Internet Measurement Conference*, 445–462.

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

Yang, H.; Ma, X.; Du, K.; Li, Z.; Duan, H.; Su, X.; Liu, G.; Geng, Z.; and Wu, J. 2017. How to learn klingon without a dictionary: Detection and measurement of black keywords used by the underground economy. In *2017 IEEE Symposium on Security and Privacy (SP)*, 751–769. IEEE.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? The data-usage agreement does not allow the deanonymization of users.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? NA.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we clarify we collect all data available except content posted in the restricted area and discuss its implications in both the Crawling Methodology and Data Collection and the Exploratory Analysis sections.
   (e) Did you describe the limitations of your work? Yes, we discuss why we were unable to access (and thus collect) data from restricted areas in the forum
   (f) Did you discuss any potential negative societal impacts of your work? Yes, there is a risk that dissemination of our work could inadvertently facilitate engagement in cybercrime by drawing attention to illicit pathways. Due to these underlying risks, we discuss the measures we put in place to avoid harm, which include a very strict sharing policy, maintained by the Cambridge Cybercrime Centre, that includes a binding legal commitment through data agreement. We discuss this in the Ethical Issues section.
   (g) Did you discuss any potential misuse of your work? The data-usage agreement prevents this.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we discuss the measures we put in place to control the access to this data, which include a very strict sharing policy, maintained by the Cambridge Cybercrime Centre, that includes a binding legal commitment through data agreement. We discuss this in the Ethical Issues section.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? We discuss ethics in the Ethical Issues Section.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? Yes, we mentioned that the data will be released through a legal agreement with Cambridge Cybercrime Center (CCC). According to such agreement, "*CCC has agreed to sub-license certain of its rights in the Data to the recipients for use in the Project on the terms set out in this Agreement*".
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, we are releasing a new dataset, for which we are providing the corresponding DOI
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes. It may be technically possible to seek informed consent from the forum members through private messaging, but this might be considered spam. Also, notifying users could affect the results (Soska and Christin 2015). Following the ethical guidelines from previous

works (Pastrana et al. 2018b; Thomas et al. 2017), informed consent is not required since the data will be used for research on collective behavior, without aiming to identify particular members. We discuss this further in the Ethical Issues Section.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, in Ethical Issues section we discuss that users of this forum generally have strong incentives to remain anonymous and rarely share personally identifiable information (PII). We have not come across PII, but we acknowledge that crawling web data is not absent of risk. Additionally, we note that the content in this forum could be considered offensive by nature. Due to these underlying risks, we discuss the measures we put in place to avoid harm, which include a very strict sharing policy, maintained by the Cambridge Cybercrime Centre, that includes a binding legal commitment through data agreement.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? We discuss this in Section Data Availability.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the (see Gebru et al. (2021))? Yes, it is appended to this checklist.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? We do not conduct research with human subjects.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? We do not conduct research with human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? We do not conduct research with human subjects.

(d) Did you discuss how data is stored, shared, and deidentified? We do not conduct research with human subjects.

Datasheets for Datasets

This is the datasheet for datasets (Gebru et al. 2021), as required by the Paper Checklist.

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Some collections of open-access forums are available for research, the biggest one being CrimeBB (Pastrana et al. 2018b). However, there is a lack of closed-access forums. We release this closed-access underground forum dataset to complement existing corpora, enabling further research of the cybercriminal ecosystem.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Mariella Mischinger (IMDEA Networks Institute & Universidad Carlos III de Madrid), Sergio Pastrana (Universidad Carlos III de Madrid), and Guillermo Suarez-Tangil (IMDEA Networks Institute).

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Guillermo Suárez Tangil is a 2020 RyC fellow RYC2020-029401-I funded by MCIU/AEI/10.13039/501100011033 and the ESF Investing in your future. The same grant has funded Mariella Mischinger's work.

Any other comments?

None

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description

The instances are textual content from forum posts, including names and titles of threads and sub-forums, collected from an underground hacking forum.

How many instances are there in total (of each type, if appropriate)?

$\approx$ 1.3M posts, $\approx$ 200k threads, 52 sub-forums

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

This dataset contains all the instances collected at the time of submission. The complete dataset is released through the Cambridge Cybercrime Center[4] under a legal agreement. A testimonial sample can be accessed through: 10.5281/zenodo.17130286

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

The dataset contains the raw data as collected from the forum. It contains also information regarding the subforums, threads and posts.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, Table 2 describes all different attributes and their variable type that can be found in the database.

[4]https://www.cambridgecybercrime.uk/data.html

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Parts of some threads can not be directly viewed by our crawlers due to permission restrictions of our account. Such content is hidden behind the text *to view this content you need to create at least x posts*, whereby x are different numbers.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Yes, the posts are grouped in threads, and have an increasing identifier which indicates which posts are replying to each other. Also, when one post cites another (even from another thread), this information is kept in the dataset.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

As mentioned previously, some post content cannot be viewed on our account due to access restrictions.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self contained. External links (URLs) mentioned in the posts are annotated, but not visited by our crawler. The availability, freshness and correctness of those links may change over time and it is beyond this dataset.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

The data is publicly available. Nevertheless, as discussed in the Ethics section, there is a risk to the privacy of forum users. Strict access control permits data access only for research and prohibits any attempt to deanonymize users.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

The data is collected from a cybercriminal underground forum. Although forum rules prohibit offensive and harassing content, collected data may include content that is illegal, offensive, insulting, threatening, or otherwise anxiety-inducing.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how

While the nature of the forum enables user anonymity, the paper discusses existing risks and the mitigations implemented.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The dataset contains no sensitive information other than elements already identified.

Any other comments?

No

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The information was directly extracted from the forum webpage as raw text.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

We developed custom software that crawls and scrapes data into a local database. The crawler was designed and implemented using robust programming techniques to detect and fix errors, and it underwent multiple debugging steps prior to deployment, using a stealthy (limited and paced daily HTTP request volume) crawl. No special hardware was required.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Only co-authors of this paper were involved in the collection process. One of the co-authors is a PhD student, the two others are senior researchers.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were crawled between 2020 and 2025, whereas the original data were posted in the forum between 2005 and 2025.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes, we discuss Ethics in section.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We collected the data from the forum website.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or other- wise reproduce, the exact language of the notification itself.

The individuals were not informed about data collection, as doing so would have led to the termination of data collection and thus affected its quality. We discuss this in the Ethics section.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

As discussed, forum users were not informed about data collection. Full disclosure would have resulted in the immediate termination of our account, thereby rendering the study infeasible. Consequently, we adhered to the ethical and legal precautions established in previous work (see Ethics section).

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes, a risk assessment was conducted and submitted before data collection began to the former institution of the senior (last) author and to the principal investigator of the project in which data collection was initiated. Our data collection followed the established procedures of CrimeBB (Pastrana et al. 2018b), a longitudinal research hub maintained by the Cambridge Cybercrime Centre. One of the authors of this paper was a co-developer of CrimeBB and guided the mitigation of ethical issues. To prevent harmful impact on forum users, the dataset is shared through the Cambridge Cybercrime Centre, and its use is governed by the legal agreement.[5]

Any other comments?

No

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing val- ues)? If so, please provide a description. If not, you may skip the remaining questions in this section.

No.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

N/A

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

N/A

Any other comments?

No

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

A subset of the dataset was used for the analysis in existing publications (Mischinger, Pastrana, and Suarez-Tangil 2024) and also ongoing research work.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No

What (other) tasks could the dataset be used for?

The dataset may be used only by the academic community for scientific research, including, but not limited to, cybercrime.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

The dataset does contain user information. However, the controlled data access through the Cambridge Cybercrime Center strictly prohibits the attempt to deanonymize users.

[5] https://www.cambridgecybercrime.uk/data.html)

This should be considered when processing the data and reporting results.

Are there tasks for which the dataset should not be used? If so, please provide a description.

This dataset is available for research purposes only and is subject to the data-sharing legal agreement. Particularly, it is forbidden to reverse engineer the data, use it for any commercial purpose, or use it for illegal activities.

Any other comments?

No

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Our dataset is available through data sharing agreements with the Cambridge Cybercrime Centre.[6]

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The data can be accessed through a custom website from the Cambridge Cybercrime Centre, with access restriction to those who have signed the data sharing agreement.

When will the dataset be distributed?

The dataset will be released upon the publication of this paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The copyright belongs to the post authors unless otherwise stated. When using the dataset, this paper should be cited.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Unknown to the authors of the datasheet.

Any other comments?

No

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The Cambridge Cybercrime Center.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

All information can be found here: https://www.cambridgecybercrime.uk/data.html. The email address for contacting is datarequest@cambridgecybercrime.uk

Is there an erratum? If so, please provide a link or other access point

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

This will be communicated and managed by The Cambridge Cybercrime Center.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced

The data contains all historical data kept by the forum at the time of this collection. There are no data retention restrictions associated to this data.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

All the posts in the dataset contain a timestamp. Accordingly, all data recipients must understand how data freshness affects the conclusions of their work (e.g., conducting research on cyberattacks using data from 8 years ago may have historical interest but may not be useful for designing modern cyberdefenses). When updated, data recipients will be individually communicated.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

The dataset could be extended by crawling later posts and providing them in the same data format. This should be discussed with the Cambridge Cybercrime Center.

Any other comments?

No

[6] https://www.cambridgecybercrime.uk/datasets.html