2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering

A Topic-based Unsupervised Learning Approach for Online Underground Market Exploration

Shin-Ying Huang
Institute for Information Industry
Taipei, Taiwan
shinyinghuang@iii.org.tw

Tao Ban
National Institute of Information and Communications Technology
Tokyo, Japan
bantao@nict.go.jp

***Abstract*—Cyber fraud has become a lucrative form of illicit business by leveraging the Internet as a communication channel and as a result, causes significant losses to the economy. Criminals in the cyber fraud underground economy use online underground markets and other forms of social media to exchange and trade illegitimate information. Due to the high variability in the marketplaces and actors therein, analyzing these underground markets is challenging. To understand more about the underground economy of cyber fraud and its actors, we propose a topic-based hierarchical self-organizing map, which can well represent and visualize actors' similarity and thus, uncover their roles in the underground markets. We compare the proposed method with a topic-based social network analysis method for identifying the key users and their roles in the cyber fraud value chain. Experiments conducted on data from several online underground markets suggest that the proposed method can aid in identifying key actors in terms of roles, influence levels, and their social relationships.**

Keywords—cybersecurity; black market; social network analysis; clustering; threat intelligence; growing hierarchical self-organizing map

I. INTRODUCTION

In recent years, cybercrime has become a critical issue that has caused significant economic losses. The global loss due to cybercrime in 2017 is estimated to be close to $600 billion [1]. Cybercriminals underground markets across various regions have demonstrated that trading illicit digital goods and services has turned out to be a lucrative business. Furthermore, the businesses that serve the cybercrime market have expanded their offerings to encompass a range of activities that allow many levels of engagement [2]. For example, consider the trade for illegitimate credit card information. Largely due to the success of point-of-way (POS) malware-related attacks [3], the proliferation of development kits for PoS RAM scrapers are witnessed in the underground markets. In the meantime, credit card companies, as well as observant card owners, are quick to spot anomalous patterns in user transactions, which entails that criminals need a steady supply of "fresh" card numbers from the underground markets. In this way, underground markets constitute a ready market for a variation of digital goods and services for buyers and sellers [4].

In China, as a clear consequence of the wide adoption of electronic and mobile payment means, the non-cash transaction volume of cybercrime-related sales has drastically increased in the past two years. For example, a credit card fraud made headlines in February 2014, when an individual from Hangzhou was trialed in the United States for a successful spam run that cost card providers a loss of roughly US$ 808,855 [5]. In recent years, China and Taiwan both rank among the top ten hacking countries and regions [6], which indicates that a great deal of illegitimate data from these regions has been distributed to the underground market. The online underground markets in China also export cybercrime technologies and equipment to international groups. For instance, crime rings involved in identity theft and credit card-forgery in New York City allegedly sourced their skimming equipment, together with the blank credit cards, from China, Lebanon, Libya, and Russia [7].

Cybercriminals rely on underground markets to trade illegitimate goods and services such as malware tools, stolen credit cards, botnets, attack services, etc. This motivated our research to develop a hands-on methodology for analyzing the activities of the underground markets and characterizing the actors therein. The following are the contributions of this paper: First, we develop a topic-based growing hierarchical self-organizing map approach for identifying key users in social network learning and to the best of our knowledge; this is the first application of this type of method in analyzing online underground markets. Second, by adopting social network visualization along with topic-based growing hierarchical self-organizing map, this approach can capture subgroup relative relationship in two-dimensional (2D) view and can easily identify the key user communities with the severity level.

The rest of this paper is organized as follows. Section II provides a brief review of related work. Section III describes the research methodology. Section IV presents the experiment for evaluating our proposed method. Section V draws the conclusions and presents the plan for future study.

II. RELATED STUDIES

In this section, we review the research study in two related areas: a criminal analysis using social network analysis and cyber fraud community analytics in online social media. We also highlight the methodologies used in the related study including a self-organizing map and growing hierarchical self-organizing map.

2324-9013/19/$31.00 ©2019 IEEE
DOI 10.1109/TrustCom/BigDataSE.2019.00036

A. Criminal analysis

There are many researchers using social network analysis (SNA) methods to conduct criminal analysis. For example, Shaabani et al. [8] studied the problem of early identification of violent gang users. Their approach relies on modified centrality measures that take additional data of the individuals in the social network of co-arrestees into account. These measures, together with other metadata, provide a rich set of features for a classification algorithm. Tayebi et al. [9] addressed an important problem of analyzing co-offending networks -networks of offenders who have committed crimes together, by proposing a framework for co-offense prediction using supervised learning.

B. Community analytics in the online underground economy

The majority of the online underground community research study can be categorized into three sub-areas. The first type of research focuses on identifying the threats found in the content and other content-related features. The second type of research focuses on identifying the most influential community users. The third type focuses on understanding the community structure and social relationships (Section C describes it in details).

1) Identifying threats

These studies primarily apply machine learning, text mining, or other network analysis methods to identify threats or related evidence in the online underground marketplace. Al-Rowaily et al. [10] presented the development of a bilingual sentiment analysis lexicon for the cybersecurity domain, which consists of a sentiment lexicon for English and Arabic. Isah et al. [11] proposed a bipartite network model for inferring hidden ties between actors who initiate an illegal interaction and objects affected by the interaction. Benjamin et al. [12] developed an automated methodology for identifying tangible and verifiable evidence of potential threats within hacker forums, IRC channels, and carding shops. Macdonald et al. [13] examined the communication activities of the actors involved in communication platforms such as discussion forums. Using automated analytical tools, the language of hackers was analyzed to identify potential threats against critical infrastructures. Benjamin and Chen [14] utilized the recurrent neural network language models to develop an unsupervised machine learning technique for learning hacker language.

2) Identifying influential community users

Some studies focus on identifying the key users and understanding users' interests by using machine learning, text mining, or network analysis approaches. Yang et al. [15] incorporated the message similarity and response immediacy features with link analysis to determine the impact and the neighborhood of the influential users. Tang et al. [16] used bipartite graph analysis and developed a user interest and topic detection model to predict user participation in the Dark Web. Li et al. [17] developed a text mining-based framework, dubbed AZSecure, for identifying and characterizing key sellers in the underground economy. Advanced text mining techniques including deep learning and topic modeling are developed to evaluate and profile sellers based on customer reviews and advertisements, respectively.

C. Detecting communities

Many prior studies have utilized network analysis with text mining techniques to study community composition. L'Huillier et al. [18] combined text mining and SNA techniques to achieve a complete understanding of what the main interests of the Dark Web community are. Ríos et al. [19] proposed a novel approach that combines topic-model based text mining techniques and traditional network analysis methods for overlapping community detection. The proposed methods were useful to mine the Dark Web portal.

1) Self-organizing map in criminal analysis

Self-organizing map (SOM) [20] is an unsupervised learning method which can project data with high dimensional features into 2D space so that the similarity of samples can be visualized properly. SOM has been widely used in the criminal analysis. Li et al. [21] proposed a framework of intelligent decision-support model based on a fuzzy self-organizing map network to detect and analyze crime trend patterns from temporal crime activity data. Poelmans et al. [22] compared the usability of emergent self-organizing map and multi-dimensional scaling as text exploration instruments in police investigations. Currently, the results of their research study are operational in the Amsterdam-Amstelland police region. Olszewski [23] proposed a fraud detection method based on user accounts visualization and threshold-type detection. To visualize the user accounts, this approach employed the SOM technique and set the classification threshold by finding the ridge in the unified distance matrix (U-matrix) of a given SOM. U-matrix is a 2D image representation of a SOM where the Euclidean distance between the vectors of neighboring neurons is depicted.

2) Growing hierarchical self-organizing map

The weaknesses of SOM include a predefined fixed topology size and its inability to enlighten hierarchical relations among samples. To address this issue, Dittenbach et al. [24] proposed the growing hierarchical self-organizing map (GHSOM). GHSOM is an extension of SOM, which can develop a multilayer hierarchical network structure. Due to the flexibility of the hierarchical structure, GHSOM is helpful to provide a more sophisticated clustering result than SOM. GHSOM has been widely applied in the fields of web mining [24][25], text mining [26][27][28], and data mining [29][30][31]. GHSOM has also been used in anomaly detection fields such as network anomaly detection [32][33][34], exploring malicious behaviors of the mobile applications [31], and financial fraud detection [35].

III. METHODOLOGY

Our proposed framework, as illustrated in Fig. 1, consists of the following steps. First, we collect data from the online underground forum. Next, we perform feature extraction through topic modeling and generate the user-based features from the historical posts. We then analyze the social media ecosystem by doing GHSOM exploration, SOM exploration, and SNA, respectively. Based on the results of social media ecosystem analytics, we extract the key users, visualize the sub-communities, and summarize

their common characteristics. Finally, we evaluate each method in the social media ecosystem analytics step by measuring the clustering quality. We compare our proposed topic-based GHSOM method with SOM and social media analysis. In particular, we compare the topology of the detected communities in the social media ecosystem analysis stage and then, we compare the quality and perform the out-of-sample test in the evaluation stage.

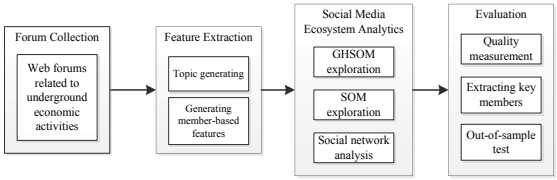


Fig. 1. The proposed framework

A. Data collection

The marketplace of interest is the Baidu forum (http://tieba.baidu.com/), the largest Chinese forum in China. The Baidu forum provides a keyword searching mechanism and lightly regulated login and post mechanism. Different from the QQ group (http://qun.qq.com/), the Baidu forum does not require permission from the forum creators to join the discussion, which means everyone can gain access to the contents. Therefore, the Baidu forum suffers from heavy abuse and become a major public accessible site where underground trading information can be collected. We collected the cyber-fraud-related Baidu forums by using the following keywords: "four pieces," "interception," "black card," "internal materials," and "cvv" (card verification value). In particular, "four pieces" refers to the account name, Social Security number, credit card number, and passwords. When forums whose titles matched with the predefined keywords are crawled, all threads posted on the forum together with the user information are retrieved. We developed a parsing engine to transform the HTML data into post records, where the attributes are stored in different fields.

B. Feature extraction

Feature extraction of the posts is obtained in two steps. First, we apply the topic modeling to generate the topic categories. Then, we generate user profiling features.

1) Topic clustering

To extract topics from textual contents, we adopt the well-known method, latent Dirichlet allocation (LDA) [36]. LDA is a probabilistic model (a Bayesian model) that relates documents and words through variables, which represent the main topics inferred from the text itself.

A document can be considered as a mixture of topics (z), represented by probability distributions that can generate the words in a document given these topics. Given the smoothing parameters and joint distribution of a topic mixture, the idea is to determine the probability distribution that generates from a set of topics T - a message composed by a set of *word,* $\mathbf{w} = (w^1,.., w^S)$. The posterior probability $p(z_s|\theta)$ can be represented by the random variable, θ_i. Statistical techniques are then utilized to learn the topic components and mixture coefficients of each document.

After the LDA analysis, we proceed to explain the results of topic modeling and generated topic-based similarity and carder-based similarity, accordingly. We label the roles based on the keywords of each topic. Generally, the roles in the cyber fraud value chain include attack originator, buyer, dropper, shopper, runner, and other sellers [37].

2) User profiling features

For the topic-based similarity features, LDA provides the most similar topic, the scores of the word coefficients and the document coefficients for each post. To extract features that can characterize the active level of a user, we calculate the following features: number of posts, number of feedbacks, number of participated forums, and the most represented topic category.

Furthermore, we develop five indices to measure the *recency*, *engagement*, *inferentiality*, and *topic severity* of a particular user, *i*. The first four indices are then combined as a *severity* index as shown in the following.

Recency: $1/((Latest\_post\_year_i – Post\_Year_i) * 360+(12- Post\_month_i) * 30))* 360$

Engagement: $Total\_post_i + Number\_of\_forums_i$

Inferentiality: $Number\_of\_feedback_i \times (Latest\_post\_year_i – Earliest\_post\_year_i)$

Severity: Normalize(*Recency*) + Normalize(*Engagement*) + Normalize(*Inferentiality*) + Normalize(*Topic severity*)

C. Social media ecosystem analytic

In the social media ecosystem analytics step, we perform GHSOM exploration, SOM exploration, and social media analysis to extract the sub-communities from users.

1) GHSOM exploration and SOM exploration

The input features for GHSOM includes topic categories, document frequency, words frequency, number of feedbacks, number of posts, number of forums, year, month, recency, engagement, inferentiality, and topic severity. The GHSOM training process contains the following four phases [25]:

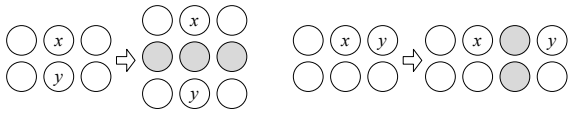
1. Initialize layer 0: Layer 0 includes a single node, the weight vector of which is initialized as the expected value of all input data. Then, the mean quantization error of layer 0 (MQE_0) is calculated, where MQE of a node denotes the sum of the deviation between the weight vector of the node and all input data mapped to the node.

2. Train individual maps: Under the competitive learning principle, only the winner and its neighboring nodes qualify for an adjustment of their weight vectors. This competition and adjusting process is repeated until the learning rate decreases to a certain value.

3. Grow each individual map horizontally: Each individual map grows until the mean value of the MQE for all nodes on the map, i.e., avg (MQE), is smaller than the product of τ_1 and the MQE of the parent node, MQE_p, as in (1). If the stop criterion is not satisfied, we find the error node that owns the largest MQE and inserts one row or column of new nodes between the error node and its

dissimilar neighbor, as shown in Fig. 2. The notation *x* indicates the error node and y indicates the dissimilar neighbor.

$$\text{avg}(\text{MQE}) < \tau_1 \times \text{MQE}_p \quad (1)$$



(a) insert a row (b) insert a column

Fig. 2. Horizontal growth of GHSOM

4. Expand or terminate the hierarchical structure: If a node with an MQE_i greater than the product of τ_2 and MQE_0, then it will be used to develop the next layer.

For the SOM exploration step, the input features for SOM are the same as GHSOM. To compare the results of GHSOM with SOM, we set the total number of units (nodes) similar to the size of the chosen GHSOM. The structure difference is that the topology yielded by SOM is 2D while that by GHSOM is 3D.

2) Social network analysis

In this step, SNA is applied to explore the relationship between active users and their possible roles. To generate a topic-based social network, we set the users and topics as nodes and set the post action as the edges. Then, we perform community detection on top of the topic-based social network to visualize the sub-communities in the online underground marketplace. We use the Louvain method proposed by Blondel et al. [38][39] to detect communities. The method is a greedy optimization method that attempts to optimize the modularity of a partition of the network. The optimization is performed in two steps. First, the method looks for small communities by optimizing modularity locally. Second, it aggregates nodes belonging to the same community and builds a new network whose nodes are the communities. These steps are repeated iteratively until a maximum of modularity is attained.

D. Evaluation

The evaluation step consists of three tasks: quality measurement, key user extraction, and out-of-sample test. These procedures are used to verify the proposed method in terms of clustering quality, practical implications, and generality.

1) Quality measurement

We perform quality measurement for each component in the last social media ecosystem analytics step. We choose the parameter settings with the best clustering quality and also compare the difference of three components in respect of clustering quality, clusters structure, and reasonability.

2) Extracting key users

The main purpose of extracting key users is to discover the active and influential actors in the communities. Generally, key users tend to post more contents and receive more replies compared to other users.

For each node of the GHSOM exploration and SOM exploration, we calculate the average severity to represent the group severity and use the majority role to represent the group role. Then, we highlight the groups with higher average severity and presume the users that belong to these groups are key actors. The group center of each node will be used to measure the similarity with other unknown samples (i.e., holdout samples) for role detection and severity prediction.

For the generated topic-based social network, we adjust the threshold of the topology degree to reveal the active nodes and their interrelationships. We visualize the key users' associated roles in online underground markets and illustrate the sub-communities among these key users using different node color.

IV. Experimental Results and Discussion

A. Data collection

Based on the associated keywords (or jargon) defined in Section 4.1, we collected 21 Baidu forums as the research testbed (Table I). There are 2,947 users, 5,131 threads, and 53,963 threads including replies. The time period covers January 2006 through March 2016. We randomly chose 10% of the users as the testing samples.

TABLE I. Forum Information

No	Forum name	Threads	No	Forum name	Threads
1	cvvvvv	6,565	12	originalChannel	977
2	cvvvvvvvvp	3,586	13	blackProduct	736
3	four	4,215	14	track	1,952
4	innersave	1,459	15	materialOwner	655
5	brave	3,082	16	jp	168
6	bank	2,745	17	pickMoney	254
7	collectMachine	796	18	interceptMaterial	151
8	wild	7,236	19	outsideCard	111
9	card	3,068	20	washIntercept	92
10	collect	2,310	21	Four pieces	12,419
11	outsideMachine	1,015			

B. Feature extraction–topic generation

To cover the general roles in the cyber fraud value chain, we set the number of topics equal to ten [30]. Table II shows the extracted topics, represented roles, descriptions, and the assigned topic severities.

TABLE II. Topic Descriptions

Topic	Role	Description	Topic severity
01	Runner/Shopper	Online banking	3
02	Other sellers	General networking	1
03	Buyer	Equipment–ATM skimmer	5
04	Runner/Shopper	Collaboration	2
05	Buyer	Equipment–PoS skimmer	5
06	Dropper	Stolen data–China region	3
07	Other sellers	Deceiver report	1
08	Dropper	Stolen data–foreign	4
09	Shoppers	Alipay fraud	3
10	Attack originator	Use/sell malware to steal data	5

C. Social network ecosystem analytics and evaluation

1) GHSOM exploration and SOM exploration

For the GHSOM parameter setting, we set τ_1 from 0.1 to 0.9, and match each τ_1 with τ_2 to 0.01, 0.03, 0.05, 0.07,

0.09, respectively. Afterward, a grid search is performed to select the parameters that can clear the following selection criteria: average quantitative error (*aqe*) < 0.05, number of clusters (nofclusters) < 40. Table III shows the partial parameter table satisfied with the selection criterion *aqe* < 0.05 and sorted by *aqe*. The parameter settings satisfied the criterion *nofclusters* < 40 are underlined.

TABLE III. GHSOM PARAMETER TABLE

t1	t2	layer	clusters	Parent	aqe
0.3	0.01	2	109	31	0.003929
0.4	0.01	2	123	20	0.006291
0.5	0.01	2	69	9	0.017119
0.5	0.03	2	60	12	0.02051
0.6	0.01	3	66	2	0.021983
0.4	0.03	2	23	1	0.032821
0.6	**0.03**	2	**39**	9	0.039489
0.5	0.05	2	25	11	0.041977
0.6	0.05	2	34	9	0.043257
0.7	0.01	3	55	4	0.054248
0.8	0.01	3	55	4	0.054248
0.9	0.01	3	55	4	0.054248

We choose the GHSOM parameter setting (τ_1=0.6, τ_2=0.03) because it satisfies the criteria *aqe* < 0.05 and nofclusters < 40 and can provide more delicate clusters for further analysis.

For the SOM parameter setting, we generate SOM which has a similar cluster size with the chosen GHSOM (39 clusters) and set the SOM map size to 6*6 (36 clusters) and 8*5 (40 clusters). We find that the clustering quality of SOM(8*5) is better than SOM(6*6), so we choose SOM(8*5) to compare it with GHSOM(0.6, 0.03). The comparison is shown in Fig. 3.

The clustering quality of GHSOM is better than SOM because GHSOM can group similar clusters together while providing hierarchical relationships. Therefore, we use the clustering results of GHSOM in further investigation.

2) SNA

We extract the key users by increasing the lower bound of degree range of the generated topic-based social network. Fig. 4 shows the identified key users under the lower bound degree parameter to seven. We can keep increasing the lower bound degree to decrease the number of remained key users.

In Fig. 4, the center parts of the filtered topic-based network are the categories of roles and the outer circle which locates the extracted key users, such layout helps to illustrate the user-topic correlation. The relationships of key users are illustrated through different node colors and connections based on their specialties and similarities. By increasing the topology degree threshold up to 7, we find 7 key users and their associated topics.

We use the severity score defined in Section IV to rank all these key users. The ranked key users are shown in Table IV. We map the key users extracted from the topic-based SNA to their belonging clusters in GHSOM. The results are shown in Fig. 5.

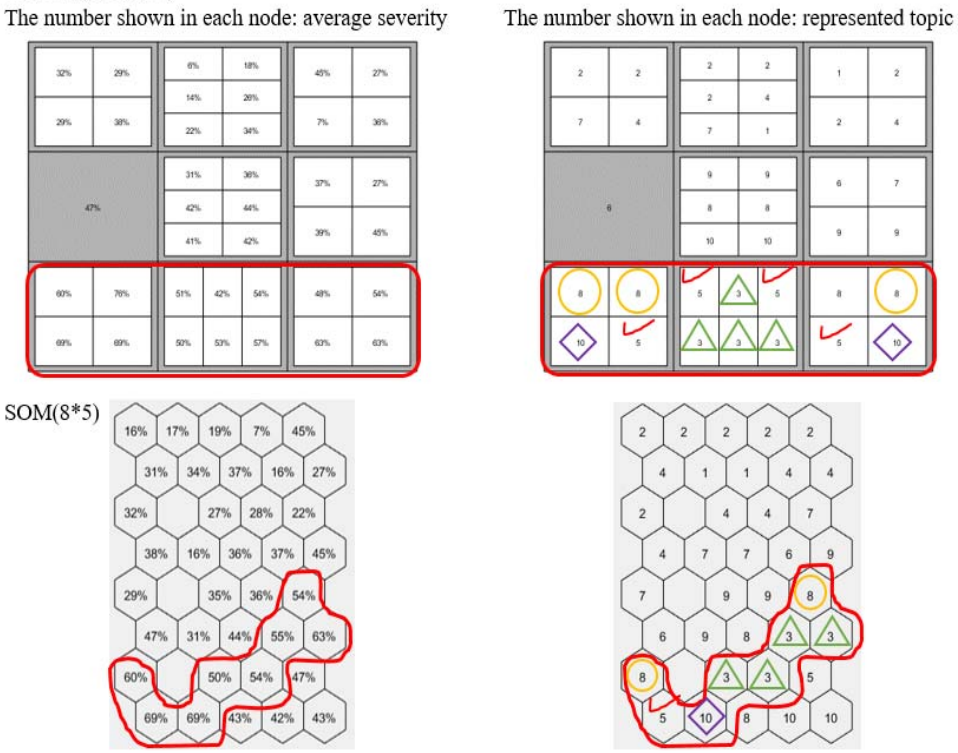


Fig. 3. GHSOM vs. SOM

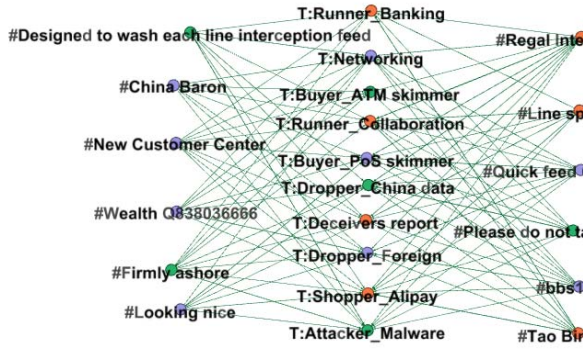


Fig. 4. The identified key users (Color: sub-communities; T: topic; #: user)

TABLE IV. KEY USERS

Rank	User
1	#Wealth Q838036666
2	#Line spacing 1
3	#Quick feed Q838036666
4	#Please do not take this sister
5	#Looking nice
6	#Firmly ashore
7	#Regal International 1
8	#bbs16163
9	#Tao Binghong
10	#Designed to wash each line interception feed
11	#China Baron
12	#New Customer Center

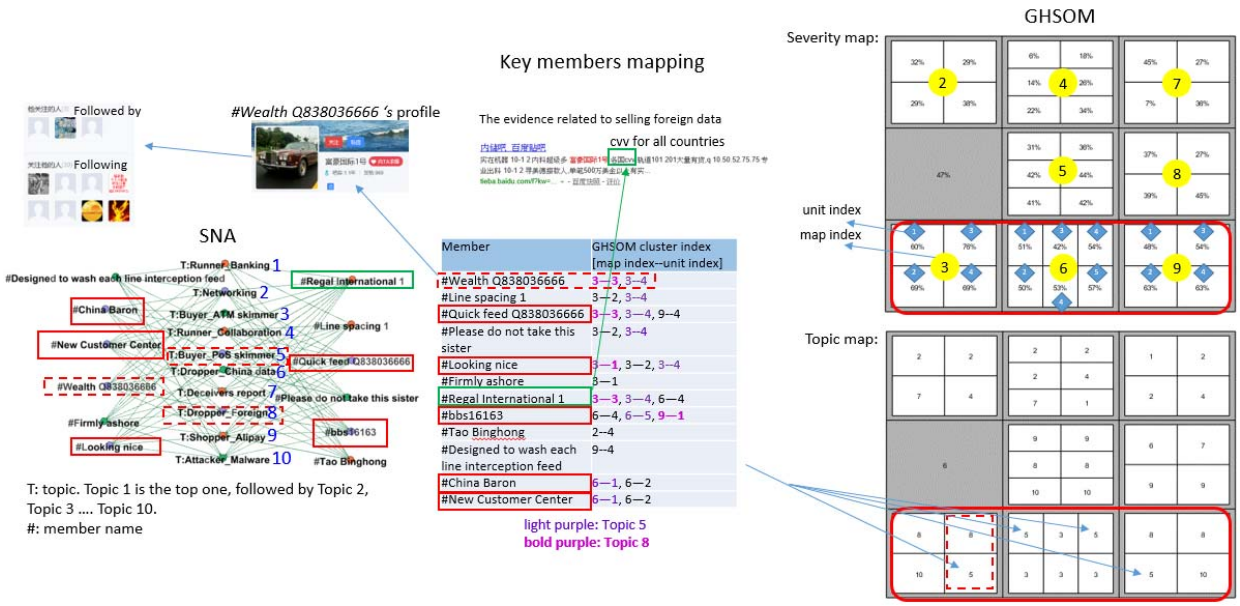


Fig. 5. Key user mapping

It is worth noting that GHSOM indicates that #Regal International 1 (marked with a green square) is also very similar to #Wealth Q838036666, but SNA did not place them in the same community. Therefore, GHSOM can help to generate clusters with distinctive characteristics and use topology layout to show the degree of difference between different clusters. The knowledge extracted from the proposed methods are listed as follows:

- Ranking of active members by severity.
- Clustering of members with categories.
- Network and hierarchical relationship with other members.

3) Out-of-sample test

We left 10% users as out-of-samples and used the trained GHSOM (0.6, 0.03) to classify these testing samples. We used Euclidean distance to measure the similarity and classify the testing sample into its closest cluster. It is worth noting that GHSOM has no information such as the historical posts about all the users in testing samples. The evaluation indices include:

- Average error of severity: average(sum(abs(true severity–predicted severity)))
- Topic precision: sum(if(true topic = predicted topic) then 1, else 0)/ number of testing sample

With this setting, we got the following result of the out-of-sample test on the 10% left out dataset: average error of severity = 3.22% and topic precision = 84.45%. As an unsupervised learning approach, the prediction result is impressive and demonstrates a high ability to recognize unknown samples.

V. Conclusion

We propose a topic-based SNA approach with unsupervised clustering methods for identifying the key users and their associated roles from the online underground markets. We perform a numerical study using data collected from the Baidu forums. The experimental results show that our proposed framework is capable of identifying the key users, profiling their roles and visualizing their relationships. We developed a severity index which is helpful to explain the results of GHSOM, SOM, and SNA, and also helps to rank the key users based on the factors of engagement, inferentiality, recency, and topic severity. Furthermore, we show that the developed GHSOM key user exploration mechanism is effective in detecting the roles of unknown users in the cyber fraud value chain. The identification can lead to further criminal investigations for China's government. Moreover, as the darkweb and black markets are constantly changed, this method can be generalized to be applied to different social media contents, for example, OSINT (open source threat intelligence) data or darkweb data.

The future work includes four possible avenues of investigation. First, including data from QQ groups and correlating the QQ accounts appearing in the Baidu forums may deepen our understanding of key users' identities, behavior, and networks. Second, it may be fruitful to apply the proposed framework to other types of online underground markets where knowing more about key users' and their networks is desirable. Third, investigating the credibility of community users by applying sentiment analysis may shed further light on key users' reputations. Fourth, studying the jargon and writing styles used in online underground markets may help in further identification of key users.

Acknowledgment

We acknowledge the financial support of Ministry of Economic Affairs Taiwan (107-EC-17-D-11-1502).

References

[1] Center for Strategic and International Studies, McAfee, Economic Impact of Cybercrime— No Slowing Down, https://www.mcafee.com/enterprise/en-us/assets/reports/restricted/rp-economic-impact-cybercrime.pdf, Feb, 2018.

[2] D. Manky, FortiGuard Labs, Cybercrime as a Service: a very modern business (2013).

[3] Trend Micro, Annual security roundup (2014).

[4] Symantec, 2015 Internet security threat report (2015).

[5] Trend Micro, Prototype nation-The Chinese Cybercriminal underground in 2015 (2015).

[6] Akamai, Q4 2015 State of the Internet – Security Report, http://www.stateoftheinternet.com/security-report (2016).

[7] M. J. Schwartz, 111 arrested in identity theft probe. InformationWeek. http://www.informationweek.com/news/security/attacks/231900438 (2011).

[8] E. Shaabani, A. Aleali, P. Shakarian, J. Bertetto, Early Identification of Violent Criminal Gang Users, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015).

[9] M. A. Tayebi, M. Ester, U. Glässer, P. L. Brantingham, Spatially embedded co-offence prediction using supervised learning, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (2014).

[10] K. Al-Rowaily, M. Abulaish, N. Al-Hasan Haldar, M. Al-Rubaian, BiSAL – A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security, Digital Investigation 14 (2015) 53-62.

[11] H. Isah, D. Neagu, P. Trundle, Bipartite Network Model for Inferring Hidden Ties in Crime Data, Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2015).

[12] V. Benjamin, W. Li, T. Holt, H. Chen, Exploring Threats and Vulnerabilities in Hacker Web: Forums, IRC and Carding Shops, Proceedings of IEEE International Conference on Intelligence and Security Informatics (ISI) (2015) 85-90.

[13] M. Macdonald, R.Frank, J. Mei, B. Monk, Identifying Digital Threats in a Hacker Web Forum, Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2015).

[14] V. Benjamin, H. Chen, Developing Understanding of Hacker Language through the User of Lexical Semantics, Proceedings of IEEE International Conference on Intelligence and Security Informatics (ISI) (2015) 79-84.

[15] C. F. Yang, X. Tang, B. M. Thuraisingham, An analysis of user influence ranking algorithms on Dark Web forums, Proceedings of the 10th ACM SIGKDD Workshop on Intelligence and Security Informatics (2010).

[16] X. Tang, C. C. Yang, M. Zhang, Who will be participating next-predicting the participation of Dark Web community, Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (2012).

[17] W. Li, H. Chen, J. F. Nunamaker, Cyber Carding Community Collection and Analytics: The AZSecure Text Mining Research Framework, Journal of Management Information Systems (2016) (In Press).

[18] G. L'Huillier, S. A. Ríos, H. Alvarez, F. Aguilera, Topic-based social network analysis for virtual communities of interests in the Dark Web, Proceeding of the 10th ACM SIGKDD Workshop on Intelligence and Security Informatics (2010).

[19] S. A. Ríos, R. Muñoz, Dark Web portal overlapping community detection based on topic models, Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (2012).

[20] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43 (1982) 59-69.

[21] S. T. Li, S. C. Kuo, F. C. Tsai, An intelligent decision-support model using FSOM and rule extraction for crime prevention, Expert Systems with Applications 37:10 (2010) 7108-7119.

[22] J. Poelmans, M. M. V. Hulle, S. Viaene, P. Elzinga, G. Dedene, Text mining with emergent self organizing maps and multi-dimensional scaling: A comparative study on domestic violence, Applied Soft Computing 11:4 (2011) 3870-3876.

[23] D. Olszewski, Fraud detection using self-organizing map visualizing the user profiles, Knowledge-Based Systems 70 (2014) 324-334.

[24] M. Dittenbach, D. Merkl, A. Rauber, The Growing hierarchical self-organizing map, Proceedings of the 2000 International Joint Conference on Neural Networks (IJCNN) (2000).

[25] M. Dittenbach, A. Rauber, D. Merkl, Uncovering hierarchical structure in data using the growing hierarchical self-organizing map, Neurocomputing 48:1-4 (2002) 199-216.

[26] A. Rauber, D. Merkl, M. Dittenbach, The Growing hierarchical self-organizing map: exploratory analysis of high-dimensional data, IEEE Transactions on Neural Networks 13(6) (2002) 1331-1341.

[27] E. Schweighofer, A. Rauber, M. Dittenbach, Automatic text representation classification and labeling in European law,

Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL), ACM Press (2001).

[28] H. C. Yang, C. H. Lee, H. W. Hsiao, Incorporating self-organizing map with text mining techniques for text hierarchy generation, Applied Soft Computing 34 (2015) 251-259.

[29] J. Y. Shih, Y. J. Chang, W. H. Chen, Using GHSOM to construct legal maps for Taiwan's securities and futures markets, Expert Systems with Applications 34:2 (2008) 850-858.

[30] R. H. Tsaih, W. Y. Lin, S. Y. Huang, Exploring Fraudulent Financial Reporting with GHSOM, Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics (PAISI), Lecture Notes in Computer Science 5477 (2009) 31-41.

[31] F. Yu, S. Y. Huang, L. C. Chiou, R. H. Tsaih, Clustering iOS Executable Using Self-Organizing Maps, Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN) (2013).

[32] Y. Huang, S. Y. Huang, Discover the dynamics of a zero-day vulnerability," Journal of the Chinese Institute of Engineers 88:4 (2015) 84-94.

[33] E. Hoz, E. Hoz, A. Ortiz, J. Ortega, A. Martínez-Álvarez, Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps, Knowledge-Based Systems 71 (2014) 322-338.

[34] E. J. Palomo, J. North, D. Elizondo, R.M. Luque, T. Watson, Application of growing hierarchical SOM for visualisation of network forensics traffic data, Neural Networks 32 (2012) 275-284.

[35] S. Y. Huang, R. H. Tsaih, Y. Fang, Topological Pattern Discovery and Feature Extraction for Fraudulent Financial Reporting, Expert Systems with Applications 41 (2014) 4360-4372.

[36] D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993-1022.

[37] A. Singh, The Underground Economy of Credit-Card Fraud. https://www.peerlyst.com/posts/the-underground-economy-of-credit-card-fraud (2016).

[38] V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, 10:10008 (2008).

[39] Gephi, https://gephi.org/