Identifying Proficient Cybercriminals Through Text and Network Analysis

Jan William Johnsen
Dep. of Information Security and Communication Technology
Norwegian University of Science and Technology
Gjøvik, Norway
jan.w.johnsen@ieee.org

Katrin Franke
Dep. of Information Security and Communication Technology
Norwegian University of Science and Technology
Gjøvik, Norway
kyfranke@ieee.org

***Abstract*—A few highly skilled cybercriminals run the Crime as a Service business model. These expert hackers provide entry-level criminals with tools that allow them to enhance their cybercrime operations significantly. Thus, effectively and efficiently disrupting highly proficient cybercriminals is of a high priority to law enforcement. Such individuals can be found in vast underground forums, though it is particularly challenging to identify and profile individual users. We tackle this problem by combining two analysis methods: text analysis with Latent Dirichlet Allocation (LDA) and Social Network Analysis with centrality measures. In this paper, we use LDA to eliminate around 79% of hacker forum users with very low to no technical skills, while also inferring the forum roles held by the remaining users. Furthermore, we use centrality measures to identify users with hugely popular public posts, including users with very few public posts who receive much attention from their peers. We study various preprocessing methods, wherein we achieve our results by following a series of rigorous preprocessing steps. Our proposed method works towards overcoming current challenges in identifying and interrupting highly proficient cybercriminals.**

***Index Terms*—Topic modelling, Latent Dirichlet Allocation, Social Network Analysis, Network centrality, Underground forum, Crime as a Service, Digital forensics.**

I. Introduction

A few very skilled cybercriminals sell their criminal spoils and technical knowledge to the larger underground market population, through the Crime as a Service (CaaS) business model [1]–[5]. Such criminal activities contribute to the estimated cybercrime cost of (at least) 45 billion US dollars in 2018 [6]. Focusing on identifying and taking down a few proficient criminal actors has desirable benefits [1], [7] such as causing more substantial disruption on the CaaS business model. Therefore, both researchers and practitioners alike are developing methods to identify those proficient actors.

In this paper, we study the combination of methodologies from Natural Language Processing (NLP) and Social Network Analysis (SNA) to identify the more prominent and popular users in such underground forum marketplaces, more specifically: Nulled.io hacker forum. By utilising complementary methods, we explore *what* users are talking about and with *whom* they communicate. Our novel approach can exclude around 79% of underground market users who demonstrate low to no technical skills and remove them from further analysis. This reduction in users increases investigation efficiency and effectiveness by reducing algorithms' execution time and focuses the analysis on prominent cybercriminals.

To achieve a reduction in users, we understand that similar users tend to use combinations of equivalent words when they write posts on the forum. Thus, we can exclude users who exclusively express gratitude towards others when we analyse all forum posts made by individual users and assign specific topics to each user. Removing lower-skilled users allows further analysis steps to focus on skilled cybercriminals. Furthermore, we can infer users' role on the forum, because user groups (such as administrators and reverse engineers) use different assortments of words.

We can also identify users with popular public posts using centrality measures; not only in terms of their connectivity but also those users with few public posts that generate a high intercommunication in the forum. A challenge with underground forum marketplaces is that they are imbalanced datasets where a few administrators and skilled users serve many thousands of users. It is, therefore, necessary to follow a series of rigorous steps to achieve our results. In this paper, we present this series of rigorous preprocessing steps to reduce the number of users to investigate in an underground forum.

II. Previous Work

We refer to our previous article [8] for an overview of related work concerning SNA and network centrality measures, as we focus this section on topic modelling. The topic modelling algorithm Latent Dirichlet Allocation (LDA) ability to find unobserved groups (i.e. identify latent topics) makes it applicable to many areas. LDA is typically used to get an overview over a large corpus of text, but can also be used to identifying key actors and hacker assets in underground forums. For example, Porter [9] utilised LDA to find keywords and trends over a period in darknet markets subreddits.

Although it is challenging to investigate underground forums due to the combination of public, restricted and private sections, Motoyam et al. [10] found that a user's reputations come from being publicly active. Their findings motivated our

The research leading to these results has received funding from the Research Council of Norway programme IKTPLUSS, under the R&D project 'Ars Forensica - Computational Forensics for Large-scale Fraud Detection, Crime Investigation & Prevention', grant agreement 248094/O70.

approach to looking at publicly available posts to identify highly proficient actors. In contrast, Marin et al. [4] used a reputation system in the underground forum to validated their results. They showed that hybridisation of features could identify key hackers more precisely, which we also suggest by combining multiple methods to explain their independent results better.

Researchers [2], [11], [12] have examined various features to identify expert hackers, such as hacker assets (number of attachments), speciality lexicons (vocabulary of a person) and forum involvements (metrics such as number of threads, posts and attachments). They found that older forum members and very active members typically have a higher reputation than other users. Pastrana et al. [12] applied a combination of SNA, topic modelling and clustering to identify features to understand better who is at risk of becoming involved in criminal activities.

Features such as keywords have been used by Benjamin et al. [13]–[15] to explore and understand hacker language and to identify keywords for potential threats. While researchers like Li et al. [16]–[18] have tried to use sentiment analysis (interpret positive, negative and neutral emotions in the text) to identify and profile top malware and carding sellers. They also mentioned that active hackers comprise of those who are more actively involved in hacker community discussions.

Samtani et al. [19]–[21] have utilised SNA techniques to identify key hackers and explore Cyber Threat Intelligence (CTI) and hacker assets. More specifically, Samtani et al. [19] looked at particular classes of networks (bipartite and monopartite) and limited key hacker identification using only betweenness centrality measure. Furthermore, they also focused their research [20], [21] on thread starters, utilising LDA to understand the topic characteristics of hacker assets and identify hacker tools, such as crypters, keyloggers, web and database exploits. In contrast, Nunes et al. [22] tried to find zero-day exploits and vulnerabilities.

Marin et al. [23] used clustering to identify hacker product categories and found that many (nearly) identical items are posted across multiple marketplaces, sometimes under the same vendor username. In comparison, Huang and Chen [24] used clustering to find key members and their roles in the cyber fraud value chain. They used SNA to identify communities and assume key members generally post more content and receive more replies compared to other members.

A challenge with previous research is that many of them gather data through (web-)crawling underground forums. This approach closely mimics real law forensic investigations, but it encounters the same problems from anti-crawling techniques, and only parts of the underground forum may be accessible. Pastrana et al. [5] rectified this problem of relying on incomplete or outdated datasets by capturing data from underground forums resulting in a dataset called CrimeBB.[1] Our access to a leaked hacker forum database has two unique opportunities: i) we can evaluate the performance of our proposed method on the best-case scenario and ii) we have some type of ground truth to base our evaluation with access to all data.

Other issues are that they only look at very tiny fractions of a network (e.g. thread starters or bipartite networks) or rely on knowledge from cybersecurity experts, knowledge which may not generalise to other underground forums. Additionally, many of them assume that higher reputations can indicate proficiency, which can be artificially increased by being a very active member or accumulate reputation over time. For example, forum administrators must be quite active to manage their forums, and they do not necessarily have to possess technical knowledge except for running a website. Thus, we disagree that proficient users are those who are extroverts and communicates a lot in the forum. Proficient hackers can also be introverts which only make a few posts with a high impact on the forum.

[1]Available after legal agreement, which has not been pursued in this moment of time.

III. Methods and Material

Nulled is a hacker forum found in the deep web, that facilitate the brokering of compromised passwords, stolen bitcoins and other sensitive data. We chose Nulled as we believe they closely resemble criminal underground forums, and we had a unique opportunity accessing this leaked database. We have access to all data contained in the database, which allow us to have some type of ground truth to verify our results. The Nulled dataset contains details about 599 085 user accounts, 800 593 private messages and 3 495 596 public messages.

Fig. 1 depicts the overall process of preprocessing and analysing our dataset. We incorporate many standard text preprocessing practices found in NLP and related literature. We also introduce some measures of our own because some type of data was producing noise in the analysis. These measures are specific to this type of dataset, which includes removing e-mail and password combinations and removing Uniform Resource Locators (URLs), HyperText Markup Language (HTML) and BBcode tags. Where related research only employ stemming/lemmatisation to normalise words in their data, we actively attempt to normalise it further. We normalised the text by extracting words with repeating patterns and replacing them with their intended word – this section details these preprocessing and analysis steps.

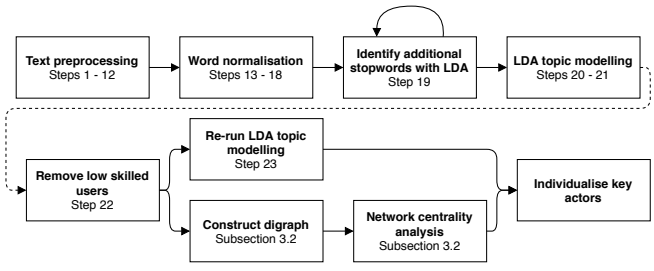

Fig. 1. Process model

A. Latent Dirichlet Allocation

We start this section with a general introduction to the LDA algorithm. Then, the following two subsections will describe: i) our LDA preprocessing steps on forum posts in more detail and ii) generate an LDA model to find new words to filter out.

LDA is one of the more popular algorithms in NLP because it is typically more effective and generalises better than other algorithms. LDA [25], [26] is a statistical model, commonly used to categorise a set of observations (i.e. text) into unobserved groups that explain why some parts of the data are similar. The result is a set of human-interpretative topics from a document corpus. The generalisability property is particularly beneficial, so our proposed method may generalise to more specific domains such as those of underground forums.

LDA is a way of 'soft clustering' using a set of documents and a pre-defined k number of topics. The two other hyperparameters α and η adjust two Dirichlet distributions. These Dirichlet distributions adjust the LDA model document-topic density and topic-word density, respectively. Thus, LDA allows for a nuanced way of categorising documents, as each document has some probability of belonging to several topics. The biggest problem with LDA is the lack of extracting semantically meaningful information [27]. However, a human analyst can deduce what the relation of the topics by studying the word-distributions.

The challenges of analysing a dataset such as Nulled includes how frequent users write with a spoken language. This means text tends to include more repetitions, incomplete sentences, slang expressions (such as 'gonna') and using repeating words/characters to add emphasis. They also tend to write short messages, such as a simple 'thanks' to express gratitude or appreciation. Finally, most users are non-native English speakers and sometimes use non-English words or frequently misspell words; either because they do not know how to spell certain words or they simply ignore misspellings.

1) Latent Dirichlet Allocation Dataset Preprocessing: We followed standard topic modelling preparation steps, while also adding a few of our own to accommodate for this type of dataset. We ensured to replace anything that was removed with whitespace to guarantee that words are not unintentionally combined. The following list is the order of our initial preparation steps:

1) Remove extra spaces and convert to lowercase.
2) Remove all URLs.
3) Remove all 'e-mail:password'-combinations.
4) Remove all e-mails.
   - Leaking of credentials is one primary focus area of this hacker forum. Consequently, it contains large dumps of e-mail and passwords which dominated (esp. mail hosting domains) the topics.
5) Remove all HTML tags (including its contents).
6) Remove all HTML entities (e.g. ' ').
7) Remove newline/tabulator characters ('\n', '\r', and '\t').
8) Remove all BBcode tags.
   - HTML tags and entities and BBcode tags are particular for this type of dataset. We removed them as they were of little use; however, it could be useful if one wants to preserve e.g. attack indicators.
9) Remove symbols, including numbers.
10) Run lemmatization.
    - We chose lemmatization instead of stemming because lemmatization considers the morphological analysis of words. In other words, it considers the structure and part of words to find their root form.
11) Remove stop words.
    - Removing over 700 of the most common English stop words, such as: able, come, do and during.
12) Remove any extra white space.

Users on this hacker forum frequently used exaggeration and abbreviations when writing. They show this by repeating characters (e.g. 'niceeee' or 'goooood') and words (e.g. 'tytyty', short for 'thanks'). Normalising this data could allow us to distinguish between low and high skilled cybercriminals more accurately. Repetition of characters and words, including word misspellings, is an obstacle for LDA [22] – as word variations exist as separate words during the analysis. Lemmatization mitigates many issues of word variations because it converts inflectional forms of words such as 'studying' and 'studies' to the base form of 'study'. However, lemmatisation fails to fix the issue word variations from misspelling and repetition.

To solve the challenge with word variations from repeating characters and words, we extracted and replaced those repeating patterns in two different processes. The first part extracts whole words by looking at repeating patterns while minimising the chance of replacing words erroneously. An example is to avoid illogical changes such as 'r*emem*ber' to 'rember'.

After extracting words with repeating characters or words, we found their shortest expressable form by allowing repetitions to occur a maximum of two times. For example, 'goooood' would have the short form 'good' and 'tytyty' would have the short form 'tyty'. This transformation allowed us to gather repeating patterns of varying length into a collective short form. Finally, the extraction part grouped common short forms and sorted them in descending order of frequency.

13) Extract all words with repeating characters.
14) Extract all words with repeating words.
15) Identify the short form of all extracted words.
16) Group and count the short form words and sort it by frequency in descending order.
17) Inspect and identify replacement words for the first 1000 words.
18) Replace those 1000 new short words with the original words in the dataset.

Our approach to finding the shortest common words made it easier and more effective to replace words with similar repeating characteristics. We manually inspected and changed the first 1000 shortest common words to ensure data quality and control in our experiment. Manual examination allowed us to avoid changing words erroneously. We would either i)

replace the short word with the intended word when it was apparent or ii) replace it with itself if the intended word was unknown or it was already an existing word.

Finding replacement words for the 1000 most frequent short words allowed us to replace 73% of the words found in that list (around 17% of words in the original dataset). The second part of the process is to use those 1000 replacement words to change the original words identified in the first part. Additionally, we replaced many common synonyms and abbreviations without any repeating patterns, for example, 'ty', 'thx', 'thnx', 'merci', and 'gracias' was replaced with 'thanks'.

2) Running the Latent Dirichlet Allocation Analysis: The previous subsection explained more general preprocessing done to forum posts. However, running the LDA algorithm on this dataset can still produce words that do not provide any significant meaning to our analysis. For example, words like 'haha', 'asp', 'tk', 'content' and 'href' are words without any meaning. To further ensure data quality in our experiment, we had to run the LDA algorithm a few times to identify these words to remove them from the final analysis.

The LDA model is very affected by the document input construction. We identified two distinct document construction approaches: i) one document is a concatenation of all messages from a single user or ii) each message is a single document. We call these approaches concatenated and singular, respectively. We keep the hyper-parameters k, α and η identical between each construction approach when running the LDA analyses.

We analysed both approaches and concluded that the singular approach had topics with more mixed words and therefore provided less coherent topics for a human analyst. Therefore, we focus the result section on showing the outcome from the concatenated approach. However, the singular approach was suitable when individualising users, as it gave LDA more documents to learn the underlying topics.

19) Run LDA analysis a few times (four times in our experiment) in order to identify additional stop words and remove them from future analyses.

After these steps, we had a suitable LDA model that can identify which topics users were primarily sending messages about. The identification process resulted in a floating-point array for every user. Each element in the array shows the similarity between the user's messages and every topic. Array elements are the sum of similarities and averaged by the number of sent messages. We also convert this array to a binary format, where topic(s) with any positive float value receives a one, and the other topics are set to zero. In other words, the threshold is anything above zero.

We continue by labelling each topic in the LDA model to find the category that best explains those word combinations. We assume that cybercriminals with low technical skills are very dependent on others with higher kills. They will, therefore, more likely, be consumers of information rather than producers of it. This should come from the way they communicate, such as expressing relatively more gratitude in their public posts. Thus, we can distinguish between users who are pure consumers and express gratitude with everyone else.

20) Categorising LDA model topics, distinguishing between the expression of gratitude versus reverse engineering.
21) Identify which topics each user mainly post messages.
22) Distinguish between users who purely express gratitude with everyone else.
23) Re-run LDA and network centrality analysis using the remaining users that are of interest.

Notably, the result of LDA should be improved when lesser skilled forums users are removed from the dataset, as this gets rid of much junk. Network centrality analysis could also benefit from this, possibly by highlighting different key actors. The main benefit for network centrality measures is fewer forum users to go through, instead of wasting time considering lesser skilled individuals. Thus, the result should be attained faster and feasible for investigators finding secondary targets to take down.

B. Centrality Measures

Network centrality measures are graph-based analysis methods found in SNA, used to identify important and influential individuals within a network. However, public forum posts do not have any natural way of constructing a directed graph (digraph). As the digraph construction will affect the centrality analysis results, we need to decide on how to best model the interaction between users. For example, should edges' direction go out from the thread starter or in towards them? Constructing accurate graphs from these forums are non-trivial, yet essential, to avoid meaningless centrality measures and attribute incorrect significance to users [2].

We denote a set of users V and a set of posts E, as the vertices and edges in a digraph $G = (V, E)$. We chose to construct digraphs with edges from a replying user to the author of forum threads. More specifically, there is a direct edge (v, v'), when user v reply to a thread started by v'. This edge represents an interest to respond on a public thread. We acknowledge this construction method does not truly reflect how forum users interact with other users, as forum threads can be used with multiple purposes, such as asking other users for advice or having unrelated discussions.

We evaluate five popular centrality measures for digraphs: in-degree (C_{deg^-}), out-degree (C_{deg^+}), betweenness (C_B), closeness (C_C) and eigenvector (C_E). They differ in their interpretation of what it means to be 'important' in a network. Thus, some vertices in a network will be ranked as more important than others, as vertices and edges affect the centrality value. We chose these centrality measures as they are in popular forensic investigation tools such as IBM i2 Analyst's Notebook. We refer the reader to SNA books, e.g. McCulloh et al. [28], for formulas and detailed explanations of centrality measures.

IV. Experiment and Results

We have two goals with the experiment: first, to distinguish the majority from the minority (i.e. consumers of content with those who produce it) and secondly, to find out better which individuals to focus investigations resources.

A. Latent Dirichlet Allocation Topic Results

We identified two distinct LDA document construction approaches. The first approach concatenated all messages from individual users into one document, while the second approach treated each message as a document. Due to the page limits, we are only showing the results for the concatenated approach.

We observe that the concatenated approach in Table I gives more coherent groups of words, compared to the singular approach. For example, topic 1 talks about various popular games (possibly sharing of e-mail/username and passwords to get access to these games); topic 2 confirms that something is working (possibly cracks); topic 3 captures various hacker tools, such as Remote Access Trojan (RAT), crypter (encrypt, obfuscate and manipulate malware, to make it harder to detect by security programs) and stealer (theft of some type of information); and topic 7 and 10 express some appreciation of someone's work or thank them for sharing. The singular approach was producing less intelligible results because it has a broader mix of words per topics. For example, appreciation words were distributed among several topics.

TABLE I
CONCATENATED TOPICS FOR ALL USERS

Topic #	Keywords
Topic 1	game, origin, email, sims, capture, key, battlefield, edition, password, country, unit, username, type, fifa, command
Topic 2	work, download, account, crack, post, file, game, time, link, help, find, update, bot, free, check, script, guy, people
Topic 3	stealer, rat, crypter, tool, phisher, scan, binder, beta, spam, user, lsie, module, power, password, ddos, public
Topic 4	script, bol, update, download, legend, enemy, work, champion, bot, version, auto, game, target, login, vip, combo
Topic 5	account, pm, sell, buy, bump, email, paypal, skype, skin, vouch, price, member, password, ban, level, information
Topic 6	add, attack, bot, troop, clashbot, play, password, base, download, set, version, bln, update, pro, feature, option
Topic 7	nice, good, work, man, share, brother, test, love, thank, hope, job, check, mate, wow, dude, lol, great, awesome
Topic 8	password, lol, xxx, minecraft, dragon, account, thank, class, brazzers, alex, fish, mofos, major, profile, cre, david
Topic 9	site, project, user, lol, password, smtp, unranked, round, username, location, try, game, modifier, key, kid, type
Topic 10	thank, test, nice, brother, work, lol, man, good, account, please, much, very, check, rt, wow, rep, dude, game, help

For each forum user, we run their messages through the LDA model (Table I) and output similarity for which of the ten topics they are most similar, as detailed in Section III. Since this similarity is a binary value, we can exclusively distinguish users into two distinct groups: assumed high skill users and low skilled users with only appreciation posts.

There are a total of 299 719 unique users on this forum that had made at least one public post. Table II shows that the concatenating approach categorised 24% of them as assumed high-skilled users. This approach probably achieved fewer skilled users because it could group words used in similar situations more appropriately than the singular approach. Thus, the concatenated approach is the best technique to employ for forensic analysts as it reduces the amount of users most.

TABLE II
NUMBER OF USERS IN HIGH- AND LOW-SKILL GROUPS

Digraph	Appreciation topics	High skill	Low skill
Concatenated	Topics 7 and 10	62 859	201 924
Singular	Topics 1, 2, 5 and 6	94 755	170 028
	Topics 2, 5 and 6	101 439	163 344
	Topics 5 and 6	115 008	149 775

However, it is hardly the case that everyone of the 24% is equally interesting for law enforcement. For example, few individuals can have skills that are very sought after by other cybercriminals or for some other reason are more attractive targets for investigations. Therefore, we need to employ network centrality measures to prioritise proficient users further.

B. Network Centrality Analysis Results

Each forum user is assigned a unique and incremental User ID (UID); this value is a positive integer based on the order they registered as users. Furthermore, they receive a rank or group from their peers (typically assigned by moderators/administrators), which indicate their position in a forum. A variety of factors enable this group position to change during a user's lifetime. We obtained the forum groups for the Nulled forum, from their database tables, as seen in Table III. This group overview gives us the ground truth to compare our findings against, which was previously lacking in many related works. The reader can refer back to Table III to find short names for groups used in this section.

TABLE III
FORUM GROUP OVERVIEW

Group	Short name	# of members
Donator	Do	1
Moderators	Mo	1
Administrators	Ad	2
Legendary Reverser	LR	2
Senior Moderator	SM	3
VIP_Plus	VIP+	3
Reverser	Re	6
Legendary	Le	7
Contributor	Co	57
Royal	Ro	63
VIP	VIP	2245
Validating	Va	98837
Banned	Ba	111967
Members	Me	385891

Constructing a network of all the public posts would yield a network size of 299 702 vertices and 2 738 710 edges. However, constructing the same network using only high skilled users (Table II) from the concatenated approach, reduced the number of vertices by 79.6% and edges by 71.7%, for this particular dataset. More specifically, this digraph had 61 127 vertices and 773 983 edges. Consequently, network centrality algorithms took a much shorter time to complete; particularly for betweenness centrality, which has a time complexity of $O(VE)$. This is significant as such digraph can now be used for time-critical investigations.

At first glance, Table IV is almost identical in the ordering of who are most central individuals as our previous research [8]. The reason for this similarity is that we use the LDA results to extract a sub-graph, which retain a selected set of vertices and all their edges. More notably, we could identify a new user (with UID 574289) higher up in our result in this paper.

TABLE IV
CONCATENATED TOP TEN CENTRALITY RESULTS NULLED (FORUM GROUP OVERVIEW IS FOUND IN TABLE III)

UID	Group	C_{deg^-}	UID	Group	C_{deg^+}
15398	LR	0.294588	1471	Ad	0.016474
574289	LR	0.136505	8	SM	0.012891
1337	Le	0.100874	193974	Mo	0.01173
4	Re	0.08561	47671	Ba	0.011141
0	N/A	0.076759	334	Ba	0.010503

UID	Group	C_B	UID	Group	C_C
15398	LR	0.029145	15398	LR	0.535465
1337	Le	0.016342	1337	Le	0.472462
1471	Ad	0.012232	0	N/A	0.470536
334	Ba	0.008738	574289	LR	0.456597
574289	LR	0.0087	8841	Le	0.454961

UID	Group	C_E
15398	LR	0.207474
1337	Le	0.177821
0	N/A	0.148149
334	Ba	0.133743
22239	Le	0.13364

Many senior ranking members respond to lower-skilled cybercriminals for various reasons, such as helping them with guidance or answering questions, which artificially increase their network centrality scores. We can reduce this spurious effect of responding to many users and have a more accurate representation of important actors. Therefore, we see that results for betweenness, closeness and eigenvector centrality change the most from our previous research [8].

Centrality measures only provide a number to represent how central a user is, as seen in Table IV, which compares the relative centrality between users. However, it does not provide any other type of information, such as why they receive this score or their forum group. LDA provides us with this ability to inspect the overall posts made by individual actors, finding whether they are to interest in a continued investigation.

We re-run the LDA analysis for those users that the network centrality measures found as most central. We train this new LDA model using the hyper-parameter values ($\alpha = 0.05$ and $\eta = 0.05$ and $k = 3$) and using the singular approach mentioned in Section III.

Table V shows a selection of the top actors from each centrality measure. We can distinctly see that UID 1337 and 1471 talks about some administration-related topics. While UID 15398 and 574289 talks about some reverse engineering-related topics. Thus, the individualised LDA results have a correlation with the groups these users has been assigned. Similar to Samtani et al. [19], our result also indicates that many key/central members are those most senior and longest participants in their community, due to their low UID numbers.

TABLE V
SAMPLE OF CENTRAL INDIVIDUAL'S TOPICS

UID	Group	Topic	Words
1337	Le	1	data, update, work, thread, nulled, press
1337	Le	2	member, account, post, scam, download
1337	Le	3	ban, post, account, thread, thing, time
1471	Ad	1	game, forum, member, nulled, time
1471	Ad	2	deny, account, member, ban, pm, solve
1471	Ad	3	bump, ban, post, long, text, deny
15398	LR	1	loader, update, open, pipe, work, crack
15398	LR	2	member, rep, allow, hack, update, paypal
15398	LR	3	work, game, inject, bol, crack, nulled
574289	LR	1	bot, application, crack, download
574289	LR	2	bot, wrobot, feature, clashbot, download
574289	LR	3	version, download, troop, improve, add

V. CONCLUSION

We combine LDA and network centrality measures to identify proficient criminals, i.e. key hackers, in a real-world hacker forum. We can remove up to 79% of uninteresting users (who only wrote appreciating messages). This allowed us to focus our investigation on the remaining and presumed high skilled cybercriminals. This reduction allowed network centrality measures to run faster on a smaller sub-graph, as a lot of vertices and edges was removed. Furthermore, utilising a leaked hacker database allowed us to examine these methods in a best-case scenario, where we have the ground truth of forum user's groups.

Recall that our digraph was the product of who responded to which forum thread (i.e. a collection of related posts), instead of the actual relationship between users. Therefore, it is essential to note that centrality measures mostly identified users with viral threads. On the other hand, these threads become popular for a reason; for example, users can acquire the threads without gaining the technical skills to acquire them by themselves. When users post something that other users desire and become famous, they could be seen as having higher skills than the rest.

The contribution of our research is manifold. First, we proposed to study the underground economy through the lens of its participants, uniquely identifying the minority group of highly skilled cybercriminals. The minority group play a pivotal role in the CaaS and observing their behaviours enriches our understanding of it, allowing us to investigate central criminals further. Secondly, we developed advanced text-mining technique capable of identifying and profiling key underground hackers, and we experimented with evaluating its effectiveness.

For future work, we find it interesting to pursue a legal agreement to get access to CrimeBB dataset and repeat experiments, as presented in this paper. Although CrimeBB does not provide us with access to ground truth, it will provide us with other real-life underground forums to analyse.

REFERENCES

[1] Europol, "The Internet Organised Crime Threat Assessment (iOCTA) 2014," tech. rep., 2014.

[2] A. Abbasi, W. Li, V. Benjamin, S. Hu, and H. Chen, "Descriptive Analytics: Examining Expert Hackers in Web Forums," in *2014 IEEE Joint Intelligence and Security Informatics Conference*, (The Hague, Netherlands), pp. 56–63, IEEE, 2014.

[3] Z. Fang, X. Zhao, Q. Wei, G. Chen, Y. Zhang, C. Xing, W. Li, and H. Chen, "Exploring key hackers and cybersecurity threats in Chinese hacker communities," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, (Tucson, AZ, USA), pp. 13–18, IEEE, 2016.

[4] E. Marin, J. Shakarian, and P. Shakarian, "Mining Key-Hackers on Darkweb Forums," in *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, (South Padre Island, TX), pp. 73–80, IEEE, 2018.

[5] S. Pastrana, D. R. Thomas, A. Hutchings, and R. Clayton, "CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, (Lyon, France), pp. 1845–1854, ACM Press, 2018.

[6] Online Trust Alliance (OTA), "2018 Cyber Incident & Breach Trends Report," 2019.

[7] Europol, "Internet Organised Crime Threat Assessment (IOCTA) 2019," 2019.

[8] J. W. Johnsen and K. Franke, "Identifying Central Individuals in Organised Criminal Groups and Underground Marketplaces," in *Computational Science – ICCS 2018*, vol. 10862, pp. 379–386, Cham: Springer International Publishing, 2018.

[9] K. Porter, "Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling," *Digital Investigation*, vol. 26, pp. S87–S97, July 2018.

[10] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference - IMC '11*, (Berlin, Germany), p. 71, ACM Press, 2011.

[11] V. Benjamin and H. Chen, "Securing cyberspace: Identifying key actors in hacker communities," in *2012 IEEE International Conference on Intelligence and Security Informatics*, (Arlington, VA), pp. 24–29, IEEE, 2012.

[12] S. Pastrana, A. Hutchings, A. Caines, and P. Buttery, "Characterizing Eve: Analysing Cybercrime Actors in a Large Underground Forum," in *Research in Attacks, Intrusions, and Defenses* (M. Bailey, T. Holz, M. Stamatogiannakis, and S. Ioannidis, eds.), (Cham), pp. 207–227, Springer International Publishing, 2018.

[13] V. Benjamin, W. Li, T. Holt, and H. Chen, "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, USA), pp. 85–90, IEEE, 2015.

[14] V. Benjamin and H. Chen, "Developing understanding of hacker language through the use of lexical semantics," in *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, (Baltimore, MD, USA), pp. 79–84, IEEE, 2015.

[15] V. Benjamin and H. Chen, "Identifying language groups within multilingual cybercriminal forums," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, (Tucson, AZ, USA), pp. 205–207, IEEE, 2016.

[16] W. Li and H. Chen, "Identifying Top Sellers In Underground Economy Using Deep Learning-Based Sentiment Analysis," in *2014 IEEE Joint Intelligence and Security Informatics Conference*, (The Hague, Netherlands), pp. 64–67, IEEE, 2014.

[17] W. Li, H. Chen, and J. F. Nunamaker, "Identifying and Profiling Key Sellers in Cyber Carding Community: AZSecure Text Mining System," *Journal of Management Information Systems*, vol. 33, no. 4, pp. 1059–1086, 2016.

[18] W. Li, J. Yin, and H. Chen, "Identifying High Quality Carding Services in Underground Economy using Nonparametric Supervised Topic Model," p. 10, 2016.

[19] S. Samtani and H. Chen, "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, USA), pp. 319–321, IEEE, 2016.

[20] S. Samtani, R. Chinn, and H. Chen, "Exploring hacker assets in underground forums," in *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, (Baltimore, MD, USA), pp. 31–36, IEEE, 2015.

[21] S. Samtani, R. Chinn, H. Chen, and J. F. Nunamaker, "Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence," *Journal of Management Information Systems*, vol. 34, no. 4, pp. 1023–1053, 2017.

[22] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart, and P. Shakarian, "Darknet and deepnet mining for proactive cybersecurity threat intelligence," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, (Tucson, AZ, USA), pp. 7–12, IEEE, 2016.

[23] E. Marin, A. Diab, and P. Shakarian, "Product offerings in malicious hacker markets," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, (Tucson, AZ, USA), pp. 187–189, IEEE, 2016.

[24] S.-Y. Huang and H. Chen, "Exploring the online underground marketplaces through topic-based social network and clustering," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, (Tucson, AZ, USA), pp. 145–150, IEEE, 2016.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," p. 30, 2003.

[26] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam, "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology," *Communication Methods and Measures*, vol. 12, no. 2-3, pp. 93–118, 2018.

[27] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," p. 10, 2009.

[28] I. McCulloh, H. Armstrong, and A. Johnson, *Social Network Analysis with Applications*. Wiley, 2013.