HackerRank: Identifying key hackers in underground forums

Cheng Huang¹,², Yongyan Guo¹,², Wenbo Guo¹ and Ying Li¹

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

¹College of Cybersecurity, Sichuan University, Chengdu, China
²Guangxi 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. Therefore, many cybersecurity researchers 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 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 and social network-based analysis. 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. 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. analyzed hacking assets within underground forums that can identify the tools which may be used in a cyber-attack, 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. 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. 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.
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 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. 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. 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. 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. 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. In general, key hackers have high network centralities, such as degree centrality, eigenvector centrality, and PageRank. Pete et al. utilized network centrality analysis to highlight the structural patterns of each network to identify important nodes and key hackers. Zhang et al. 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. 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 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

![Figure 1. The framework of HackerRank.](image)
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) 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.

**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, 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 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 \( j \)th metric of the \( i \)th user, \( max_j \) is the maximum value of
the jth metric, and \( \min x_j \) is the minimum value

\[
x_{ij} = \frac{x_{ij} - \min x_j}{\max x_j - \min x_j}
\]

(1)

- Calculate the information entropy of the jth metric

\[
e_j = -k \sum_{i=1}^{n} p_{ij} \ln(p_{ij})
\]

(2)

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

- Calculate the weight of each metric

\[
w_j = \frac{1 - e_j}{\sum_{j=1}^{m} (1 - e_j)}
\]

(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

\[
U_i = \sum_{j=1}^{m} x_{ij} \cdot w_j
\]

(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 \( \alpha \) and \( \beta \). 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 \( \alpha \) to generate the topic distribution \( u_i \) of document \( i \).
- Take samples from the topic polynomial distribution \( u_{zi,j} \) to generate the topic \( z_{i,j} \) of the jth word for document \( i \).
- Take samples from the Dirichlet distribution \( \beta \) to generate the word distribution \( \varphi_{zi,i} \) of the topic \( z_{i,j} \).
- Take samples from the words polynomial distribution \( \varphi_{zi,j} \) and finally generating words \( w_{zi,j} \).

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 (6)), and the topic distribution of the document (equation (7)).
(6), where \(C_{wk}\) represents the times that the word \(w\) is assigned to the topic \(K\)

\[
p(k|w) = \frac{C_{wk} + \beta}{\sum_{k=1}^{K} C_{wk} + K\beta}
\]

(5)

\[
p(k|d) = \frac{C_{dk} + \beta}{\sum_{k=1}^{K} C_{dk} + K\beta}
\]

(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 “for a document, how uncertain the LDA model is that it belongs to 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

\[
\text{perplexity} = \exp - \frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}
\]

(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

\[
\text{coherence} = \sum_{i=2}^{N} \frac{\log D(w_i, w_j) + 1}{D(w_j)}
\]

(8)

where \(D_w\) 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)

\[
N_{ij} = U_j \cdot w_{ij}
\]

(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 \(\alpha\) to communicate with other users next time. At this time, the users rank can be presented as equation (10)

\[
\text{Rank} = (1 - \alpha)M \ast \text{Rank} + \alpha \vec{v}
\]

(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 \(p = [1/n_n \times 1]\)
\[ v = \begin{cases} \frac{1}{|K_j|} & i \in K_j \\ 0 & i \notin K_j \end{cases} \quad (11) \]

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.\cite{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 \( W_T \) with dimension \( W \times T \), where \( W \) is the word frequency of a topic, and \( T \) is the count of topics. \( W_{Tj} \) represents the times that the word \( w_i \) is assigned to the topic \( t_j \). The following formula is used to calculate \( r_t \)

\[ r_t = \frac{\sum_{i=1}^{W} T_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{T} W_{Tij}} \quad (12) \]

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

\[ \text{HackerRank} = \sum_t r_t \cdot \text{Rank} \quad (13) \]

**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.

![Figure 4. LDA model's coherence of different number of topics.](image)

![Figure 5. LDA model's perplexity of different number of topics.](image)

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...
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 \( \tau \) ranges from \(-1\) to 1. Two sequences are exactly the same when \( \tau = 1 \). Two sequences are opposite when \( \tau = -1 \). The greater \( \tau \) 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 \( \tau \) 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, 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.

---

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

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

**Table 4.** Correlation between rank lists by different methods.

<table>
<thead>
<tr>
<th>Forum</th>
<th>HR vs CA</th>
<th>HR vs SNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nulled</td>
<td>-0.0922</td>
<td>0.0465</td>
</tr>
<tr>
<td>HackThisSite</td>
<td>-0.2244</td>
<td>0.0824</td>
</tr>
<tr>
<td>HiddenAnswers</td>
<td>0.0122</td>
<td>0.0644</td>
</tr>
<tr>
<td>BreachForum</td>
<td>-0.0151</td>
<td>0.1311</td>
</tr>
<tr>
<td>Raid</td>
<td>-0.0294</td>
<td>0.0422</td>
</tr>
</tbody>
</table>

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 5.** Top five key hackers in each forum.

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank</th>
<th>Nulled</th>
<th>HackThisSite</th>
<th>HiddenAnswers</th>
<th>Raid</th>
<th>BreachForum</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>1</td>
<td>Zaida</td>
<td>Goatboy</td>
<td>anonymous</td>
<td>BoringApe</td>
<td>Daemon</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Veterun</td>
<td>limdis</td>
<td>Man</td>
<td>bitsandbytes</td>
<td>Rape</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Psych0path</td>
<td>thedotmaster</td>
<td>ForTheLuks</td>
<td>mantext</td>
<td>Kevin</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>K33P0</td>
<td>WallShadow</td>
<td>voh20</td>
<td>rwkregime</td>
<td>KANANSTARKS</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Nord</td>
<td>godocereal</td>
<td>jeliavlov</td>
<td>$2a$45</td>
<td>Syrup</td>
</tr>
<tr>
<td>SNA</td>
<td>1</td>
<td>Zaida</td>
<td>Goatboy</td>
<td>anonymous</td>
<td>bitsandbytes</td>
<td>Kevin</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Nord</td>
<td>limdis</td>
<td>Man</td>
<td>$2a$45</td>
<td>Daemon</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Veterun</td>
<td>godocereal</td>
<td>jeliavlov</td>
<td>MrSimpleA</td>
<td>LSDoom</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>CaptainSperg</td>
<td>nuclearhaxor</td>
<td>ForTheLuks</td>
<td>Already</td>
<td>Gucci</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Psych0path</td>
<td>moonlightkiller</td>
<td>ForTheLuks</td>
<td>Already</td>
<td>Gucci</td>
</tr>
<tr>
<td>CA</td>
<td>1</td>
<td>coolmodee</td>
<td>Goatboy</td>
<td>anonymous</td>
<td>mantext</td>
<td>shopsocks5.com</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>lsqal</td>
<td>tremor77</td>
<td>voh20</td>
<td>bgn91</td>
<td>vn5socks.net</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>khalidF</td>
<td>WallShadow</td>
<td>Overf10w</td>
<td>rwkregime</td>
<td>dichvuousocks</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>johnluke</td>
<td>limdis</td>
<td>ArronOTTor</td>
<td>Based Stick141</td>
<td>tisocks</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>MatheuZ</td>
<td>-Ninjex-</td>
<td>WickedX</td>
<td>meggiescouldm</td>
<td>Spastic</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Username</th>
<th>HR</th>
<th>SNA</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Zaida</td>
<td>0.01456</td>
<td>0.00414</td>
<td>59.996</td>
</tr>
<tr>
<td>2</td>
<td>Veterun</td>
<td>0.01333</td>
<td>0.00257</td>
<td>47.7261</td>
</tr>
<tr>
<td>3</td>
<td>Psych0path</td>
<td>0.01085</td>
<td>0.00234</td>
<td>57.0223</td>
</tr>
<tr>
<td>4</td>
<td>K33P0</td>
<td>0.00972</td>
<td>0.00232</td>
<td>31.2062</td>
</tr>
<tr>
<td>5</td>
<td>Nord</td>
<td>0.00523</td>
<td>0.00337</td>
<td>29.0523</td>
</tr>
</tbody>
</table>
“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. “Psychopath” 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
