Challenges of Collecting Cyber Threat Intelligence from Hacker Forums

Ashwini Dalvi[1*], Nishita Dhote[2], S G Bhirud[1]
Veermata Jijabai Technological Institute, Mumbai, India
[1] *aadalvi_p19@ce.vjti.ac.in
1sgbhirud@ce.vjti.ac.in
K. J. Somaiya College of Engineering, Mumbai, India
nishita.dhotre@somaiya.edu

----------ABSTRACT----------

Hacker and dark web forums have become popular among cybercriminals, where a range of topics can be discussed freely, including tutorials and materials for learning hacking or even launching a cyber-attack. Researching and monitoring these forums can help cybersecurity professionals gain valuable insights about the trending terms, vulnerabilities, and exploits discussed in dark web forums. Additionally, Hacker forum investigation can help proactively detect cyber-attack and situational awareness.

Hacker forums available on the open, deep, and dark web are potential cyber threat intelligence (CTI) sources. However, the challenges in collecting and analyzing data may make CTI goals and objectives difficult to achieve with hacker forums.The present work discussed the limitations and challenges associated with using data from hacker forums to CTI.

Keywords: **Hacker forum, dark web, cyber threat intelligence, data collection, hacker forum datasets**

1 INTRODUCTION

The cybersecurity community develops cyber threat intelligence (CTI) systems based on publicly available data to combat the rapidly evolving landscape of cyber threats. Using data collected from vendors and research, CTI identifies threat events, threat techniques, and threat tactics.Additionally, gathering information from forums, social media platforms, news websites, hidden services from the dark web, and other related sources may prove helpful in identifying potential threats and their purposes, tools, and methods. Collecting and collating information from different sources to construct threat intelligence demands topic specific attention. For example, by the nature of hacker forums, information collected from it would help comprehend trends of attack vectors like ransomware and malware or discussion on zero-day exploits.

Researchers attempted hacker form investigation with different approaches like performing content analysis on collected text data and attempting potential threat actors from forum data. In addition, limited research attempts discussed comprehension of hacker psychology. The topical discussions among hackers. Using Natural Language Processing (NLP) and text mining, a foresight study can be conducted on hacker forum data. However, a significant processing challenge involves processing the raw data in hacker forums.

2 RELATED WORK

2.1 Hacker Forum Dataset

Representative data samples are required to be analyzed and labeled to train ML models. However, the first limitation in comprehending hacker forum data motivation for forum data research is to collect information proactively, and proactive actions require a capable backbone to convert data into information. One such proactive CTI strategy would be the identification of hackers within online hacking forums.

Services like hackerforums, InternetRelay-Chat (IRC), carding shops, and Dark Net Marketplaces (DNMs) are proliferating with hackers [1]. However, penetrating these services is challenging because hacker communication is concealed behind anti- crawling mechanisms. Still, security professionals attempt to reach hacker forums continuously to gain usable CTI because hackers share direct useable malicious artifacts like exploits and zero-day vulnerabilities on hacker forums. Researchers mentioned collecting information about hacking tools, tutorials to conduct malicious activities, and sharing credit card information and other related goods [2].

Manual analysis of hacker forums is error-prone, resource-intensive, and time- consuming. Therefore, machine learning algorithms are used to automate the search and clustering of hacker forum posts to identify the most relevant comments and estimate the

The scope of the present work is to discuss if security professionals and researchers aim to use hacker forum data, then what potential limitations and challenges they need to comprehend for applying machine learning models.

is labeled data source. Table I summarizes hacker forum data sources and the purpose of forum investigation.

Table 1. I Hacker Forum Data Set

References	Data set	Purpose of study
[3]	GitHub and five forums Ethicalhacker.net Hackthissite.org Offensivecommunity.net Wildersecurity.com Mpgh.net	Comprehension of hackers' online footprints on GitHub and forums
[4]	Nulled hacker forum	Identification of skilled cybercriminals
[5]	Public repositories: Seebug, ExploitDB, Packet Storm, Metasploit, Vulnerlab, Zero science	Hacker exploit source code classification
[6]	53 Dark Web hacking forums retrieved from CYR3CON NVD (CVE, CPE) ExploitDB 230 records from an enterprise log	Prediction of hacker strategies
[7]	CYR3CON Dark Web marketplaces and forums, NVD (CVE, CPE)	Prediction of cyberattacks
[8]	NVD, CVE, ExploitDB, Zero Day Initiative, Sark Web and Deep Web marketplaces and forums	Predicting vulnerabilities that could be exploited mentioned in hacker forums
[9]	Nulled.IO	Detection of cyber threats with SVM and CNN
[10]	Nulled.io	Detecting threats from the posts on the hacker forum
[11]	Dataset provided by the University of Arizona's Artificial Intelligence Lab	Discovery of crucial hackers and threats
[12]	HackHound forum, University of Arizona Hacker database	Analyzing hacker roles and behaviors
[13]	Hacker forums from the CrimeBB dataset from the Cambridge Cybercrime Centre	Identifying threats, key actors, and potential future vital actors, as well as their evolving expertise and interests
[14]	CrimeBB dataset from the Cambridge Cybercrime Centre	Finding trending terms in cybersecurity in longitudinal historical noisy text data of an underground hacking forum

Some of the datasets mentioned for forum investigation are university datasets. Cambridge Cybercrime center does not scrape underground marketplace websites, currently. Nevertheless, earlier scraped data on underground forums are available for researchers to fulfill legal compliance. In addition, researchers can access AZSecure Hacker Assets Portal, which provides researchers with attachments, and source code samples compiled from hacker forums [15]. The Artificial Intelligence Lab at the University of Arizona collected dark web forum data through 2012 for its project on Jihadi social media. The collected data is available on AZSecure Hacker Assets Portal.
Though Cambridge Cybercrime Center and AZSecure Hacker Assets Portal offer forum data, the data is historical.

2.2 ML-enabled hacker forum analysis

Researchers attempted to generate forum data other than proprietary and open- source hacker forum datasets with different means. Researchers used text mining techniques to identify terms that frequently appear on dark web forums [16]. Based on information received from social media and dark net forums, alerts were generated regarding the anticipated threat. The proposed method linked text mining results with a group of words with contextual semantic relationships to interpret the warnings and observe the evolution of activities related to the discovered terms on the dark web. The dataset used in the research was tweeted on Twitter.
Using incremental crawlers with anti-crawling mechanisms, researchers crawled hacker forums and categorized exploits based on a deep learning algorithm to understand exploits [17]. Researchers identified trending terms, exploits, and attachments from hacker forums during the classification phase. Moreover, a visualization dashboard was developed to analyze individual exploit postings and author activities by year and type of exploit.
Researchers employed the LDA topic modeling algorithm on five hacker forums
[18]. The study aimed to comprehend hacker assets and language analysis of source code in the forums.
The research aimed to identify trending cyber security terms based on longitudinal historical noisy text data from an underground hacking forum using weighted log odds ratio methodologyand compare these results to the ones obtained based on TFIDS [19]. Compared to TFIDS, the weighted log odds ratio methodology produced more salient terms.
Researchers developed hacker rank to detect key hackers in underground forums, combining content analysis and social network analysis [20]. An LDA model was used to analyze the topic preferences based on content analysis. Further, a network was constructed that represented relationships among users and helped to identify each user. The final step was applying a page rank algorithm to extract users' rankings, and the users with higher rankings were identified as key hackers. An analysis of the interaction between the users resulted in the construction of a social network graph. The users' activity was evaluated by analyzing the number of postings and replies they made. The results indicated that different hackers had high social network influence and distinctive topic preferences.
In one attempt, researchers attempt unsupervised learning on dark Web forums to identify potential data breaches [21].

Researchers extracted cyber security information from unstructured hacker forums using a hybrid method which combines text tagging, clustering analysis techniques and an improved data processing method.22]
In one study, researchers studied online forums to detect and analyse users posting hyperlinks with malicious intension.[23]

Researchers employed different machine learning techniques to address critical questions from forum discussions, such as key hackers and types of content. Nevertheless, a machine learning model needs to be trained on the relevant data set in order for it to be successful.Machine learning algorithms are unable to perform efficiently without high-quality training data.Researchers presented the web-based tool 'POSTCOG' for hacker forum data exploration [24]. POSTCOG cites hacker forum datasets based on whether the data source is leaked or scraped, also whether the dataset is complete or partial. Following a signing agreement with the Cambridge Cybercrime Centre, POSTCOG becomes available for academic research.

3 CHALLENGES WITH HACKER FORUM DATA COLLECTION

The major challenge of hacker forum investigation is the scarcity of open source,and annotated datasets for training machine learning enabled models. Researchers attempt to create forum datasets by scraping hacker forums. However, most forums have been unavailable, and even available forums have been dormant for the past few years.
Table II shows the availability or accessibility of the hacker forum dataset.

Table 2. II Hacker Forums

References	Forum name	Accessible/ availability	Open Source Data Acces
[25]	Nulled forum	Yes	No
[26]	Hack this site	No	---
[27]	Hidden answers	No	---
[26]	Breach forum	No	---
[27]	Raid	No	No
[28]	Eternia	y	No
[29]	Antionline	y	Log in Required
[30]	crackingzilla	n	---
[31]	Webcracking	n	---
[34]	Hacksden	y	Log in Required

Scraping forums from table II or refereeing content from them could not contribute much to CTI because the quality of collected data from mentioned forums tends to be minimalistic.

Further, challenges with hacker forum data analysis are listed as follows:

∟ Language analysis

The users of the hacker community intentionally use inconsistent language as an anonymity method. Grammatical mistakes, spelling mistakes, and idiomatic content are also intentional while communicating over these forums.

∟ Anti crawling Mechanism

In order to prevent automated, large-scale data collection, Hacker Community platforms employ several anti-crawling measures. A robust anti-crawling mechanism forces researchers to collect data manually.

∟ CAPTCHA enabled login

Dark web crawlers are frequently interrupted by text-based CAPTCHAs, resulting in manual intervention and inhibiting large-scale data collection [35].

CAPTCHAs could be of different formats. For example, figure 1 shows a different form of CAPTCHA to login into the forum.

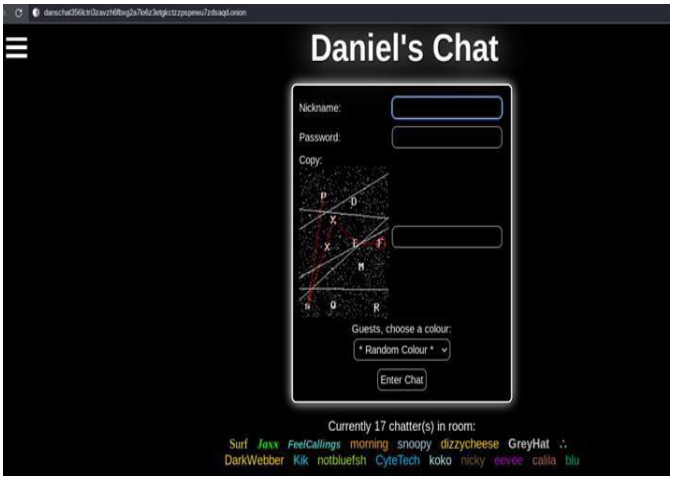

Fig. 1. A different form of CAPTCHA

4 CONCLUSION

CTI enables organizations to identify and address threats that may have affected and are likely to affect them. Providing CTI facilitates the identification and prevention of cyberthreats.

However, data quality, relevance, timeliness, and intelligence are essential in usingdata as CTI. Unfortunately, the current state of hacker forum investigation research lacks an updated backbone to facilitate forum data to be CTI. As a result, it may be challenging to use hacker forums data due to the difficulties in data collection from forums, data labeling, and technology limitations to reach live hacker forums.

REFERENCES:

1. Du, P. Y., Zhang, N., Ebrahimi, M., Samtani, S., Lazarine, B., Arnold, N., & Chen, H. (2018, November). Identifying, collecting, and presenting hacker community data: Forums, IRC, carding shops, and DNMs. In 2018 IEEE international conference on intelligence and security informatics (ISI) (pp. 70-75). IEEE.
2. Ebrahimi, M., Samtani, S., Chai, Y., & Chen, H. (2020, May). Detecting cyber threats in non-English hacker forums: an adversarial cross-lingual knowledge transfer approach. In 2020 IEEE Security and Privacy Workshops (SPW) (pp. 20-26). IEEE.
3. Islam, R., Rokon, M. O. F., Darki, A., &Faloutsos, M. (2021). Hackerscope: The dynamics of a massive hacker online ecosystem. Social Network Analysis and Mining, 11(1), 1-12.
4. Johnsen, J. W., & Franke, K. (2020, November). Identifying proficient cybercriminals through text and network analysis. In 2020 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 1-7). IEEE.
5. Ampel, B., Samtani, S., Zhu, H., Ullman, S., & Chen, H. (2020, November). Labeling hacker exploits for proactive cyber threat intelligence: a deep transfer learning approach. In 2020 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 1-6). IEEE.
6. Marin, E., Almukaynizi, M., &Shakarian, P. (2019, November). Reasoning about future cyber-attacks through socio-technical hacking information. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 157-164). IEEE.
7. Almukaynizi, M., Marin, E., Nunes, E., Shakarian, P., Simari, G. I., Kapoor, D.,&Siedlecki, T. (2018, November). Darkmention: A deployed system to predict enterprise - targeted external cyberattacks. In 2018 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 31-36). IEEE.
8. Almukaynizi, M., Nunes, E., Dharaiya, K., Senguttuvan, M., Shakarian, J., &Shakarian, P. (2017, November). Proactive identification of exploits in the wild through vulnerability mentions online. In 2017 International Conference on Cyber Conflict (CyCon US) (pp. 82-88). IEEE.
9. Deliu, I., Leichter, C., & Franke, K. (2017, December). Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 3648-3656). IEEE.
10. Deliu, I., Leichter, C., & Franke, K. (2018, December). Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent dirichlet allocation. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 5008-5013). IEEE.
11. Zenebe, A., Shumba, M., Carillo, A., & Cuenca, S. (2019). Cyber threat discovery from dark web. EPiC Series in Computing, 64, 174-183.
12. Biswas, B., Mukhopadhyay, A., & Gupta, G. (2018, January). "Leadership in Action: How Top Hackers Behave" A Big-Data Approach with Text-Mining and Sentiment Analysis. In Proceedings of the 51st Hawaii international conference on system sciences.
13. Pastrana, S., Hutchings, A., Caines, A., & Buttery, P. (2018, September). Characterizing eve: Analysing cybercrime actors in a large underground forum. In International symposium on research in attacks, intrusions, and defenses (pp. 207-227). Springer, Cham.
14. https://www.cambridgecybercrime.uk/datasets.html
15. https://www.azsecure-data.org/hacker-assets-portal.html
16. Sapienza, A., Bessi, A., Damodaran, S., Shakarian, P., Lerman, K., & Ferrara, E. (2017, November). Early warnings of cyber threats in online discussions. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 667-674). IEEE.
17. R. Williams, S. Samtani, M. Patton, and H. Chen, "Incremental hacker forum exploit collection and classification for proactive cyber threat intelligence: an exploratory study," in Proceedings of the 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), Miami, FL, USA, November 2018.
18. Samtani, S., Chinn, R., & Chen, H. (2015, May). Exploring hacker assets in underground forums. In 2015 IEEE international conference on intelligence and security informatics (ISI) (pp. 31-36). IEEE.
19. Hughes, J., Aycock, S., Caines, A., Buttery, P., & Hutchings, A. (2020, November). Detecting trending terms in cybersecurity forum discussions. In Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020) (pp. 107-115).
20. Huang, C., Guo, Y., Guo, W., & Li, Y. (2021). HackerRank: identifying key hackers in underground forums. International Journal of Distributed Sensor Networks, 17(5), 15501477211015145.
21. Nazah, S., Huda, S., Abawajy, J. H., & Hassan, M. M. (2021). An Unsupervised Model for Identifying and Characterizing Dark Web Forums. IEEE Access, 9, 112871-112892.
22. Chen, Chia-Mei, et al. "Retrieving Potential Cybersecurity Information from Hacker Forums." Int. J. Netw. Secur 23 (2021): 1126-1138.
23. Islam, Risul, et al. "HyperMan: detecting misbehavior in online forums based on hyperlink posting

behavior." *Social Network Analysis and Mining* 12.1 (2022): 1-14..
24. Pete, I., Hughes, J., Caines, A., Vu, A. V., Gupta, H., Hutchings, A., ...& Buttery, P. (2022, June). PostCog: A tool for interdisciplinary research into underground forums at scale. In 2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) (pp. 93-104). IEEE.
25. https://www.nulled.to/index.php
26. https://www.hackthissite.org/
27. https://onion.live/site/hidden-answers
28. https://breached.to/
29. https://raidforums.com/index.php
30. https://eternia-to-fj.veno2.cn/
31. https://www.antionline.com/
32. http://www.crackingzilla.net
33. https://hackforums.net/showthread.php?tid=2900093
34. https://hackforums.net/
35. Zhang, N., Ebrahimi, M., Li, W., & Chen, H. (2022). Counteracting dark Web text-based CAPTCHA with generative adversarial learning for proactive cyber threat intelligence. ACM Transactions on Management Information Systems (TMIS), 13(2), 1-21.