DMDb: Uncovering Criminal Hacking on the Dark Web to Enhance Cyber Threat Intelligence Research

Wesley H. Kwan
Cal Poly Pomona
whkwan@cpp.edu

Lynn K. Takahashi
Cal Poly Pomona
lktakahashi@cpp.edu

Nathan Pham
Cal Poly Pomona
nathanpham@cpp.edu

Apurva Sista
Cal Poly Pomona
apurvasista@cpp.edu

Minh Khoi Tran
Cal Poly Pomona
minhkhoitran@cpp.edu

Vincent C. Lee
Cal Poly Pomona
vclee@cpp.edu

Siwen (Victor) Wang
Cal Poly Pomona
siwenwang1@cpp.edu

Ericsson Marin
Cal Poly Pomona
santanamarin@cpp.edu

Abstract

The emergence of the dark web has enabled hackers to anonymously exchange information and trade malware worldwide, exposing organizations to an unprecedented number of threats. Without visibility into this offensive base, defenders are often left to mitigate damage. While prior cyber-threat intelligence research has been valuable, it has been constrained by incomplete, outdated, and noisy datasets. In this paper, we detail our efforts to build a comprehensive repository that illuminates the current plans of cyber-attackers. We achieve this by designing and deploying DarkMiner, a system that regularly scrapes the Tor network to populate the DarkMiner Database (DMDb). DMDb offers researchers a structured criminal hacking data collection enhanced with non-textual fields and object change tracking capabilities. To show its potential, we present three case studies analyzing: 1) cyber threat market fluctuations, 2) image-based vendor attribution, and 3) software vulnerability targeting.

Keywords: Dark web, hacking, threats, database.

1. Introduction

With the recent rise in cyber-attacks, cybersecurity concerns have reached unprecedented levels. Kaspersky Lab reported that 1,198,396 cyber-attacks were repelled daily by the company in 2023 (Kaspersky, 2023), reflecting the scale of criminal cyber threat activity. Worldwide, cyber-attacks cost organizations 8 trillion US dollars in 2023 (7% of the global GDP), and 10.5 trillion US dollars in costs are expected annually by 2025 (Cybersecurity Ventures, 2023).

A credible explanation for this threatening scenario is that malicious hackers increasingly use the dark web to share knowledge and achieve their goals (Bermudez-Villalva & Stringhini, 2021; Boshmaf et al., 2023; Robertson et al., 2017; Schäfer et al., 2019). Cybercriminals who want to operate under the radar of law enforcement agencies can now use platforms to shield their communications and product transactions, making what was once a hard-to-penetrate business accessible to a much broader population. This scenario creates a paradoxical condition in cyberspace that fosters malware distribution while giving defenders a valuable resource to mine emerging cyber threats (Almukaynizi et al., 2018; Marin et al., 2021; Nunes et al., 2016). However, large-scale analysis of hacker communication, especially on the dark web, has been limited due to a lack of available datasets that are complete, authorized (not leaked), and up-to-date (Boshmaf et al., 2023; Hughes et al., 2024). Lastly, these datasets often contain non-hacking content, such as drug-related material or pornography, adding extra challenges to cyber threat intelligence research.

To effectively address these challenges, this research aims to build a comprehensive database that reveals the current activities of cyber-attackers on the dark web. We present our methodology for collecting, processing, and storing hacking-related data through the design and deployment of a multi-component system named DarkMiner. Since September 2023, we have been executing DarkMiner weekly to scrape popular forums and marketplaces on the Tor network, generating a robust and up-to-date criminal hacking repository—the DarkMiner database (DMDb). Currently, the system is acquiring both textual and non-textual data from English-speaking platforms, incrementally updating DMDb. We have developed machine learning classifiers and integrators that accurately identify hacking data in 87% of cases and track object changes over time.

As of this writing, the database contains 31k threads including 790k posts produced by more than 110k users, as well as 14k items offered by more than 1.8k vendors discussing or trading a variety of cyber threats such as infiltration, data theft, and evasion tools. Over 40 pieces of information compose

URI: https://hdl.handle.net/10125/109319
978-0-9981331-8-8

DMDb—e.g., thread titles, post content, and timestamps or item names, descriptions, and prices, with the database being fully available to researchers under an End-User License Agreement (EULA)—redistribution and commercial use are prohibited. To showcase DMDb's capabilities, we present three case studies that: 1) analyze cyber threat market fluctuations driven by demand; 2) investigate how images can be used for vendor attribution; and 3) inspect software vulnerabilities targeted for exploitation.

2. Research background

This section provides the research background, introducing the dark web, cyber threat intelligence, and ethical and security-related issues relevant to our work.

2.1. Dark web

Dark web networks, such as The Onion Router (Tor), are increasingly being adopted by users who want to conceal their online activities (Jeziorowski et al., 2020; Marin et al., 2021; Robertson et al., 2017). Tor, the largest dark web network (Robertson et al., 2017), enables users to browse the internet anonymously by routing their traffic through a series of volunteer-operated "nodes," which obscure the origin and destination of the data (Dingledine et al., 2004). Additionally, Tor enables users to host anonymous, theoretically untraceable websites by implementing extra security measures that help protect onion service providers (Schäfer et al., 2019).

Although the dark web provides a legitimate environment for circumventing censorship, tracking, and surveillance, it has also attracted cybercriminals who seek to keep their activities hidden from law enforcement agencies (Bermudez-Villalva & Stringhini, 2021; Broadhurst et al., 2018; Robertson et al., 2017; Schäfer et al., 2019). The architecture of these underground networks allows for the hosting of platforms, such as forums and marketplaces (Nunes et al., 2016), where users can anonymously engage in illegal activities related to drugs, weapons, pornography, and frauds (Intelliagg, 2016; Soska & Christin, 2015). According to The Tor Project (2024), there were more than 2.5 million daily visitors to the dark web in 2023, with cybercriminals generating about 62% of the current dark web activities (Boshmaf et al., 2023). The global dark web economy was valued at 520 million US dollars in 2023 and is projected to surpass 1 billion US dollars by 2027 (Borgeaud, 2024), making cybercrime one of the biggest challenges of humanity (Morgan, 2019).

Research has revealed that many offerings on dark web marketplaces (Broadhurst et al., 2018; Marin et al., 2016; Meland & Sindre, 2019) and discussions in forum threads (Bermudez-Villalva & Stringhini, 2021; Pete et al., 2022; Schäfer et al., 2019) are highly relevant for cybersecurity. These include information about exploits, stolen datasets containing login credentials, ransomware, botnets available for hire, among other threats. We refer to this type of cybercrime as *criminal hacking*, formally defining the concept below:

Definition 1 *Criminal hacking comprises the subsection of cybercrime activities leveraged to compromise data, reputation, financial security, or disrupt online services—e.g., phishing, carding, keylogging, scams, exploits, ransomware, DDoS, etc.*

To illustrate websites hosted on Tor, Figure 1 showcases a live forum and a live marketplace as they appear as of this writing. Note how malicious hackers are trading or engaging in discussions about malware.

2.2. Cyber threat intelligence

While the popularity of dark web forums and marketplaces contributes to malware proliferation, it also provides intelligence for defenders. The digital traces left by hackers—e.g., interests, plans, targets, and assets—offer critical insights into evolving cyber threats, thereby fueling the cyber threat intelligence industry (Hughes et al., 2024; Robertson et al., 2017).

In summary, cyber threat intelligence (CTI) involves the systematic analysis of cyber threat related data to generate insights on future cyber-attacks (Nunes et al., 2016). From this perspective, CTI differs from other

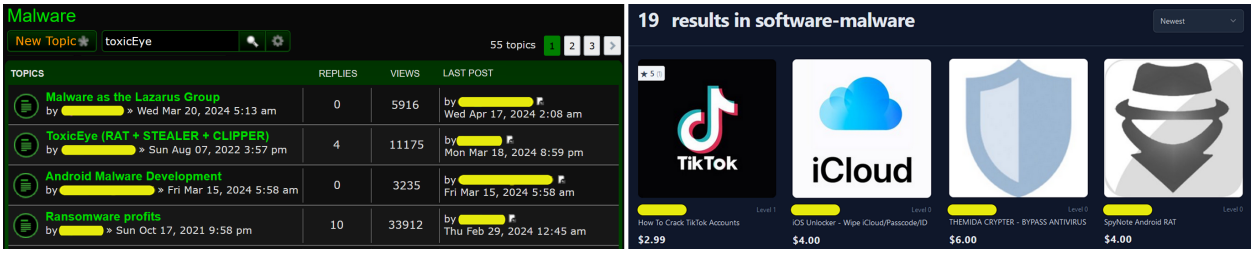

Figure 1: Examples of forum threads (left) and marketplace items (right) from websites hosted on Tor (onion services).

cybersecurity fields by emphasizing prevention over remediation. The field has emerged from the recognition that defensive measures alone are insufficient to address cybersecurity challenges (Marin et al., 2021).

Extensive research has implemented various CTI approaches to combat cybercriminals, including the detection of key hackers (Marin, Shakarian, & Shakarian, 2018; Samtani & Chen, 2016), the prediction of software vulnerability exploitation (Almukaynizi et al., 2017; Bullough et al., 2017), the anticipation of real-world attacks (Almukaynizi et al., 2018; Marin et al., 2019), and the prediction of hacker engagement (Marin, Almukaynizi, et al., 2018). These studies combined machine learning, natural language processing, and social network analysis to conduct descriptive and predictive CTI-related tasks using dark web data—see (Hughes et al., 2024) for an overview.

2.3. Ethical and security considerations

Ethical and security issues should be carefully considered when dealing with the dark web (Pastrana et al., 2018). As reported by (Hughes et al., 2024), only a minority of articles discussed ethics, with 13% receiving approval and 8.7% receiving an exemption.

For this study, we formaly submitted a research protocol to our Institutional Review Board (IRB). The project was granted "exempt" status due to the passive nature of the research, as it only analyzes publicly available data. Thus, no manual or automated interactions with posts or payments are made during the data scraping process. Despite the exemption, the IRB provided specific recommendations, which we are diligently following: 1) researchers should not attempt to identify users personally, 2) images should never be stored in their original form to prevent the possession of illegal content, such as those involving child exploitation material, 3) when reporting findings, the names of platforms and their users should be anonymized to protect privacy, 4) the built database should only be shared with professionals from accredited institutions, who must sign an agreement to prevent misuse.

In addition, the IRB also provided the following security measurements to safeguard our infrastructure and research environment: 1) access to the dark web should only be done through virtual machines to protect the host operating system, 2) attachments should never be downloaded, as they may contain hidden malware, 3) researchers should not provide any indicators of their identities when registering accounts online.

3. Related work

A key challenge in cybercrime investigations is the acquisition of credible data. While researchers may be reluctant to share their datasets due to privacy or copyright concerns, publicly available datasets are often incomplete, outdated, or obtained through unauthorized access like leaks (Hughes et al., 2024). Consequently, individuals may need to develop their own data collection infrastructure to explore CTI-research.

We acknowledge existing research introducing scraping tools that automate dark web data collection, such as BlackWidow (Hughes et al., 2024) and PostCog (Pete et al., 2022). However, we compare our work only with projects that have both collected and shared dark web data, as many researchers, especially those outside computer science or related fields, may lack the resources and expertise to run and manage the entire scraping processes. Our final list of papers is presented in Table 1, where we assess whether these projects meet seven proposed CTI-based database design criteria.

These criteria include: 1) *Targeted Platforms*—the types of platforms scraped, 2) *Scope*—the number of websites covered, 3) *Data Format*—how the data was organized, 4) *Type of Data*—whether images were provided or represented alongside the text, 5) *Classification*—whether hacking content was identified or categorized, 6) *Maintenance*—whether the data is regularly updated, and 7) *IRB Review*—whether the

Table 1: Research papers that collect and share dark web data.

Research papers	Targeted Platforms: ● Marketplaces and Forums ◐ Marketplaces or Forums	Scope: ● Multiple websites ◐ Few websites	Data Format: ● Structured ◐ Semi-Structured	Type of Data: ● Textual and Non-textual ○ Only textual	Classification: ● Filter hacking ○ Do not filter hacking	Maintenance: ● Up-to-date ○ Outdated	IRB Review: ● Approved or Exempt ○ Not mentioned
(Branwen et al., 2015)	●	●	◐	○	○	○	○
(Nunes et al., 2016)	●	●	●	○	●	○	○
(AZSecure, 2017)	●	●	◐	○	○	○	○
(Pastrana et al., 2018)	◐	●	●	○	○	●	●
(Bhalerao et al., 2019)	◐	◐	◐	○	●	○	●
(Campobasso & Allodi, 2023)	◐	◐	●	○	○	○	●
(Boshmaf et al., 2023)	●	●	◐	●	○	●	●
The DarkMiner Database (DMDb)	●	●	●	●	●	●	●

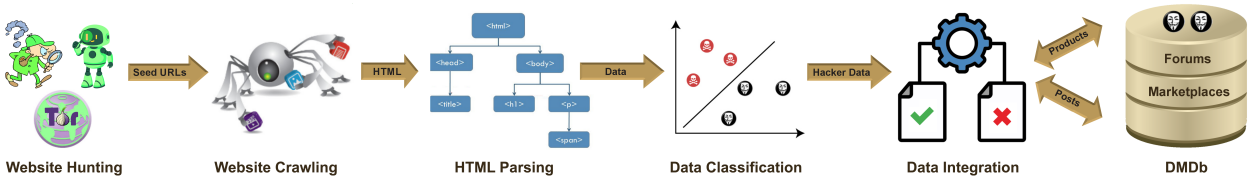

Figure 2: Methodology for data collection, processing, and storage.

research was approved or exempted by IRB. As noted, only DMDb satisfies the seven criteria, underscoring the relevance of the database for CTI-related projects.

4. Methodology

In this section, we present our methodology for data collection, processing, and storage, detailing the sequential tasks of Figure 2. We describe how these tasks are accomplished using DarkMiner, a system designed to update DMDb regularly. The system has been operational on the Tor network since September 2023, collecting about 400 new forum threads (13k posts) and 200 new marketplace items weekly.

4.1. Website hunting

To continuously feed DMDb, DarkMiner uses two types of website hunters: human analysts and robots. Analysts manually use dark web search engines, wikis, and catalogs to locate new forums and marketplaces hosted on the Tor network. They may create random accounts to access websites and assess whether they include hacking data. If so, their URLs are saved. For the automated part, a robot uses regular expressions to search for onion links in forum posts stored in DMDb. When a link is found, the post and its neighborhood content are saved for manual analysis, providing context for what is being offered. Figure 3 presents an instance where a responding hacker shares an explicit onion link to "site hacking" learning resources.

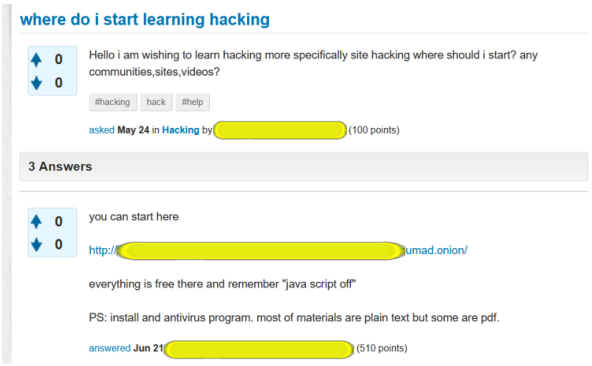

Figure 3: Example of a forum post with .onion links.

4.2. Web crawling

DarkMiner uses Tor-based crawlers implemented with Selenium[1] to automatically access and traverse websites, downloading HTML pages for processing. Due to platform differences, we designed a custom crawler for each website, following the data collection goals outlined by Pastrana et al. (2018) below.

Completeness. Forum boards or marketplace categories have listing pages that display multiple threads or items, each linked to one or more description pages with detailed information. Starting from a seed URL, DarkMiner employs targeted crawling to visit all listing and description pages within the sections of a website containing hacking-related information.

Incremental crawling. Incremental crawling is an efficient strategy for collecting web data (Pastrana et al., 2018). However, it has a significant drawback: it misses object data changes. For instance, if an item's price is modified, incremental crawlers may overlook the modification by skipping previously visited URLs, potentially leading to database freshness issues. To address this, DarkMiner fully re-crawls targeted sections of websites to capture any object modifications.

Accessibility. During the authentication process for multiple websites, Darkminer's crawlers automatically log in using credentials created by human analysts and semi-automatically solve CAPTCHAs. Currently, we achieve 80% accuracy in automatically solving text-based CAPTCHAs using deep learning techniques, such as YOLOv8[2], as illustrated in Figure 4. Advanced types of CAPTCHAs[3] are being solved manually.

Figure 4: Example of real-world CAPTCHA being solved.

Flexibility. Adding new websites to DarkMiner requires minimal development effort. We designed an architecture that separates the scraper (crawler

[1] https://www.selenium.dev/
[2] https://docs.ultralytics.com/
[3] Image, audio, video, and puzzle-based CAPTCHAs.

and parser) implementation from other system functionalities, such as scheduling, classification, integration, and error logging. This allows for testing new scrapers without impacting other components.

Verbosity. Crawling and error logs help diagnose issues, implement fixes, and ensure reliable data collection. DarkMiner generates and monitors these logs to quickly detect events or changes that may impact scrape quality. For example, a drop in downloaded pages or an increase in errors could indicate a structural change in the target site, requiring adjustments.

Stealthiness. If a crawler is detected while traversing a website, the session may be terminated or the account banned, preventing a complete data collection process. To mitigate this, DarkMiner mimics human behavior by randomizing navigation patterns and introducing random time delays for resource requests.

Efficiency. To speed up crawling, DarkMiner runs its crawlers in parallel, collecting data from multiple websites simultaneously. Each crawler also tracks visited URLs during a session, avoiding revisits.

Non-textual content. Handling non-textual content is important for both data security and comprehensiveness. To avoid downloading harmful data, DarkMiner sanitizes web pages by removing scripts and attachments. Additionally, images are converted to screenshots, encrypted, and re-embedded into the HTML pages before they are saved.

4.3. HTML parsing

As with crawlers, we designed a custom parser for each website. DarkMiner uses BeautifulSoup[4] to navigate the HTML structure of webpages, extracting data for the objects and fields listed in Table 2.

Table 2: Objects and fields extracted by parsers.

Forums

Object	Fields
Thread	title, date, n_views, n_posts, board, URL, author
Post	thread, content, date, image, feedback, user
User	name, status, reputa., interest, signat., image

Marketplaces

Object	Fields
Item	name, description, price, n_sold, n_left, n_reviews, n_views, ship_from, ship_to, rating, image, category, URL, vendor
Vendor	name, n_transactions, rating, image

[4]https://beautiful-soup-4.readthedocs.io/en/latest/.

To enable the safe sharing of dark web image data, DarkMiner decrypts the collected images to generate informative, yet abstract image representations using computer vision tools, including ResNet-50 and SIFT. ResNet-50 is a deep convolutional neural network that excels at extracting high-level features for image classification (He et al., 2016). As Wang et al. (2018) demonstrated its effectiveness at linking vendor profile images across marketplaces, we use the output of ResNet-50's last hidden layer to represent each image. Additionally, we store SIFT (Scale-Invariant Feature Transform) key points and descriptors (Lowe, 1999), which are useful for object detection tasks like logo matching (Bharathidevi et al., 2017). Figure 5 provides an example of both image representation processes.

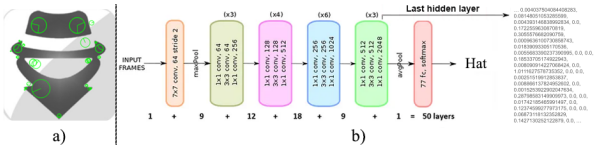

Figure 5: SIFT key points detection (a) and ResNet-50 image encoding and classification (b) examples.

4.4. Data classification

Despite implementing targeted crawling, DarkMiner often collects non-hacking-related data due to unorganized website sections. To address this, we designed two classifiers for the system: one for forums and other for marketplaces. These classifiers analyze threads or items to assign them a probabilistic score based on their relevance to hacking. A score of 0.5 or higher indicates a positive instance, though the threshold can be adjusted for stricter filtering.

Labeling. Following a supervised learning approach, we randomly labeled 1,000 threads from 10 forums and 1,000 items from 25 marketplaces as positive or negative according to **Definition 1**. Only the thread's title and first post, as well as the item's name and description, were considered, as illustrated in Table 3. Note that negative samples may also represent other forms of cybercrime or benign programming activities.

Text cleaning. The textual data collected from websites often contains noise, e.g., ***NEW***, making it unsuitable for direct use in classification models. Thus, we remove non-alphanumeric characters and stop words to eliminate insignificant terms.

Feature extraction. We compare two word-embedding techniques for feature extraction: Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec (W2V) (Manning et al., 2008; Mikolov et al., 2013). For TF-IDF, we use

Table 3: Examples of labeled forum threads and marketplace items.

	Threads		Items	
Hacking	**Title**	**Post**	**Name**	**Description**
YES	How To Spread Your Viruses	This tutorial shows how to spread your trojans/viruses ...	.Lnk exploit Builder	QuantumBuilder will make your payload looks anything ...
NO	A guide for beginners to C	hello there many members ask where to find good material ...	2.5 g Bio GELATO 33	Gelato 33 is a hybrid marijuana strain made by ...

character n-grams to handle misspellings and word variations while preserving word boundaries—e.g., "exploit" generates "expl," "xplo," "ploi," and "loit." For W2V, we apply spell-checking and lemmatization before training a skip-gram model to capture word meanings and relationships. With both methods, features are independently extracted from the text fields, resulting in a thread title and post vectors for forums and an item name and description vectors for marketplaces. Following Nunes et al. (2016), we concatenate the vectors from each platform to preserve context.

Model training and testing. We employ five-fold cross-validation to measure the F1 score of our classifiers. The F1 score is the harmonic mean of precision (the fraction of predicted positive cases that are actually positive) and recall (the fraction of positive cases correctly identified). We leverage Scikit-Learn[5] to deploy multiple machine learning algorithms, including logistic regression (LR), linear support vector machine (LSVM), and radial basis function support vector machine (RSVM), exploring the following configurations:

- **TF-IDF:** character n-gram range (min and max character length of terms), alpha threshold (term filter threshold for the ANOVA F-value of term features and labels), and learning algorithm.
- **W2V:** window size (number of positive words when forming word pairs) and negatives (number of negative words for each positive word), vector size, aggregation method (sum or average of word vectors), and learning algorithm.

Results. For marketplaces, configurations using TF-IDF and LSVM achieve the best overall performance—see Figure 6 on the left. The classifier with a [2, 6] character n-gram range—see Figure 6 on the right—and an alpha threshold of 1.0 achieves the highest average F1 score of 0.84, recalling 82% of the relevant threads with a precision of 86%. We obtain similar results for forums, with a highest average F1 score of 0.84. Through error analysis, we noted that many "account," "card," and "cashout" samples were mislabeled as negative. Therefore, we re-labeled them as positive and retrained our classifiers, improving the F1 score to 0.87 for forums and 0.88 for marketplaces.

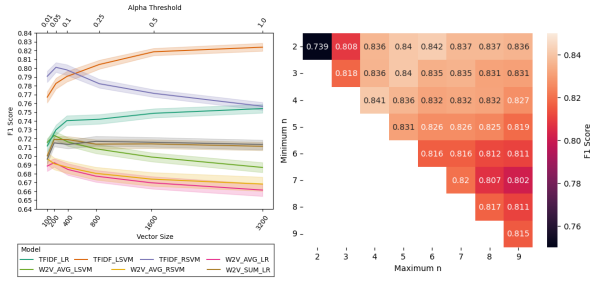

Figure 6: Classification performance based on the feature extraction method and learning algorithm (left) and character n-gram range (right) for marketplace items.

4.5. Data integration

The integrator ensures DMDb receives all scraped data objects and their changes. When a new object, such as an item, is obtained, it is added to the item table. However, if an existing object is received, the integrator checks for any object changes—e.g., an increased item price—updating the corresponding object table while storing its previous version in the history table. Thus, if an item's price is modified four times, DMDb will contain five records: four in an item history table and the most recent one in the item table.

5. DarkMiner database (DMDb)

DMDb is a comprehensive and up-to-date criminal hacking repository populated with dark web data. Implemented using PostgreSQL[6], the database is the result of our continuous data collection efforts while deploying DarkMiner. Table 4 shows the considerable growth of DMDb since its launch in September 2023.

[5]https://scikit-learn.org/stable/

[6]https://www.postgresql.org/

Table 4: Volume of data in DMDb.

Forums

	Sep. 2023	Nov. 2023	Jan. 2024	Mar. 2024	May 2024	Jul. 2024
Forums	8	9	9	10	11	11
Threads	15k (9.6k)	25k (17k)	26k (17k)	42k (27k)	45k (29k)	48k (31k)
Posts	160k (130k)	330k (290k)	340k (300k)	700k (640k)	810k (740k)	870k (790k)
Users	54k (49k)	77k (71k)	78k (73k)	110k (100k)	116k (110k)	121k (110k)
Images	10k	17k	18k	24k	26k	27k

Marketplaces

	Sep. 2023	Nov. 2023	Jan. 2024	Mar. 2024	May 2024	Jul. 2024
Markets	14	20	22	25	26	32
Items	6.0k (2.5k)	17k (7.1k)	18k (7.7k)	21k (9.5k)	32k (12k)	35k (14k)
Vendors	535 (286)	752 (467)	786 (495)	1.1k (646)	3.9k (1.6k)	4.2k (1.8k)
Images	4.2k	9.2k	9.5k	12k	12k	14k

As of July 2024, DMDb contains about 31k threads and 14k items that are hacking-related. These values are shown within parentheses, while values outside indicate the full amount of data collected. We also list in Table 4 the data volume of other associated hacking-related objects—i.e., users, posts, and vendors. Note that for forums, only 64% of the threads and 90% of the posts and users are relevant for criminal hacking, which indicates how much cybercrime data is being filtered. For marketplaces, this filter is even stronger, suggesting that 60% of the items and 58% of the vendors should be discarded for CTI-research.

To analyze DMDb data, we apply k-means to group its hacking-related threads and items into disjoint clusters. In the absence of a single standardized taxonomy for categorizing cyber threat tools and techniques, we defined a taxonomy of seven groups subdivided into 18 subgroups, with the latter used for clustering. We initialize the 18 centroids of k-means by filtering samples that match predefined keywords, such as "xss" for "Exploits." Figure 7 presents the results, showing that "Malware," "Exploits," "Financial Fraud," and "Data Stealers" are the most popular clusters.

We set up a page at https://www.cpp.edu/calsys/database.shtml for those interested in exploring DMDb, offering sample (.csv) files that illustrate the database's structure and content. For full access, researchers can contact the authors to complete an online screening process. DMDb is available in multiple formats to accredited researchers and professionals under an End-User License Agreement (EULA)—redistribution and commercial use are prohibited—with weekly updates to support CTI-research projects or applications.

6. Case studies

To show DMDb's potential for CTI-research, we now present three case studies that analyze: 1) fluctuations of cyber threats, 2) vendor attribution with images, and 3) software vulnerability targeting.

6.1. Fluctuation of cyber threats

In this first case study, we examine whether DMDb forum data can indicate cyber threat fluctuations over time, an approach commonly used by researchers

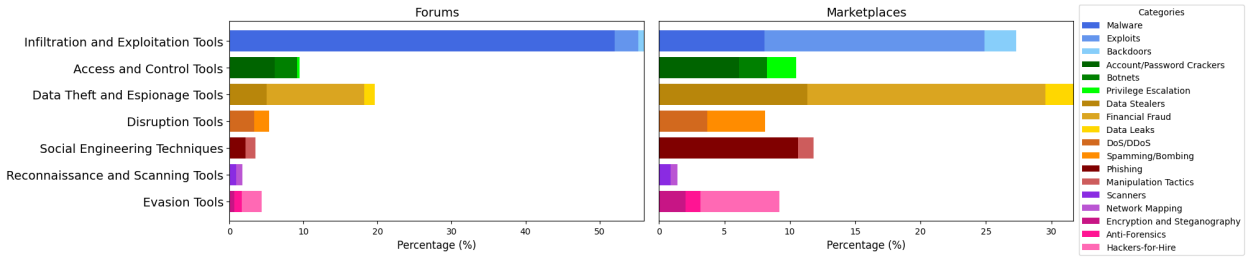

Figure 7: Distribution of cyber threats across all forum threads and marketplace offerings in DMDb.

analyzing market dynamics (Campobasso & Allodi, 2023; Schäfer et al., 2019). These dynamics can reveal correlations between hacker discussions and real-world incidents, often highlighting spikes in hacker activity. If identified early, such spikes may help anticipate future cyber-attacks (Almukaynizi et al., 2018).

To conduct this investigation, we analyze hacker conversations within the 18 subgroups of our cyber threat tools and techniques taxonomy, searching for trends related to specific topics. As shown in Figure 8, there is a slight upward trend in DDoS discussions during June and July 2024, correlating with Cloudfare's Q2 threat report that revealed a 2.1% increase in customers being targeted with ransom DDoS attacks (Staff, 2024). Also, a notable spike in DDoS discussion occurs in August 2024, coinciding with Akamai's report on preventing one of the largest DDoS attacks ever observed on July 15, 2024 (Dummer, 2024).

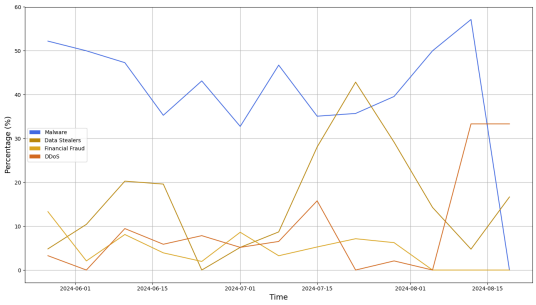

Figure 8: Top four most popular topics of Q3 2024.

During Q3 of 2024, our analysis reveals another spike of hacking activity regarding "Data Stealers." This aligns with the broader trend observed in the first half of 2024 where ransomware such as RisePro, Lumma, and Vidar became increasingly prevalent (Cybersixgill, 2024). The increasing interest may also correlate with the Indian crypto company WazirX that confirmed a data breach in the same month, resulting in the loss of nearly half of its reserves and the theft of 200 different cryptocurrencies (Singh, 2024).

6.2. Vendor attribution with images

Hackers on the dark web frequently manage multiple accounts across forums and marketplaces with varying usernames, making it difficult to infer relationships and detect coordinated criminal activities. The large volume of data on these platforms makes manually investigating and associating accounts labor-intensive. Traditional CTI-research has relied on usernames as ground truths to identify cross-marketplace vendors (Jeziorowski et al., 2020; Marin et al., 2016), but this approach is prone to false positives due to the ease with which usernames can be taken or mimicked. Since it requires more effort to mimic another vendor by copying their images, this case study explores whether images serve as a more reliable method to conduct vendor attribution. Specifically, we analyze whether vendors with similar images also tend to use similar usernames across multiple marketplaces, and vice versa.

In this context, we leverage DMDb data to perform a pairwise comparison of vendors across different marketplaces. For each pair of vendors, two similarity scores are calculated. The first score, LEV, measures one minus the normalized Levenshtein distance between their usernames. The second score, COS, represents the maximum cosine similarity of ResNet-50 image representations from items across vendor pairs.

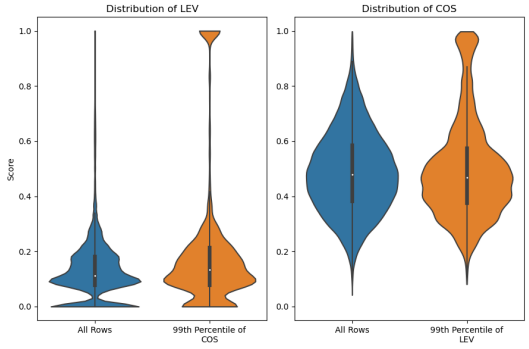

Figure 9: Distribution of LEV and COS scores.

By comparing the overall distribution of LEV scores within the 99th percentile of COS scores, as well as the overall distribution of COS scores within the 99th percentile of LEV scores, we observe that the constrained distributions exhibit higher values compared to the overall distributions—see Figure 9. More importantly, the distribution of LEV scores shift to higher values when constrained by high COS scores than the distribution of COS scores when constrained by high LEV scores. See how the average LEV score for comparisons with a COS score of 0.84 or higher is 0.19, notably higher than the overall average LEV score of 0.13. However, the average COS score for comparisons with a LEV score of 0.63 or higher is 0.50, slightly higher than the overall average COS score of 0.49. This suggests that images may be more reliable to perform vendor attribution across marketplaces than usernames.

6.3. Software vulnerability targeting

As the number of discovered software vulnerabilities grows each year, organizations face an increased risk of cyber-attacks. Vulnerability exploit prediction is a key CTI-task aimed at identifying which vulnerabilities

are likely to be exploited, helping security specialists prioritize patching efforts. Since research has demonstrated the value of incorporating dark web data to enhance these predictive models (Almukaynizi et al., 2017; Bullough et al., 2017), this case study analyzes vulnerability exploitation by answering two questions: 1) What is the ratio of exploited CVEs[7] compared to those disclosed in DMDb? 2) How does DMDb's ratio compare to other data sources in this context?

To conduct this analysis, we gathered data on CVE disclosures from several sources: NVD[8], ZDI[9], EDB[10], DMDb, and KEV[11]. As NVD, ZDI, and EDB are repositories with different goals, their CVE disclosing date provides complementary information. From DMDb, we extract CVEs from forum posts, setting the disclosure date as the date of the first post mentioning the CVE. Lastly, KEV indicates which CVEs were actively exploited and when. After merging this data, we plot the percentage of CVEs exploited from 2018 to 2024, starting from 0 days after their disclosure and extending up to four years (1,460 days) post-disclosure.

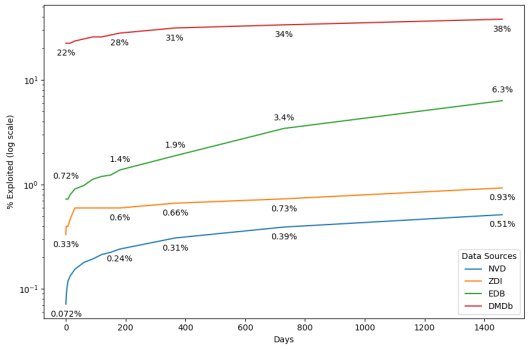

Figure 10: Percentage of CVEs exploited after disclosure.

Figure 10 shows that DMDb has the highest final percentage of exploited CVEs, followed by EDB, ZDI, and NVD. This aligns with the analysis by Almukaynizi et al. (2017), which shows that CVEs disclosed on the dark web are more likely to be exploited. We note that our exploitation percentage of 0.51% for NVD is lower than the 3% observed by Almukaynizi et al. One explanation for this is that CISA under-reports CVE exploitation in KEV, as the catalog only receives CVEs that have remediation guidelines. Despite the potential under-reporting, the exploitation percentage in DMDb remains high. We hypothesize that capturing the hackers' intentions on the dark web, information not available in ZDI or EDB, overcomes the effect of under-reporting. Therefore, DMDb serves as a valuable data source for predicting vulnerability exploitation.

[7]Common Vulnerabilities and Exposures (CVE) - unique and standardized identifier assigned to a specific vulnerability cataloged in the CVE List.

[8]National Vulnerability Database (NVD) - an augmented software vulnerability database built on the CVE List and kept by the U.S. government.

[9]Zero Day Initiative (ZDI) - a research program that focuses on identify and report zero-day (unknown) vulnerabilities.

[10]Exploit Database (EDB) - an archive of proof-of-concept (PoC) exploits and documentation for known vulnerabilities.

[11]Known Exploited Vulnerabilities (KEV) - Cybersecurity and Infrastructure Security Agency's (CISA) catalog of vulnerabilities exploited in attacks.

7. Conclusion

This paper introduces DMDb (DarkMiner Database), a criminal hacking repository of dark web forum and marketplace data fully available to cybersecurity researchers. The database is the result of our efforts to build a comprehensive data collection that illuminates the current plans of cyber-attackers. We achieve this by designing and deploying DarkMiner, a system that regularly scrapes the Tor network to populate DMDb. To enhance the value of the database, we address common repository limitations such as incomplete, unauthorized, and noisy datasets. By offering structured, up-to-date data—including text, images, and object change tracking—DMDb enables in-depth analysis of cybercrime activities involving criminal hacking, as demonstrated in this work through multiple case studies. As it grows, incorporating foreign websites and other dark web networks, DMDb will remain a critical tool for advancing CTI-research.

8. Acknowledgements

This work is supported by the National Science Foundation (NSF) under the grant No 2246220.

References

Almukaynizi, M., Marin, E., Nunes, E., Shakarian, P., Simari, G. I., Kapoor, D., & Siedlecki, T. (2018). Darkmention: A deployed system to predict enterprise-targeted external cyberattacks. *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*.

Almukaynizi, M., Nunes, E., Dharaiya, K., Senguttuvan, M., Shakarian, J., & Shakarian, P. (2017). Proactive identification of exploits in the wild through vulnerability mentions online. *2017 International Conference on Cyber Conflict (CyCon U.S.)*

AZSecure. (2017). Intelligence and Security Informatics Data Sets [https://www.azsecure-data.org/].

Bermudez-Villalva, A., & Stringhini, G. (2021). The shady economy: Understanding the difference in trading activity from underground forums in different layers of the web. *2021 APWG Symposium on Electronic Crime Research*.

Bhalerao, R., Aliapoulios, M., Shumailov, I., Afroz, S., & McCoy, D. (2019). Mapping the underground: Supervised discovery of cybercrime supply chains. *2019 APWG Symposium on Electronic Crime Research*.

Bharathidevi, B., Chennamsetty, L., Prasad, A., & Balijepalli, A. (2017). Logo matching for document image retrieval using sift descriptors. *International Journal of Engineering Research and Applications*, *07*.

Borgeaud, A. (2024). *Dark web intelligence market revenue worldwide from 2022 to 2032* [https://www.statista.com/statistics/1461403/global-dark-web-intelligence-market-size/]. Statista.

Boshmaf, Y., Perera, I., Kumarasinghe, U., Liyanage, S., & Al, H. (2023). Dizzy: Large-scale crawling and analysis of onion services. *Proceedings of the 18th International Conf. on Availability, Reliability and Security*.

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., & Goode, S. (2015, June). Dark net market archives, 2013-2015 [https://gwern.net/dnm-archive/].

Broadhurst, R., Lord, D., Maxim, D., Woodford-Smith, H., Johnston, C., Chung, H., Carroll, S., Trivedi, H., & Sabol, B. (2018). *Malware trends on darknet crypto-markets: Research review*.

Bullough, B. L., Yanchenko, A. K., Smith, C. L., & Zipkin, J. R. (2017). Predicting exploitation of disclosed software vulnerabilities using open-source data. *Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics*.

Campobasso, M., & Allodi, L. (2023). Know your cybercriminal: Evaluating attacker preferences by measuring profile sales on an active, leading criminal market for user impersonation at scale. *32nd USENIX Security Symposium (USENIX Security 23)*.

Cybersecurity Ventures. (2023). *Cybercrime to cost the world $9.5 trillion usd annually in 2024* [https://www.esentire.com/web-native-pages/cybercrime-to-cost-the-world-9-5-trillion-usd-annually-in-2024/]. The 2023 Annual Cybercrime Report.

Cybersixgill. (2024). State of the underground 2024: Combating risepro, lumma, vidar, and other top stealer malware [https://cybersixgill.com/news/articles/combating-stealer-malware].

Dingledine, R., Mathewson, N., & Syverson, P. (2004). Tor: The second-generation onion router. *Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13*, 21.

Dummer, S. (2024). Akamai blocked 419 tb of malicious traffic in a 24-hour ddos attack [https://www.akamai.com/blog/security/akamai-blocked-419-tb-of-malicious-traffic]. *Akamai*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conf on comp. vision and pattern recognition*, 770–778.

Hughes, J., Pastrana, S., Hutchings, A., Afroz, S., Samtani, S., Li, W., & Santana Marin, E. (2024). The art of cybercrime community research. *ACM Comput. Surv.*, *56*(6).

Intelliagg. (2016). *Deeplight: Shining a light on the dark web*. University of Melbourne.

Jeziorowski, S., Ismail, M., & Siraj, A. (2020). Towards image-based dark vendor profiling: An analysis of image metadata and image hashing in dark web marketplaces. *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*.

Kaspersky. (2023). *Kaspersky security bulletin 2023. statistics* [https://securelist.com/ksb-2023-statistics/111156/].

Lowe, D. (1999). Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, *2*.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press (CUP).

Marin, E., Almukaynizi, M., Nunes, E., Shakarian, J., & Shakarian, P. (2018). Predicting hacker adoption on darkweb forums using sequential rule mining. *ISPA/IUCC/BDCloud/SocialCom/SustainCom*.

Marin, E., Almukaynizi, M., Sarkar, S., Nunes, E., Shakarian, J., & Shakarian, P. (2021). *Exploring malicious hacker communities: Toward proactive cyber-defense*. Cambridge University Press (CUP).

Marin, E., Almukaynizi, M., & Shakarian, P. (2019). Reasoning about future cyber-attacks through socio-technical hacking information. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*.

Marin, E., Diab, A., & Shakarian, P. (2016). Product offerings in malicious hacker markets. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*.

Marin, E., Shakarian, J., & Shakarian, P. (2018). Mining key-hackers on darkweb forums. *2018 1st International Conference on Data Intelligence and Security (ICDIS)*.

Meland, P. H., & Sindre, G. (2019). Cyber attacks for sale. *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR, Workshop Track Proceedings*.

Morgan, S. (2019). *Cybercriminal activity is one of the biggest challenges that humanity will face in the next two decades*. Cybersecurity Ventures.

Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., Robertson, J., Shakarian, J., Thart, A., & Shakarian, P. (2016). Darknet and deepnet mining for proactive cybersecurity threat intelligence. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*.

Pastrana, S., Thomas, D. R., Hutchings, A., & Clayton, R. (2018). Crimebb: Enabling cybercrime research on underground forums at scale. *Proceedings of the 2018 World Wide Web Conference*.

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

Robertson, J., Diab, A., Marin, E., Nunes, E., Paliath, V., Shakarian, J., & Shakarian, P. (2017). *Darkweb cyber threat intelligence mining*. Cambridge University Press.

Samtani, S., & Chen, H. (2016). Using social network analysis to identify key hackers for keylogging tools in hacker forums. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*.

Schäfer, M., Fuchs, M., Strohmeier, M., Engel, M., Liechti, M., & Lenders, V. (2019). Blackwidow: Monitoring the dark web for cyber security information. *11th International Conference on Cyber Conflict*, *900*.

Singh, M. (2024). Wazirx halts withdrawals after losing 230 million, nearly half its reserves [https://techcrunch.com/2024/07/18/indias-wazirx-confirms-security-breach-after-230-million-suspicious-transfer/]. *TechCrunch*.

Soska, K., & Christin, N. (2015). Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. *24th USENIX Security Symposium*.

Staff, C. (2024). Cloudflare report reveals q2 2024 ddos attack trends [https://www.cxoinsightme.com/news/cloudflare-report-reveals-q2-2024-ddos-attack-trends/]. *CXO Insight Middle East*.

The Tor Project. (2024). *Tor metrics* [https://metrics.torproject.org/userstats-relay-country.html/].

Wang, X., Peng, P., Wang, C., & Wang, G. (2018). You are your photographs: Detecting multiple identities of vendors in the darknet marketplaces. *Asia Conference on Computer and Communications Security*.