Big Data and Machine Learning for Sensor Network Security - Research Article

International Journal of
Distributed
Sensor Networks

International Journal of Distributed Sensor Networks
2021, Vol. 17(5)
© The Author(s) 2021
DOI: 10.1177/15501477211015145
journals.sagepub.com/home/dsn
SAGE

HackerRank: Identifying key hackers in underground forums

Cheng Huang[1,2], Yongyan Guo[1], Wenbo Guo[1] and Ying Li[1]

Abstract

With the rapid development of the Internet, cybersecurity situation is becoming more and more complex. At present, surface web and dark web contain numerous underground forums or markets, which play an important role in cybercrime ecosystem. Therefore, cybersecurity researchers usually focus on hacker-centered research on cybercrime, trying to find key hackers and extract credible cyber threat intelligence from them. The data scale of underground forums is tremendous and key hackers only represent a small fraction of underground forum users. It takes a lot of time as well as expertise to manually analyze key hackers. Therefore, it is necessary to propose a method or tool to automatically analyze underground forums and identify key hackers involved. In this work, we present HackerRank, an automatic method for identifying key hackers. HackerRank combines the advantages of content analysis and social network analysis. First, comprehensive evaluations and topic preferences are extracted separately using content analysis. Then, it uses an improved Topic-specific PageRank to combine the results of content analysis with social network analysis. Finally, HackerRank obtains users' ranking, with higher-ranked users being considered as key hackers. To demonstrate the validity of proposed method, we applied HackerRank to five different underground forums separately. Compared to using social network analysis and content analysis alone, HackerRank increases the coverage rate of five underground forums by 3.14% and 16.19% on average. In addition, we performed a manual analysis of identified key hackers. The results prove that the method is effective in identifying key hackers in underground forums.

Keywords

Underground forum, key hacker, content analysis, social network analysis

Date received: 7 September 2020; accepted: 12 April 2021

Handling Editor: Yanjiao Chen

Introduction

In the current cybersecurity situation, it is increasingly difficult to guard against advanced attacks or exploits. Hackers have a lot of funds, superb technology, and rich experience. They could not only improve their attack techniques but also are good at finding the weak point in the real enterprise network, including management and personnel.[1] In the face of such complex network attack and defense status, one way to deal with problems is to identify key hackers and then mine emerging cyber threats.

At present, surface web and dark web contain numerous underground forums or markets, which play an important role in the cybercrime ecosystem.[2] These underground forums are popular places for hackers to conduct activities such as learning, communication for information, vulnerability disclosure, tools exchange, and also a distribution center for cybercrime.[3,4] Many

[1]College of Cybersecurity, Sichuan University, Chengdu, China
[2]Guangxi Key Laboratory of Cryptography and Information Security, Guilin, China

Corresponding author:
Yongyan Guo, College of Cybersecurity, Sichuan University, Chuanda Road, Shuangliu County, Chengdu 610027, China.
Email: guoyongyan1998@gmail.com

forums are also dedicated to providing underground transactions for trading malware, information theft, and other services.[5] Therefore, many cybersecurity researchers focus on hacker-centered research on cybercrime, trying to find key hackers and extract credible cyber threat intelligence from them.[6]

The data scale of underground forums is tremendous and key hackers represent only a small fraction of underground forum users. Identifying key hackers in such a situation is a great challenge. It takes a lot of time as well as expertise to manually analyze these key hackers. Therefore, it is necessary to propose a method or tool to automate the analysis of underground forums and identify key hackers involved.

In existing research, two main methods have been used to identify key hackers in underground forums: content-based analysis[7–9] and social network-based analysis.[10–12] Content-based approaches analyze user data based on selected evaluation metrics, such as activity and content quality. Social network-based approaches build a social network on an underground forum in which key hackers have a high degree of network centrality, with common approaches including degree centrality, eigenvector centrality, and PageRank. In general, content analysis (CA) is relatively comprehensive but complex. Social network analysis (SNA) can directly reflect the posting frequency and relationship of users. It is more objective but ignores users' attribute information.

In this work, we present HackerRank (HR), an automatic method for identifying key hackers. HR combines the advantages of CA and SNA. First, evaluation metrics of underground forum users are computed to generate a comprehensive evaluation. Second, topic analysis of the data generated by users is performed to obtain their topic preferences. Finally, an improved Topic-specific PageRank algorithm is used to fuse the comprehensive evaluation and topic preferences for SNA to obtain a ranking of users, with higher-ranked users being considered as key hackers. To demonstrate the validity of our method, we applied HR to different underground forums separately, comparing it with the method using CA or SNA alone. Besides, we performed a manual analysis of identified key hackers. The results prove that our method is effective in identifying key hackers in underground forums.

The specific contributions of this work are the following:

- This article proposes a framework for automatically analyzing key hackers in underground forums. HR can automatically collect data from underground forums and analyze key hackers among them.
- Key hacker identification combines methods based on CA and SNA. This method first extracts the user's comprehensive evaluation metrics and topic preferences based on CA and then applies our improved Topic-specific PageRank for SNA.
- In order to verify the effectiveness and portability of HR, we conducted experiments on five popular underground forums, and the results showed that the user coverage was higher than only using CA or SNA.

The rest of this article is organized as follows. Section "Related work" presents related work. Section "Methodology" details the implementation process of the HR framework. Section "Experiments" presents the experiments and analyses. Section "Conclusion" summarizes the conclusion and proposes future works.

Related work

We review existing works from two perspectives, including research on underground forums and key hacker identification. Key hacker identification is a branch of research on underground forums.

Research on underground forums

Due to the increasing link between underground forums and cybercrime, researchers have conducted many studies on underground forums. Related research includes the identification of underground forums, extracting cyber threat intelligence, hacker assets, and so on. Du et al.[13] proposed a method for systematically identifying and automatically collecting a large-scale of underground forums, carding shops, Internet Relay Chat (IRC), and Dark Net Marketplaces. Samtani et al.[14,15] analyzed hacking assets within underground forums that can identify the tools which may be used in a cyberattack, provide knowledge on how to implement and use such assets. They developed AZSecure Hacker Assets Portal, which uses the latest machine learning technology to collect and analyze malicious assets from online hacker communities. Deliu et al.[16] explored the potential of machine learning methods to rapidly sift through underground forums for relevant cyber threat intelligence using text data from real underground forums. Benjamin et al.[17] combined machine learning methods with information retrieval techniques to build an automated method for identifying tangible and verifiable evidence of potential threats within underground forums, IRC channels, and carding shops.

Key hacker identification

Existing methods for identifying key hackers fall into two main categories: content-based and social network-based analysis.

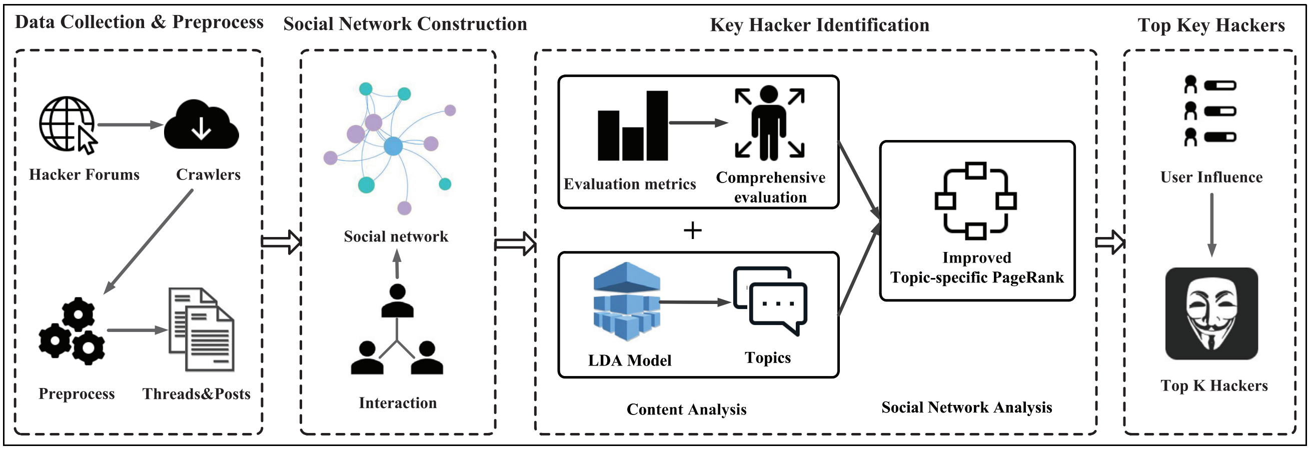


Figure 1. The framework of HackerRank.

Users of underground forums generate a lot of data, such as created threads, posts, comments, and uploaded attachments. Content-based analysis refers to mining these data[18–20] and constructing user evaluation metrics to discover key users among them. Common evaluation metrics include activity level, content quality, and so on. Different studies have chosen different evaluation metrics. For example, Marin et al.[7] analyzed content features, seniority features, and social network features among underground forums. They used an optimization meta-heuristic to identify key hackers and proposed a systematic method based on reputation to validate the results. Fang et al.[8] developed a framework with a set of topic models for extracting popular topics, tracking topic evolution, and identifying key hackers with their specialties. They identified key hackers in each expertise area by utilizing Latent Dirichlet Allocation (LDA), Dynamic Topic Model, and Author Topic Model. Zhang et al.[9] analyzed the knowledge transfer of user posts in underground forums and classified users into four types: expert, casual, learning, and novice hackers. Expert hackers act as key knowledgeable and respectable members in the communities, increasingly acting as knowledge providers. Content-based analysis builds metrics that directly reflect the influence of users by mining user data from underground forums. Although content-based analysis is very comprehensive, it is more complicated and the selection of evaluation metrics requires professional participation and verification.

In contrast to content-based analysis, social network-based analysis focuses on user interactions in underground forums.[21–23] User behavior in underground forums is used to construct a social network graph, which is then used to identify key users using graph-based analysis.[24,25] In general, key hackers have high network centralities, such as degree centrality, eigenvector centrality, and PageRank. Pete et al.[26] utilized network centrality analysis to highlight the structural patterns of each network to identify important nodes and key hackers. Zhang et al.[10] proposed a new heterogeneous information network (HIN) embedding model named ActorHin2Vec to learn the low-dimensional representations for the nodes in HIN, and then a classifier was built for key actor identification. Grisham et al.[11] used a state-of-the-art neural network architecture model to identify mobile malware attachments and then social network-based analysis techniques to determine key hackers disseminating mobile malware. Samtani and Chen[12] analyzed user interactions by leveraging metrics such as network diameter and average path length, and quantified the importance of each user using centrality measures. Social network-based analysis is common across different social platforms but ignores information about the attributes specific to underground forum users. Different from these above works, we combine the advantages of content-based and social network-based analysis to build a framework for automated analysis of key hackers in underground forums.

Methodology

In this section, we describe HR in detail, a framework for automatically analyzing key hackers in the underground forums. The high-level design of HR is illustrated in Figure 1. *Data Collection and Preprocess* collects the content of the underground forums and preprocesses the collected data. *Social Network Construction* generates a social network graph based on the interaction among users. *Key Hacker Identification* combines analysis based on content and social network. Content-based analysis constructs a comprehensive evaluation based on the user characteristics of underground forums and analyzes the users' topic preferences based on the LDA model. Then, we perform SNA through the improved Topic-specific PageRank algorithm based on the results of CA and generate users' influence. Finally, we get the Top K key hackers from the ranking based on their user influence.

Data collection and preprocess

In this section, we collect the content from underground forums and users' interaction. In underground forums, discussions are all organized as threads (i.e. a user initiates a thread and create a post, then other users reply it, discussing various hacker-related information posted by community members). While crawling the data of forums, we also collect them like this. In other words, we get all the threads from the forum first, and then we collect all the posts under the thread, including the username, profile, content, order, and time of the post. In addition, we also consider some mechanisms to deal with the anti-crawler mechanisms of the underground forums.

As for the crawled raw data, the data are not well-formatted. In order to perform the text analysis better, we conduct data preprocessing here. First, we convert all the data to lowercase to keep the data format consistent. Second, we delete non-ASCII characters and punctuation marks. Finally, we use the natural language toolkit (NLTK)[27] module to segment the text and delete the stop words. Also, word lemmatization is necessary here.

Social network construction

SNA studies the relationship between social entities based on graph structure. In a graph, there are two components: nodes and edges. Here, the nodes represent the user of underground forum, and the edges represent the social relationships among users.

The social network graph is displayed in Figure 2. We define the graph as a directed weighted graph $G=(V,E)$, where G represents a weighted directed graph, V represents a vertex set, and E is the edge set. In underground forums, each user in the underground forum represents a vertex $v_i \in V$. If $<v_i, v_j> \in E$, it means that there is an interactive relationship between user v_i and user v_j. The weight W of the edge is the number of interactions between users. For example, in Figure 2, there is an edge weight of $<v_A, v_D>$ in user A and user D with w_{AD}, which means that user A has replied to user D's post with a frequency w_{AD}. What should be noted here is that the thread initiator initiates a thread, and other users discuss it in this thread in underground forums. By default, other users' replies are for the thread initiator, and the connection should be established with the thread initiator. However, there are also some situations that users discuss with others directly in the thread. In this condition, the connection should be established according to the reply object specified by the user.

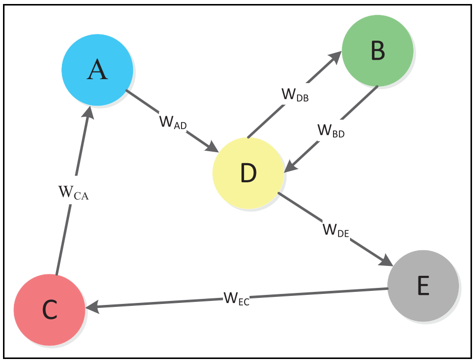


Figure 2. Social network graph.

Key hacker identification

User evaluation metrics construction. In order to dig out the relevant features and behaviors of key hackers, there have been various works to explore and study the users' characteristics of underground forums or online forums. As shown in Table 1, we summarize the common features. The related works mainly portray users from three aspects, including activity, content quality, and knowledge dissemination ability. Activity is reflected by the number of posts, the more active the user, the more the number of replies and threads in the forums. Users with high-quality speeches have longer posts, and also involve a lot of hacker jargons, technical jargons, and threat intelligence. In addition, users' interaction is usually along with knowledge transfer (knowledge acquisition and provision), and key hackers are often the core of knowledge transfer.

Based on the previous works,[8,9,18,19,28–31] we construct a user evaluation metric system based on CA, and extract some features from the collected data as users' evaluation metric. According to the characteristics of entropy, calculating the entropy value could evaluate the randomness and disorder of an event, or the degree of dispersion for some metric. The more discrete the metric, the greater the influence (weight) of the metric on the comprehensive evaluation. Therefore, we adopt entropy weight method[32] to assign weights to various metric to generate a comprehensive evaluation for each user. The calculation process is as follows:

- Data standardization: as illustrated in equation (1), we use minimum and maximum method to standardize the data since the measurement units of various indicators are not uniform, and the data dimensions and data levels are quite different. In equation (1), x_{ij} represents the jth metric of the ith user, $maxx_j$ is the maximum value of

Table 1. Content analysis metrics.

Category	Feature	Description
Activity	Start topics[8,28]	Total number of topics created by the hacker
	Start replies[8,28]	Total number of replies created by the hacker
Content quality	Length of topics[19,29]	The average length of the thread created by the hacker (i.e. the number of words contained)
	Length of replies[19,29]	The average length of the replies created by the hacker (i.e. the number of words contained)
	Length difference[18]	The ratio of the length of the reply post to the length of the topic post
	Technical jargon[18]	Count of technical terms included in the post such as computer and program
	Hacker jargon[30]	Count of posts including hacker jargons such as Attack, penetration, XSS, and SQL inject
	IOC share[31]	The number of IOCs included in the post, which indicates that hackers may participate in cybercrime or share resources, including IP, Hash, domain name, and so on
Knowledge dissemination ability	Replies with knowledge provision[9]	The number of knowledge-providing keywords contained in the reply post, such as answers, guide recommend, and follow
	Replies with knowledge acquisition[9]	The number of knowledge acquisition keywords contained in the reply post such as request, need, and doubt
	Topics with knowledge provision[9]	The number of knowledge provision keywords contained in the thread
	Topics with knowledge acquisition[9]	The number of knowledge acquisition keywords contained in the thread

the jth metric, and $minx_j$ is the minimum value

$$x_{ij} = \frac{x_{ij} - minx_j}{maxx_j - minx_j} \tag{1}$$

- Calculate the information entropy of the jth metric

$$e_j = -k \sum_{i=1}^{n} p_{ij} ln(p_{ij}) \tag{2}$$

where $k = 1/ln(n)$ and $p_{ij} = x_{ij} / \sum_{i=1}^{n} x_{ij}$.

- Calculate the weight of each metric

$$w_j = \frac{1 - e_j}{\sum_{j=1}^{m} (1 - e_j)} \tag{3}$$

where m is the count of metrics.

- Perform a weighted summation of the weights of each metric to generate a comprehensive evaluation of underground forum users as

$$U_i = \sum_{j=1}^{m} x_{ij} \cdot w_j \tag{4}$$

LDA-based underground forum topic discovery. In this section, we build a topic discovery model to analyze users' topic preferences. We use the LDA algorithm for topic modeling, which is actually a three-layer Bayesian probability model containing words, document structure, and topics.[33] If a document is considered as a set of word vectors, then for a document, the document and topic satisfy a polynomial distribution, and the words in the topic and vocabulary also satisfy a polynomial distribution. The two polynomial distributions are both Dirichlet distribution with hyperparameters α and β. As for the document, we just consider whether a word appears, rather than the order of its occurrence. In LDA model, a document is generated as Figure 3, and the process is as follows:

- Take samples from the Dirichlet distribution α to generate the topic distribution θ_i of document i.
- Take samples from the topic polynomial distribution θ_i to generate the topic $z_{i,j}$ of the jth word for document i.
- Take samples from the Dirichlet distribution β to generate the word distribution $\varphi_{z_{i,j}}$ of the topic $z_{i,j}$.
- Take samples from the words polynomial distribution $\varphi_{z_{i,j}}$ and finally generating words $w_{i,j}$.

In underground forums, users usually post more than once. In order to understand the user's topic preference, we group one's all posts into a document d. Through LDA, we could get the probability distribution of words on the topic (equation (5)), the probability distribution of the article on the topic (equation

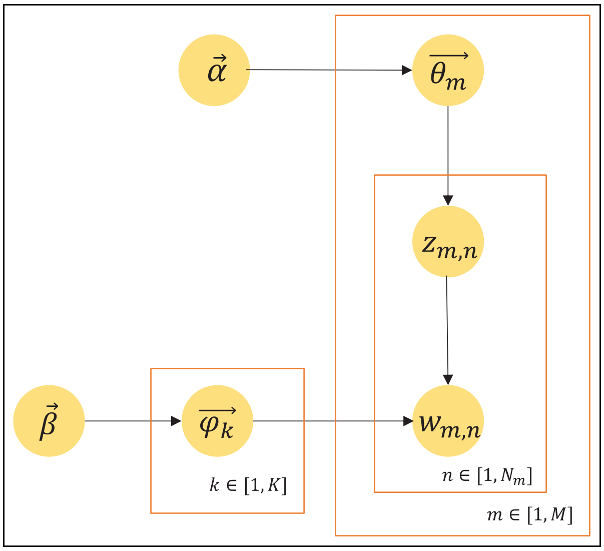


Figure 3. A document's generation in LDA model.

(6)), where C_{wk} represents the times that the word w is assigned to the topic K

$$p(k|w) = \frac{C_{wk} + \beta}{\sum_{k=1}^{K} C_{wk} + K\beta} \quad (5)$$

$$p(k|d) = \frac{C_{dk} + \beta}{\sum_{k=1}^{K} C_{dk} + K\beta} \quad (6)$$

To train LDA model, the election of the number of topics is essential. At present, perplexity and coherence are often used to determine the number of topics. Perplexity means that "for a document, how uncertain the LDA model is that it belongs to a some topic." The more topics, the lower the perplexity,[33] but the model is more likely to be over-fitting. So, when understanding the approximate range of the number of topics from the perplexity, coherence[34] can be used to select more suitable topics from this range. The calculation of perplexity is illustrated as follows

$$perplexity = exp - \frac{\sum_{d=1}^{M} logp(w_d)}{\sum_{d=1}^{M} N_d} \quad (7)$$

where M represents the count of documents in text sets, N_d is the length of document m, and $p(w_d)$ represents the probability of text.

The coherence can be calculated as follows

$$coherence = \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{D(w_i, w_j) + 1}{D(w_j)} \quad (8)$$

where D_{w_j} is the document frequency of word w_j, and $D(w_i, w_j)$ represents the co-document frequency of word w_i and w_j.[35]

We choose the best number of topics to train the LDA model through the comprehensive assessment of coherence and perplexity.

SNA based on improved Topic-specific PageRank. In sections "User evaluation metrics construction" and "LDA-based underground forum topic discovery," we construct user comprehensive evaluation metrics and topic preferences based on CA. In this section, our algorithm is improved from the Topic-specific PageRank algorithm.[36] In our method, we combine the results of the above CA for SNA. Then, we obtain the final user influence value, the HR value.

According to the social network diagram constructed in section "Social network construction," its weight is the number of interactions between users. Since the user's influence is different, we need to consider the asymmetric delivery of each node (user). Here, we define the weight of the edge in the social network graph as equation (9)

$$N_{ij} = U_j \cdot w_{ij} \quad (9)$$

where U_j is the comprehensive evaluation based on user j's activity, posts content quality, and knowledge dissemination ability. w_{ij} is the interaction frequency between user i and user j.

Next, we construct a transition matrix; the transition of user's state (i.e. the user will communicate with which user next time) is related to the current state, but not the past state. For user j, each user i pointed to by the outgoing link has $M_{ij} = N_{ij} / \sum_k N_{jk}$. Each user has the probability of α to communicate with other users next time. At this time, the users rank can be presented as equation (10)

$$Rank = (1 - \alpha)M * Rank + \alpha\vec{v} \quad (10)$$

Based on the LDA topic discovery model mentioned in section "LDA-based underground forum topic discovery," in HR, we first use a series of topics to generate the topic vector $\vec{v}$ ($\vec{v}$ is used to record the relationship among all users and topics, each topic maintains a $\vec{v}$ vector). Let K_j be the user set in a topic T_j, then when calculating the PageRank vector of topic T_j, replace the uniform damping vector $p = [1/n]_{n \times 1}$

$$v = \begin{cases} \frac{1}{|K_j|} & i \in K_j \\ 0 & i \notin K_j \end{cases} \tag{11}$$

As mentioned above, we have generated a set of topic-specific Rank vectors, which could basically measure the user's influence in each topic. In addition, we refer to the approach of Weng et al.[37] to get the overall influence of users, and calculate a weight r_t for the ranking under topic t. Besides, we need to build a matrix WT with dimension $W \times T$, where W is the word frequency of a topic, and T is the count of topics. WT_{ij} represents the times that the word w_i is assigned to the topic t_j. The following formula is used to calculate r_t

$$r_t = \frac{\sum_{i=1}^{W} T_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{T} WT_{ij}} \tag{12}$$

In summary, the calculation of the user's overall influence is shown in equation (13)

$$HackerRank = \sum_{t} r_t \cdot Rank \tag{13}$$

Experiments

Data sets

In this study, we conduct experiments through five different mainstream underground forums. According to the data collection methods described in section "Data collection and preprocess," the crawler is designed and developed. Since each forum has a different structure, we adapt it on each forum. The data set is shown in Table 2. In addition to the data we collected, the "Nulled" forum also contains the data leaked in 2016.

Analysis of LDA experimental results

In the process of key hacker identification, we choose LDA topic model to extract users' topic preferences. Instead of training the LDA model for each underground forum separately, we use all the data in Table 2 to train a general model suitable for underground forum topic analysis. During the training of the LDA model, choosing an appropriate topic number has a great influence on the model. In this article, coherence and perplexity are the indicators we choose to evaluate the performance of the model. In the experiment, the topic number is set to 2–10 (interval 1) and 15–50 (interval 5). Figures 4 and 5 show the curve of coherence and perplexity under different topic numbers, and in Figure 5, when the number of topics ranges from 2 to 10 (step = 1), the change in perplexity is on the upper right.

Table 2. Underground forum data sets.

Forum	Threads	Posts	Users
Nulled	52,707	230,934	89,671
HackThisSite	1827	9407	2022
HiddenAnswers	19,950	62,706	11,814
BreachForum	2018	8660	1233
Raid	362	4319	1722

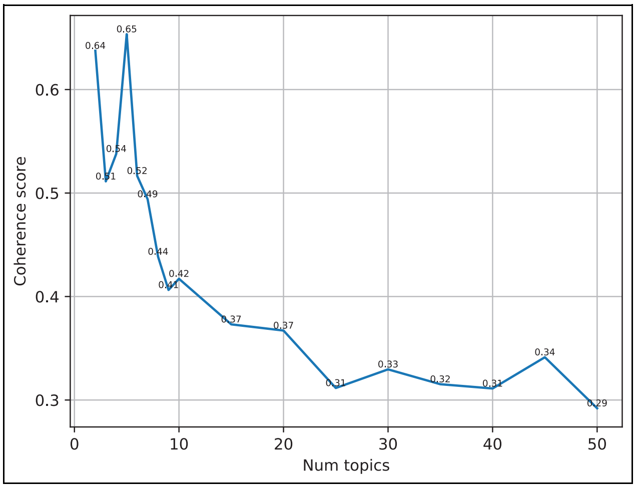


Figure 4. LDA model's coherence of different number of topics.

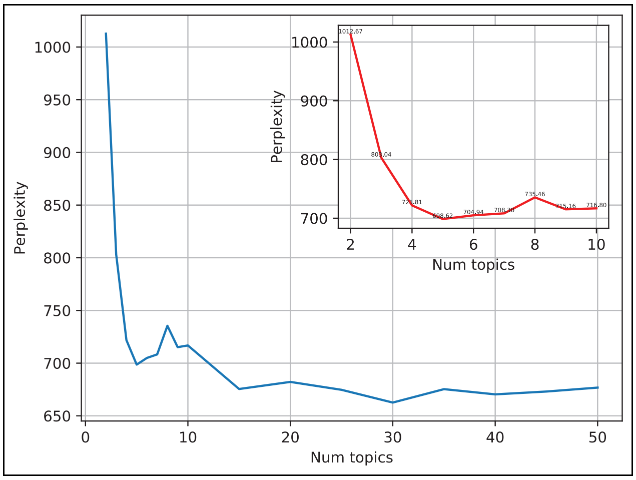


Figure 5. LDA model's perplexity of different number of topics.

In Figure 4, when the number of topics is 5, coherence reaches the maximum value, and the number of topics ranges from 5 to 50, the value of coherence decreases as a whole. As can be seen in Figure 5, the number of topics ranges from 2 to 5, and the perplexity shows a downward trend. When the number of topics

Table 3. Five topics generated by the LDA model and their representative words.

Topic	Words
Mobile security	android, apk, zombie, mobile, sdk
Data leakage	price, icq, balance, info, account
Hacking tutorial	tool, source, tutorial, executables, shared
System vulnerability	worm, dll, antivirus, spybot, hijack
Network security	com, http, www, php, ssh

Table 4. Correlation between rank lists by different methods.

Forum	HR vs CA	HR vs SNA
Nulled	−0.0922	0.0465
HackThisSite	−0.2244	0.0824
HiddenAnswers	0.0122	0.0644
BreachForum	−0.0151	0.1311
Raid	−0.0294	0.0422

HR: HackerRank; CA: content analysis; SNA: social network analysis.

goes from 6 to 8, the perplexity increases slightly; the number of topics ranges from 15 to 50 (interval 5), the perplexity is stable at 670 to 720, and the trend of change is relatively gentle. Although the perplexity is not the minimum when the number of topics $K = 5$, it is already a local minimum, and when the number of topics increases, the trend of perplexity is very small. Combining the results of Figures 4 and 5, we choose the number of topics $K = 5$.

Using the trained LDA topic model, we extract the five most representative words under each topic. As shown in Table 3, we summarize the topic name and representative words of each topic.

Effect of HR

Comparison with related algorithms. To validate HR, we set up comparison experiments. HR combines CA and SNA, so we compare methods that use CA or SNA alone.

- CA: users are ranked according to their comprehensive evaluation in section "User evaluation metrics construction."
- SNA: users are ranked according to their PageRank value.

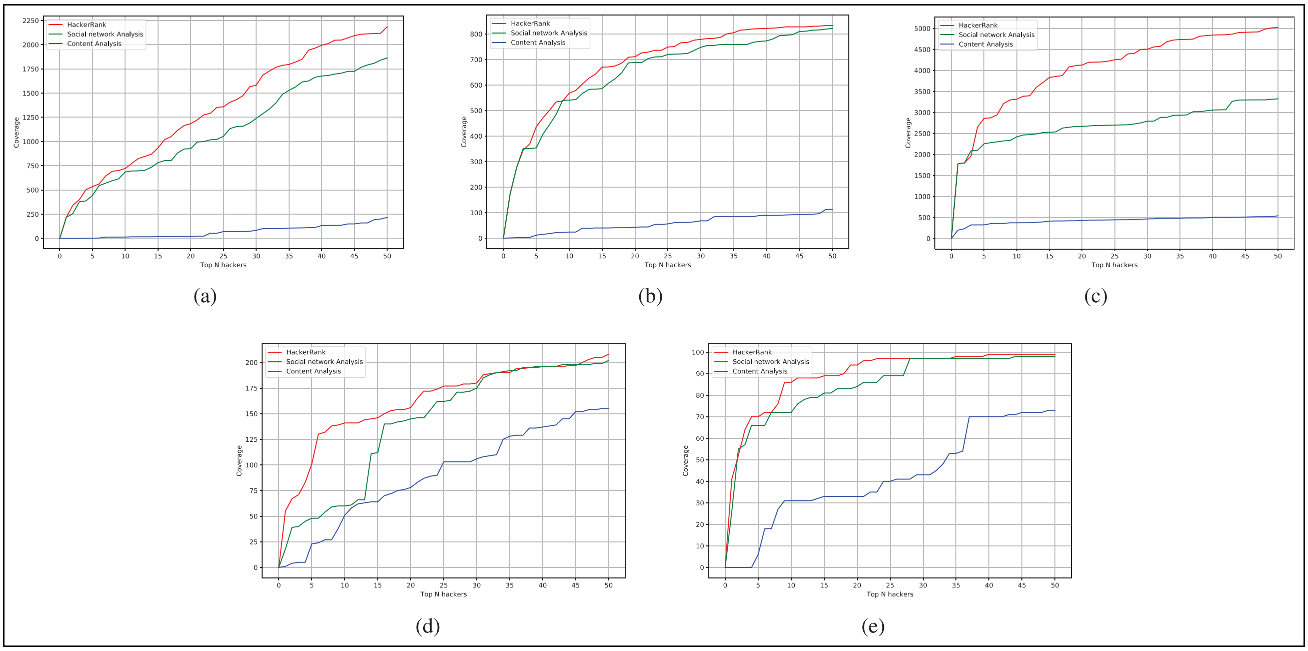
In the above methods, the damping factor has a large effect on PageRank and HR, which is a balancing parameter between the effectiveness of the algorithm and the speed of convergence. In the experiment, the damping factor is set to 0.85, which is an empirical value. With a damping factor of 0.85, it can converge to the PageRank vector in about 100 iterations. When the damping factor is close to 1, the number of iterations required will increase abruptly, and the sorting will be unstable.

Kendall correlation. Kendall correlation is used to measure the correlation between two random variables. The value of Kendall correlation τ ranges from -1 to 1. Two sequences are exactly the same when $\tau = 1$. Two sequences are opposite when $\tau = -1$. The greater τ is, the higher correlation between two sequences. In this section, we analyze the correlation between HR and rank lists generated by CA and SNA through Kendall correlation. We find the same trend in different underground forums. As shown in the Kendall correlation in Table 4, HR has a difference in the rank list generated by other methods. At the same time, it can be observed that the correlation τ of HR versus SNA is higher than HR versus CA. This is because different methods use different characteristics and analysis methods to evaluate user influence.

Coverage analysis. To validate the effectiveness of HR, we evaluate HR using coverage,[38,39] which is commonly used in the field of key user identification, as an evaluation metric. Coverage measures the effectiveness of key user identification from the network topology formed by user interactions, by counting the number of affected users.

This article compares the coverage of three methods on underground forum top 50 key hackers. To fully validate the performance of HR, the experiments are conducted on five different underground forums. As shown in Figure 6, HR's coverage of top 50 hackers in all five underground forums is higher than that using SNA or CA alone. Specifically, compared to using SNA and CA alone, HR has increased the coverage rate (*coverage number/total number of forum users*) of five underground forums by 3.14% and 16.19% on average, which proves the validity and portability of our method. It can be seen from Figure 6 that the HR coverage curve increases rapidly from 1 to 20, and then the growth rate slows down. The top 20 hackers have correlated most of the users in the forum, which shows that in underground forums, a small fraction of key hackers has high influence. In addition, it can be observed that the effect of only using CA is poor. This is due to the fact that CA only considers the text features of users but ignores the interaction among users.

Key hacker identification results. In this section, we show the top five key hackers for each forum obtained using HR, SNA, and CA, as shown in Table 5. It can be seen that the results obtained by the different methods have some similarities as well as some differences.



(a) (b) (c)

(d) (e)

Figure 6. Top 50 key hackers' coverage of five underground forums (a) Nulled, (b) HackThisSite, (c) HiddenAnswers, (d) Raid, and (e) BreachForum.

Table 5. Top five key hackers in each forum.

Method	Rank	Nulled	HackThisSite	HiddenAnswers	Raid	BreachForum
HR	1	Zaida	Goatboy	anonymous	BoringApe	Daemon
	2	Veterun	limdis	Man	bitsandbytes	Rape
	3	Psych0path	thedotmaster	ForTheLuks	mantext	Kevin
	4	K33P0	WallShadow	v0h20	rwkregime	KANANSTARKS
	5	Nord	godofcereal	jeliavlov	$2a$45	Syrup
SNA	1	Zaida	Goatboy	anonymous	bitsandbytes	Kevin
	2	Nord	limdis	Man	$2a$45	Daemon
	3	Veterun	godofcereal	jeliavlov	MrSimpleA	LSDoom
	4	CaptainSperg	nuclearhaxor	pio123498765	a115567926	Rape
	5	Psych0path	moonloghtkiller	ForTheLuks	Already	Gucci
CA	1	coolmodee	Goatboy	anonymous	mantext	shopsocks5.com
	2	lasqal	tremor77	v0h20	bgn91	vn5socks.net
	3	khalidf1	WallShadow	Overfl0w	rwkregime	dichvusocks
	4	johnluke	limdis	ArronOfTor	Based_Stick141	tisocks
	5	MatheuZ	-Ninjex-	WickedX	meggiescouldm	Spastic

HR: HackerRank; CA: content analysis; SNA: social network analysis.

In order to better verify the effectiveness of HR, we manually checked the above results. Taking the Nulled forum as an example, Table 6 shows the top five key hackers and the results of the three analysis methods. Here, we analyze the top five key hackers. "Zaida" hacks into a large number of accounts (such as mailboxes) and sells them publicly in the forum, attracting a large number of buyers to conduct transactions.

Table 6. Nulled forum top five key hacker analysis results.

Rank	Username	HR	SNA	CA
1	Zaida	0.01456	0.00414	59.996
2	Veterun	0.01333	0.00257	47.7261
3	Psych0path	0.01085	0.00234	57.0223
4	K33P0	0.00972	0.00232	31.2062
5	Nord	0.00523	0.00337	29.0523

"Veterun" often publishes high-quality hacking tutorials in the forum and shares related hacking resource links. At the same time, he also conducts in-depth technical exchanges with other users in the forum. "Psych0path" is engaged in software cracking and private data transactions. It has completed up to 880 transactions in the forum and has a high reputation. "K33P0" is very active under themes such as games (such as CSGO) and digital currencies (such as BTC, ETH, and LTC). "Nord" focuses on program cracking and participates in activation key trading activities, and has released many illegally obtained program keys. It can be seen from the above analysis that key hackers not only have high social network influence but also the content they publish also has high-quality and distinctive topic preferences. Therefore, HR can more accurately identify key hackers based on CA and SNA.

Conclusion

In this article, we propose a key hacker identification framework for underground forums, HR. This framework combines CA and SNA. First, we mine the user characteristics of underground forums and construct a comprehensive evaluation. Second, the LDA model is used to predict users' topic preferences. In SNA, user influence is obtained using an improved Topic-specific PageRank algorithm based on comprehensive evaluations and topic preferences. Through user influence ranking, we can identify key hackers in underground forums. In our experiments, we compare HR with methods that use CA or SNA alone. The results prove that HR has a significant advantage in identifying key hackers. At present, HR can identify key hackers based on historical data of underground forums but lacks consideration of forum evolution. Also, HR can only identify key hackers in a single forum. In the future, we will work on building a real-time key hacker identification framework based on dynamic graphs and study the identity linkage across different forums.

Acknowledgements

The authors thank anonymous reviewers and editors for provided helpful comments on earlier drafts of the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported in part by National Natural Science Foundation of China (no. 61902265), the Sichuan Science and Technology Program (nos 2019YFG0407 and 2020YFG0047), the Guangxi Key Laboratory of Cryptography and Information Security (no. GCIS201921), and the Fundamental Research Funds for the Central Universities.

ORCID iDs

Cheng Huang https://orcid.org/0000-0002-5871-946X
Yongyan Guo https://orcid.org/0000-0001-6623-7201

References

1. Tounsi W and Rais H. A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Comput Secur* 2018; 72: 212–233.
2. Thomas K, Huang D, Wang D, et al. Framing dependencies introduced by underground commoditization. In: *Proceedings of the workshop on economics of information security 2015*, https://cseweb.ucsd.edu/~savage/papers/WEIS15.pdf
3. Algarni A and Malaiya Y. Software vulnerability markets: discoverers and buyers. *Int J Comput Inf Sci Eng* 2014; 8(3): 71–81.
4. Macdonald M, Frank R, Mei J, et al. Identifying digital threats in a hacker web forum. In: *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining*, Paris, 25–28 August 2015, pp.926–933. New York: ACM.
5. Yue WT, Wang QH and Hui KL. See no evil, hear no evil? Dissecting the impact of online hacker forums. *MIS Quart* 2019; 43(1): 73–95.
6. Nunes E, Diab A, Gunn A, et al. Darknet and deepnet mining for proactive cybersecurity threat intelligence. In: *2016 IEEE conference on intelligence and security informatics (ISI)*, Tucson, AZ, 28–30 September 2016, pp.7–12. New York: IEEE.
7. Marin E, Shakarian J and Shakarian P. Mining key-hackers on darkweb forums. In: *2018 1st international conference on data intelligence and security (ICDIS)*, South Padre Island, TX, 8–10 April 2018, pp.73–80. New York: IEEE.
8. Fang Z, Zhao X, Wei Q, et al. Exploring key hackers and cybersecurity threats in Chinese hacker communities. In: *2016 IEEE conference on intelligence and security informatics (ISI)*, Tucson, AZ, 28–30 September 2016 IEEE, pp.13–18. New York: IEEE.
9. Zhang X, Tsang A, Yue WT, et al. The classification of hackers by knowledge exchange behaviors. *Inform Syst Front* 2015; 17(6): 1239–1251.
10. Zhang Y, Fan Y, Ye Y, et al. Kadetector: automatic identification of key actors in online hack forums based on structured heterogeneous information network. In: *2018 IEEE international conference on big knowledge (ICBK)*, Singapore, 17–18 November 2018, pp.154–161. New York: IEEE.
11. Grisham J, Samtani S, Patton M, et al. 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), Beijing, China, 22–24 July 2017, pp.13–18. New York: IEEE.

12. Samtani S and Chen H. Using social network analysis to identify key hackers for keylogging tools in hacker forums. In: *2016 IEEE conference on intelligence and security informatics (ISI)*, Tucson, AZ, 28–30 September 2016, pp.319–321. New York: IEEE.
13. Du PY, Zhang N, Ebrahimi M, et al. Identifying, collecting, and presenting hacker community data: forums, IRC, carding shops, and DNMs. In: *2018 IEEE international conference on intelligence and security informatics (ISI)*, Miami, FL, 9–11 November 2018, pp.70–75. New York: IEEE.
14. Samtani S, Chinn K, Larson C, et al. AZSecure hacker assets portal: cyber threat intelligence and malware analysis. In: *2016 IEEE conference on intelligence and security informatics (ISI)*, Tucson, AZ, 28–30 September 2016, pp.19–24. New York: IEEE.
15. Samtani S, Chinn R and Chen H. Exploring hacker assets in underground forums. In: *2015 IEEE international conference on intelligence and security informatics (ISI)*, Baltimore, MD, 27–29 May 2015, pp.31–36. New York: IEEE.
16. Deliu I, Leichter C and Franke K. Extracting cyber threat intelligence from hacker forums: support vector machines versus convolutional neural networks. In: *2017 IEEE international conference on big data (Big Data)*, Boston, MA, 11–14 December 2017, pp.3648–3656. New York: IEEE.
17. Benjamin V, Li W, Holt T, et al. Exploring threats and vulnerabilities in hacker web: forums, IRC and carding shops. In: *2015 IEEE international conference on intelligence and security informatics (ISI)*, Baltimore, MD, 27–29 May 2015, pp.85–90. New York: IEEE.
18. Abbasi A, Li W, Benjamin V, et al. Descriptive analytics: examining expert hackers in web forums. In: *2014 IEEE joint intelligence and security informatics conference*, The Hague, 24–26 September 2014, pp.56–63. New York: IEEE.
19. Benjamin V and Chen H. Securing cyberspace: identifying key actors in hacker communities. In: *2012 IEEE international conference on intelligence and security informatics*, Washington, DC, 11–14 June 2012, pp.24–29. New York: IEEE.
20. Kigerl A. Behind the scenes of the underworld: hierarchical clustering of two leaked carding forum databases. *Soc Sci Comput Rev*. Epub ahead of print 5 June 2020. DOI: 10.1177/0894439320924735.
21. Li W, Chen H and Nunamaker JF Jr. Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. *J Manage Inform Syst* 2016; 33(4): 1059–1086.
22. Zhang X and Li C. Survival analysis on hacker forums. In: *SIGBPS workshop on business processes and service*, Milan, Italy, 15 December 2013, pp.106–110. Citeseer. Berlin, German: Springer.
23. Lu Y, Luo X, Polgar M, et al. Social network analysis of a criminal hacker community. *J Comput Inform Syst* 2010; 51(2): 31–41.
24. Sarvari H, Abozinadah E, Mbaziira A, et al. Constructing and analyzing criminal networks. In: *2014 IEEE security and privacy workshops*, 17–18 May 2014, San Jose, CA, pp.84–91. New York: IEEE.
25. Zhang Y, Fan Y, Ye Y, et al. Key player identification in underground forums over attributed heterogeneous information network embedding framework. In: *Proceedings of the 28th ACM international conference on information and knowledge management*, Beijing, China, 3–7 November 2019, pp.549–558. New York: ACM.
26. Pete I, Hughes J, Chua YT, et al. A social network analysis and comparison of six dark web forums. In: *2020 IEEE European symposium on security and privacy workshops (EuroS&PW)*, Genoa, 7–11 September 2020, pp.484–493. New York: IEEE.
27. Loper E and Bird S. Nltk: the natural language toolkit. *arXiv preprint cs/* 0205028, 2002.
28. Pal A and Counts S. Identifying topical authorities in microblogs. In: *Proceedings of the fourth ACM international conference on web search and data mining*, Hong Kong, 9–12 February 2011, pp.45–54. New York: ACM.
29. Agichtein E, Castillo C, Donato D, et al. Finding high-quality content in social media. In: *Proceedings of the 2008 international conference on web search and data mining*, Palo Alto, CA, 11–12 February 2008, pp.183–194. New York: ACM.
30. Huang SY and Ban T. A topic-based unsupervised learning approach for online underground market exploration. In: *2019 18th IEEE international conference on trust, security and privacy in computing and communications/13th IEEE international conference on big data science and engineering (TrustCom/BigDataSE)*, Rotorua, New Zealand, 5–8 August 2019, pp.208–215. New York: IEEE.
31. Liao X, Yuan K, Wang X, et al. Acing the IOC game: toward automatic discovery and analysis of open-source cyber threat intelligence. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, Vienna, Austria, 24–28 October 2016, pp.755–766. New York: ACM.
32. Cruz JD, Bothorel C and Poulet F. Entropy based community detection in augmented social networks. In: *2011 international conference on computational aspects of social networks (CASoN)*, Salamanca, 19–21 October 2011, pp.163–168. New York: IEEE.
33. Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res* 2003; 3: 993–1022.
34. Newman D, Lau JH, Grieser K, et al. Automatic evaluation of topic coherence. In: *Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics*, Los Angeles, California, 2–4 June 2010, pp.100–108. Stroudsburg, PA: ACL.
35. Mimno D, Wallach H, Talley E, et al. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, 27–31 July 2011, pp.262–272. Stroudsburg, PA: ACL.
36. Haveliwala TH. Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. *IEEE T Knowl Data En* 2003; 15(4): 784–796.
37. Weng J, Lim EP, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers. In: *Proceedings of the third ACM international conference on web search and data mining*, 3–6 February 2010, pp.261–270. New York: ACM.

38. Miao Q, Zhang S, Meng Y, et al. Domain-sensitive opinion leader mining from online review communities. In: *Proceedings of the 22nd international conference on world wide web*, Rio de Janeiro, Brazil, 13–17 May 2013, pp.187–188. New York: ACM.

39. Song X, Chi Y, Hino K, et al. Identifying opinion leaders in the blogosphere. In: *Proceedings of the sixteenth ACM conference on information and knowledge management*, Lisbon, 6–10 November 2007, pp.971–974. New York: ACM.