Chinese Underground Market Jargon Analysis Based on Unsupervised Learning

Kangzhi Zhao, Yong Zhang, Chunxiao Xing
Research Institute of Information Technology for
Information Science and Technology, Department of
Computer Science and Technology, Tsinghua University
Beijing, China
zkz15@mails.tsinghua.edu.cn,
{zhangyong05, xingcx}@tsinghua.edu.cn

Weifeng Li, Hsinchun Chen
Department of Management Information Systems
The University of Arizona,
Tucson, AZ 85721, USA
weifengli@email.arizona.edu, hchen@eller.arizona.edu

***Abstract*—With the rapid growth of online population, China has become the world's largest online market. This also gives rise to the Chinese underground market, which has facilitated many of the cybercrimes in China. Consequently, there is a need for research scrutinizing Chinese underground markets. One major challenge facing cybersecurity researchers is to understand the unfamiliar cybercriminal jargons. To this end, we are motivated to analyze jargons in Chinese underground market. Particularly, we utilize the recent advancements in unsupervised machine learning methods, word embedding and Latent Dirichlet Allocation. We evaluate our work on a research testbed encompassing 29 exclusive underground market QQ groups with 23,000 members. Specifically, we test the ability of the proposed approach to learn semantically similar words of known cybersecurity-related jargons. Results suggest the state-of-the-art unsupervised learning approaches can help better understand cybercriminal language, providing promising insights for future research on Chinese underground markets.**

Keywords—Chinese underground market; cybersecurity; language model; unsupervised learning

I. Introduction

While the internet has greatly facilitated ecommerce over the past decades, the internet has also become a major platform for the underground market, where cybercriminals exchange malicious products and services. In 2013, global cybercrime was estimated to cause $ 375 billion in loss, almost as much as the cost of drug trafficking[1]. With the rapid growth of online population, China has become the world's largest online market. This also gives rise to the Chinese underground market, which has facilitated many of the cybercrimes in China. In 2015, cybercrime has caused an estimate of 32% of the Chinese Internet users to suffer a total loss of 805 billion yuan ($124 billion) [2]

Lately, research on underground markets has become increasingly popular. While various topics in this area have been investigated to benefit cybersecurity, there are several open challenges facing cybersecurity researchers. One of the major challenges is the constantly emerging covert, underground jargons that cybersecurity researchers or investigators are usually unfamiliar with. Due to the exclusiveness of underground markets, Chinese cybercriminals have developed a unique language system that is mostly composed of jargons. Jargons are used to refer to a certain product, service, and other cybercriminal-specific concept. Understanding jargons is a critical problem for mining underground markets. However, with enormous posts arising in underground markets every day, we have limited understanding of hacker jargons in a scalable manner.

We are motivated to study jargons of Chinese underground markets by developing an automated method. Recent progress of unsupervised learning on lexical semantics has provided us with a plausible way to automatically understand words and expressions. For example, Word2vec [19] using recurrent neural network provides good features to represent terms. Moreover, Latent Dirichlet Allocation (LDA) [20] can be seen as a conceptual clustering method, which automatically groups semantically related terms to form high-level cybercrime related features. As such, unsupervised learning has the potential to help identify and better understand jargons in underground markets. Hence, we propose the development of a Chinese hacker language framework based on unsupervised learning.

The rest of the paper is organized as follows. In Section 2, we review the related work and then proposes research gaps and questions. In Section 3, we describe our design for jargon analysis of Chinese underground markets. In Section 4, we show the experiments and discuss the results. In Section 5, we summarize our research.

II. Literature Review

We review the prior works in the following research areas to form the basis of our work: Underground Market and Lexical Semantics.

A. *Underground Market*

Underground markets refer to cyber black markets for crime-related products or services [2]. There are three main underground market platforms: hacker forums, Internet-Relay-Chat (IRC) channels, and carding shops [5]. For years, cybercriminals have been advertising and trading their malicious products and services in underground markets [1]. In particular, major products include credit card numbers, personal account credentials, scam (phishing) kits, and botnets [7]. Major services include cashing from stolen bank accounts, DDoS attacks, sending phishing emails, and fraudulent purchases using stolen cards [4]. Researchers are increasingly

[1] http://www.mcafee.com/us/resources/reports/rp-economic-impact-cybercrime2.pdf

[2] http://www.12321.cn/pdf/2015zgwmqybhdcbg.pdf

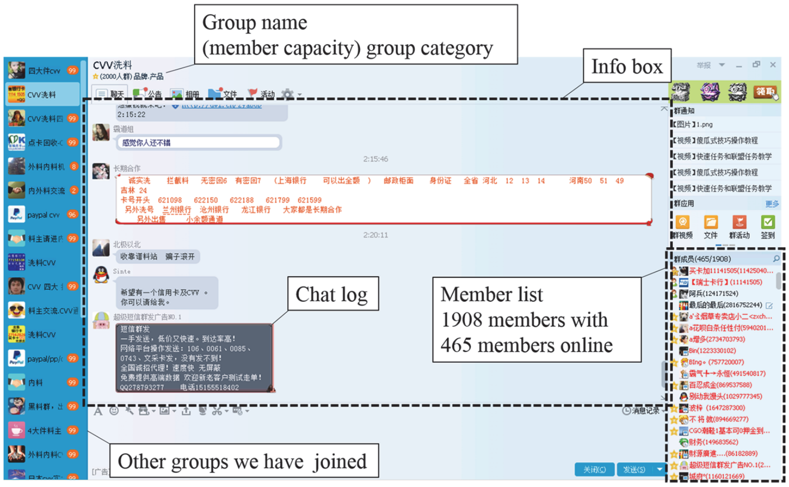
978-1-5090-3865-7/16/$31.00 ©2016 IEEE

interested in underground markets in recent years [5]. Thomas and Martin [7], Moshchuk et al. [8] and Wang et al. [9] are among the first researchers to analyze the underground economy. A wide range of topics in this field have been studied. For example, Motoyama et al. [1], Yip et al. [10], and Odabas [11] analyzed the underground markets in terms of stolen information circulating the market, social network characteristics, and market mechanisms. The identification of top sellers in underground markets using text mining is discussed in [12]. Moreover, Benjamin et al. [5] studied the hacker language and concepts, proposing a state-of-the-art language model to understand words and expressions in underground markets.

International underground markets have become a topic of significant interest. Holt et al. explored the network characteristics of Russian hacker social networks [14]. Zhuge et al. first conducted a general analysis of the Chinese cybercrime [6]. However, the research on Chinese cybercrime is still in its infancy. Chinese underground markets have facilitated a unique set of cybercrimes, including real asset theft, virtual property theft, Internet service abuse, and cybercrime training [6]. Further, Chinese underground markets differs from underground markets in other countries in terms of language usage and communication platform.

One of the major characteristics of Chinese underground market is language usage. In China, cyber language is often different from colloquial expressions and formal texts [15]. Cyber language usually include jargons made up of symbols, numerals, acronyms and new characters. Moreover, certain jargons originated from English and Pinyin homonym. In Chinese underground markets, cybercriminals heavily use jargons to describe certain products and services. As reported by Zhuge [6], jargons used by cybercriminals have been impeding cybersecurity researchers and investigators to understand hacker discussion content. For example, "料," which literally translates to "material," refers to the data of stolen cards. "洗," which literally translates to "wash," means money laundering. "四大件," which used to refer to four types of household appliances, now refers to four types of key victim information: victim's ID, bank account, password, and cellphone number. Further, "马" (which literally translates to "horse") is a homophonic Chinese character of "码" (meaning code). Consequently, cybercriminals use this character to refer to malware. In an underground criminal chain, malware writers develop and sell "horse," which is then used to steal victim's "four items." "Four items" are further "washed" into cash. Most of the previous studies do not take into account the analysis of Chinese cybercrime language.

As for communication platform, QQ group is a major platform for Chinese cybercriminals to advertise and trade malicious products and services [6]. QQ is one of the largest instant-message applications in China. QQ group is a service provided by QQ for group communication. The exclusiveness of QQ group service appeals to cybercriminals. Cybercriminals can only join the QQ group with the permission of the group moderator; moreover, the group moderator can also expel group members who violate the group rules. Because of this exclusiveness, group members usually post higher quality advertisements than in other un-moderated underground market platforms [2]. Fig. 1 shows an example QQ group. The group name is "CVV 洗料" ("CVV monetizing"). The name itself includes cybercrime jargons "CVV" and "洗料" whose semantics will be discussed later. Underground QQ group names are usually formed with such cybercrime jargons to attract experienced sellers and buyers. Moreover, the group shown in Fig.1 has the largest group capacity that QQ permits, 2,000 members, which is determined by the seniority of the group moderator as a QQ user. Right-hand side of the interface lists every member of the group. The group has 1,908 members and 465 members were online when the screenshot was taken. The communication needed for conducting criminal chain procedures are conveniently supported by QQ groups. While cybercriminal jargons have been a great challenge facing researchers and investigators, analyzing QQ group discussion allows us to find cues for interpreting jargons, and thereby better understand the underground market [3] [4]. Therefore, the chat logs of QQ groups provide valuable information for us to understand cybercriminal jargons. Given the scale and the growth of Chinese cybercrime, it is imperative to automatically extract and understand hacker jargons from hacker posts. Hence, we further review lexical semantics literature.



2:15:22

霸道姐
感觉你人还不错

2:15:46

长期合作
诚实洗　拦截料　无密回6　有密回7　（上海银行　可以出全额　）　邮政柜面　身份证　全省 河北 12 13 14　河南50 51 49
吉林 24
卡号开头 621098 622150 622188 621799 621599
另外洗号 兰州银行 沧州银行 龙江银行 大家都是长期合作
另外出售 小余额通道

2:20:11

北极以北
收靠谱料站 骗子滚开

Sinte
希望有一个信用卡及CVV 。
你可以请给我。

超级短信群发广告NO.1
短信群发
一手发送，低价又快速。到达率高！
网络平台操作发送：106、0061、0085、0743、文采卡发，没有发不到！
全国诚招代理！速度快 无屏蔽
免费提供高端数据 欢迎新老客户测试走单！
QQ278793277　电话15155518402

买卡加11141505(11425040...
【瑞士卡行】(11141505)
阿兵(124171524)
最后的最后(2816752244)
a'幺烟草专卖店小二<zxch...
a花呗白条任性付(5940201...
a增多(2734703793)
Bin(1223330102)
BIng。(757720007)
霸气十→永恒(491540817)
百忍成金(869537588)
别动我馒头(1029777345)
波梓 (1647287300)
不 将 就(894669277)
CGO潮鞋1基本可0押金到...
财务(149683562)
财源廣進....(86182889)
超级短信群发广告NO.1(2...
城府°(1160121669)

Figure 1. An Example of QQ group with hacker contents

B. *Lexical Semantics*

The lexical semantics literature studies the meanings of words. Since the lexical semantics literature is quite broad, we focus our review on promising techniques that are suitable for interpreting lexical semantics on large scale underground market data. Specifically, as underground market data usually lacks annotated text for jargons, we focus our review on unsupervised learning-based methods. The two major unsupervised learning-based methods are Neural Network Language Model (NNLM) and Latent Dirichlet Allocatoin (LDA).

NNLM aims to represent each words as a probability distribution over fixed-dimensional features. Usually, the dimensionality is much smaller than the size of the vocabulary in order to overcome the curse of dimensionality [17]. Word embedding is an NNLM technique with shallow, two-layer neural networks. Word embedding represents the semantics of the word as a word vector in k-dimensional space, whose direction is based on their semantics in the context. The similarity of the word vectors suggests the closeness of their semantics. Word embedding has the ability of additive compositionality, e.g., vector ('Paris') − vector ('France') + vector ('Italy') ≈ vector ('Rome'). There are two common word embedding model architectures: Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram Model (Skip-gram) [18]. CBOW seeks to predict the probability of a central word given the contextual words. Skip-gram seeks to predict the probability of the contextual words given the central word. Skip-gram works well on a small amount of training data, and has good performance on rare words. CBOW is much faster to train and has better accuracy for the frequent words than Skip-gram. Further, word embedding often employs two estimation algorithms: hierarchical softmax and negative sampling. Hierarchical softmax is an efficient approximation of the probability distribution method by building a Huffman tree. Negative sampling approximates the probability distribution by sampling [18].

On the other hand, LDA is a hierarchical Bayesian model that treats the corpus as a mixture over a set of latent topics [20]. Each underlying topic is further defined as a probability distribution over a set of words. Topics are usually interpreted by the top words with the highest probability weights. Hence, words with similar semantics usually have similar probability weights in the same topic. Therefore, the semantics of a word can be interpreted from other words with similar probability weights in certain topics.

C. *Research Gaps and Questions*

Based on our literature review, we find that limited research has focused on Chinese underground markets. Specifically, few studies have analyzed cybercriminal jargons in Chinese underground markets. On the other hand, NNLM is the state-of-the-art unsupervised approach for lexical semantics, but little research has explored the application in Chinese underground markets. Thus, we are motivated to analyze and compare unsupervised lexical semantics approaches for understanding cybercriminal jargons in Chinese underground markets. Specifically , we propose the following research questions:

- How can we leverage unsupervised lexical semantics approaches to understand Chinese cybercriminal jargons?
- How effective are these approaches in helping understand cybercrime jargons?

III. Research Design

In this section, we represent the framework of analyzing cybercrime jargons in Chinese underground markets. Our proposed framework includes four components: data collection, data preprocessing, model training and evaluation (Figure 2).

In data collection, we collect Chinese underground markets hosted on QQ groups through keyword searching [21] and snowball collection [22]. In particular, we first collect keywords from prior literature, such as cybercrime-related news and Chinese underground market research papers [6]. In community detection, we identify underground market communities through searching the collected keywords with QQ group search functionality. We extract the QQ group chat log after we join these groups.

Data preprocessing consists of two tasks: data scrubbing and word segmentation. Data scrubbing removes irrelevant information such as duplicate posts, log metadata, system message and picture posts. Word segmentation is a key step for Chinese text analysis. We employ the well-established Chinese Lexical Analysis System provided by Institute of Computing Techonology (ICTCLAS) for Chinese word segmentation [23].

In model training, we adopt two typical language models: word embedding and LDA. For word embedding, we interpret the semantics of the jargon by examining the words that is the closest to the jargon in the word embedding vector space. In particular, we use the Word2vec implementation of word embedding. We formally define word embedding in a given corpus C as a mapping $V \xrightarrow{C} \mathbb{R}^D : w \xrightarrow{C} \vec{w}$ that maps corpus C from vocabulary V to the D-dimensional word embedding space $\mathbb{R}^D$. Specifically, each word w is mapped to a word vector $\vec{w}$. Following previous work of Mikolov et al. [18], we use cosine similarity to measure the distance between the semantics of two words w_1 and w_2:

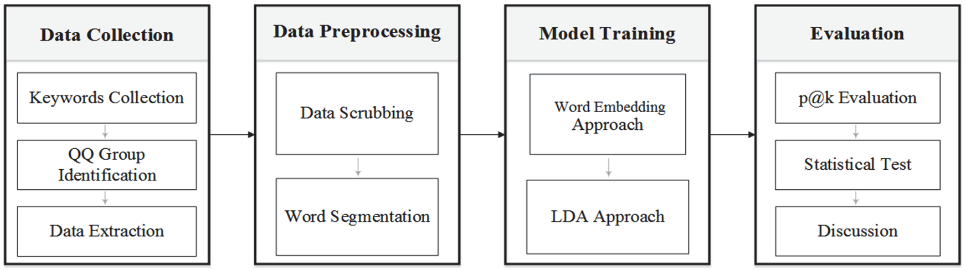


Figure 2. The Framework for Chinese Hacker Language Modeling

$$similarity(w_1, w_2) = \frac{\overrightarrow{w_1} \cdot \overrightarrow{w_2}}{\|\overrightarrow{w_1}\| \cdot \|\overrightarrow{w_2}\|}$$

Formally, let J be the set of jargons and let jargon word $w_g \in \mathrm{V} \cap \mathrm{J}$ be the *query word*. The sequence of semantically similar words $w_{(1)}^g, w_{(2)}^g, \cdots, w_{(n)}^g \in \mathrm{V}\backslash\{w_g\}$ is the ordered sequence of the words in $\mathrm{V}\backslash\{w_g\}$ based on $similarity(w_i^g, w_g)$. Further, we optimize the performance of word embedding by experimenting all four different settings with two architectures (CBOW and Skip-gram) and two learning algorithms (hierarchical softmax and negative sampling).

For LDA, we interpret the semantics of the jargon by examining the major topics that the jargon belongs to. In LDA, each document is represented with the vocabulary through a V-dimensional vector $\vec{V} = (w_1, w_2, \cdots, w_V)$. LDA generates T latent topics $\overrightarrow{\varphi_1}, \overrightarrow{\varphi_2}, \cdots, \overrightarrow{\varphi_T}$. Each topic is a probability distribution over the vocabulary $\vec{V}$: $\overrightarrow{\varphi_t} = (\varphi_{t1}, \varphi_{t2}, \cdots, \varphi_{tV})$. Similarly, we suppose that jargon word w_g is the *query word*. As mentioned in the literature review, words with similar semantics have similar probability weights in the same topic. Particularly, since φ_{ij} represents how much word w_j reflects the semantics of topic i, we pick the most similar latent topic $\overrightarrow{\varphi_{s^g}}$ such that $\mathrm{s}^g = \underset{i\in[0,K]}{\arg\max}\, \varphi_{ig}$. As a result, the sequence of semantically similar words $w_{(1)}^g, w_{(2)}^g, \cdots, w_{(n)}^g \in \mathrm{V}\backslash\{w_g\}$ is an ordered sequence of all the words in $\mathrm{V}\backslash\{w_g\}$ sorted by $\varphi_{s^g i}$. In additional, we vary the number of topics K to avoid overfitting.

Evaluating unsupervised learning models has been challenging. Ideally, evaluation is conducted by comparing the benchmark metrics on well-known testbeds. However, since Chinese cybersecurity text mining is an emerging research field, finding benchmark for our testbed can be difficult. To this end, we propose to evaluate our proposed methods using Precision-at-K (P@K). P@K measures the word similarity in terms of semantics. In particular, we first define *indicator function* based on the semantic relation:

$$I_{w_0}(w) = \begin{cases} 1, & w\ is\ semantically\ related\ to\ w_0, \\ 0, & else. \end{cases}$$

Further, P@K for *query word* w_g is defined as:

$$P@K(w_g) = \frac{\sum_{i=1}^{K} I_{w_g}\left(w_{(i)}^g\right)}{K}$$

P@K(w_g) evaluates how a set of words generated by a particular approach can explain the semantics of the *query word* w_g. To evaluate the overall performance of the approach for understanding jargons, we choose a fixed set of jargons as *query words* and calculate the average of P@K to evaluate the overall performance:

$$performance = \overline{P@K} = \frac{\sum_{i=1}^{n} P@K(w_{g_i})}{n}$$

IV. Experiments

A. Research Testbed

We collected the chat logs from underground QQ groups to extract and analyze jargons. We first searched group names with commonly used cybercriminal lexicons from prior literature and requested to join the groups. We managed to join 29 QQ groups. The total number of members was 23,000. Since cybercriminals usually joined multiple groups, we removed duplicated members across groups. Table I shows the statistics of the major QQ groups. The underground QQ groups seemed quite active because most of the groups we joined were nearly full.

TABLE I. Matadata of Underground QQ Groups

Group id	Group name	Total member	Member capacity
518270961	四大行 (four big banks)	1,885	2,000
372779882	外料内料 CVV 等交流 (exchange foreign and domestic materials, CVV)	609	1,000
472812351	洗料 CVV (money laundering, CVV)	1,885	2,000
518471495	洗料 CVV (money laundering, CVV)	1,885	2,000
76649054	CVV 洗料四大 (CVV, money laundering and four items)	1,926	2,000
426176059	国内 CVV (domestic CVV)	2,000	2,000
196653656	CVV 四大 拦截料 (CVV, four items, intercept materials)	1,902	2,000
517530328	CVV 洗料 (CVV, money laundering)	1,908	2,000
197313973	点卡回收-CVV 交流 (recycle click card, exchange CVV)	984	1,000
484681593	WEBshell 黑产交易群 (WEBshell black-market goods transaction group)	869	1,000

Overall, we collected more than 90,000 text posts with 18,800 unique posts, most of which are advertisements. We use the advertisement from the group member “长期合作” (translated as “Long Term Cooperation”) as an example. The jargons are highlighted in bold. This member posted, “诚实洗拦截料 无密回 6 有密回 7 (上海银行 可以出全额)…卡号开头 621098 622150…” (translated as “Honestly looking for intercept **materials** to **wash**. Reply 6 for **materials** without password, reply 7 for **materials** with password (Shanghai Bank, **full account** is available)…The Bank cards numbers start with 621098 622150…”). In our research testbed, we can find various jargons describing cybercrime products and services.

B. Evaluation and Discussions

We performed the word similarity experiments to evaluate the effectiveness of our proposed method in understanding jargons as described in our research design. Specifically, we evaluated the performance of the four settings of Word2vec and the LDA approach on our underground market QQ group testbed. We selected 70 typical Chinese underground jargons (as much as possible) and listed 10 most similar words generated by each approach. We then manually calculated the P@10 metric for each jargon by checking if the retrieve words are semantically related to the input one. We averaged all P@10 metrics for each approach. A number of selected jargons have similar semantics so that we often can use one to explain another.

We adjusted the parameter such as window size to make sure each approach has the best configuration. Table II shows two examples of our Word2vec experiment with CBOW+NS setting.

"洗料" ("material washing") and "四大件" ("four items") are two jargon words we selected as *query words* w_g. We sorted top 10 similar terms based on their cosine similarity. These are the output of Word2vec with CBOW+NS setting. Semantically related words are bold and italic in the table. The semantics of "洗料" ("material washing") is monetizing stolen data. We expected the output of our approach would include relevant words describing data or methods. Among the 10 most similar terms, "出料" ("material selling") is the synonym. "机子" ("machine/terminal") and "基站" ("base station") describe the method to monetize data. "拦截料" ("intercept material") is a crucial type of stolen data. They all have high semantical relevance with the input jargon.

TABLE II. EVALUATION OF JARGONS

	洗料 (material washing)		四大件 (four items)	
	Term	Similarity	Term	Similarity
1	***出料 (material selling)***	0.916172	***广发 (Guangfa Bank)***	0.892864
2	***机子 (machine/ terminal)***	0.877228	***四大 (four items)***	0.868508
3	试单 (try order)	0.869973	***4 大 (4 items)***	0.865067
4	有意者 (interested person)	0.869257	大额 (big amount)	0.857389
5	***基站 (base station)***	0.866542	二十万 (two hundred thousand)	0.855467
6	航空机票 (airline ticket)	0.858605	***广发四大 (four items of Guangfa Bank)***	0.854781
7	者 (man)	0.850602	***交通 (Bank of Communications)***	0.849424
8	***拦截料 (intercept material)***	0.849222	***民生 (Minsheng Bank)***	0.838702
9	寻找 (find)	0.844255	***招商 (China Merchants Bank)***	0.833639
10	三五万 (thirty/fifty thousand)	0.839619	附近 (nearby)	0.833082
P@10	40%		70%	

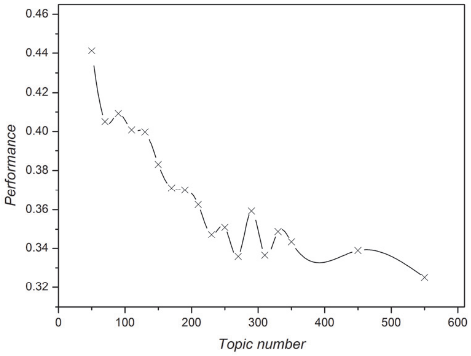


Figure 3. P@10 for LDA over topic number

The term "四大件" ("four items") refers to four types of stolen personal information, including bank accounts. The returned terms include many bank names along with other acronyms of "four items". We calculate the P@10 by manually counting the number of semantically related terms. In this setting, "洗料" has 4 relevant terms out of 10 outputs, thus getting 40% as its Precision-at-10 (P@10). "四大件" produces the result with 70% P@10.

We continued with the remaining 68 jargons and tested on all the four settings of Word2vec and the LDA approach. We used one-tailed paired sample t test to compare different approaches. The overall performance are shown in Table III. All four settings of Word2vec performs significantly better than the LDA approach. The best setting of Word2vec is CBOW+NS, which is nearly 20% higher than the LDA approach. We explored different combination of architectures and algorithms in the Word2vec approach. In Table III, CBOW significantly outperforms skip-gram for both architectures, and negative sample algorithm enlarges the gap. Overall, the performance ranking of the four settings of the Word2vec approach follows CBOW+NS > CBOW+HS > SG+HS > SG+NS.

For the LDA approach, the topic number parameter T is a key value for the performance. We tested for different numbers of topics (Fig.3). In general, having fewer topics helps increase the performance. This is because the number of topics in underground QQ group discussion is not huge; therefore, having too many topics would not help increase the performance of the understanding of *query word*.

TABLE III. P@10 RESULTS OF DIFFERENT APPROACHES WITH TEN REPEATED TESTS

Approach			**P@10**	**P-value**
Word2vec	Hierarchical Softmax (HS)	CBOW	**57.91%*** **	0.000172694
		Skip-gram	55.56%	
	Negative Sampling (NS)	CBOW	**60.56%*** **	<0.0001
		Skip-gram	50.59%	
Topic Model	LDA (90 topics)		40.9%	-----

[a] *p<0.05; **p<0.01; ***p<0.001

V. Conclusions

This paper represents our proposed research framework on automatically understanding Chinese cybercriminal jargons in large-scale cybercriminal posts. We incorporate two unsupervised lexical semantics approaches: Word2vec and LDA. We evaluated the ability of Word2vec and LDA to understand jargon words in textual posts of underground market QQ groups. Specifically, we tested different settings of each approach to ensure the best performance. Result showed that Chinese cybercriminal jargons can be well understood by our proposed approach. The proposed approach would benefit underground market researchers and further be helpful to secure China's e-commerce environment.

Acknowledgment

This work was supported by the National High-Tech R&D Program of China (Grant No. SS2015AA020102), the 1000-Talent program, Tsinghua University Initiative Scientific Research Program.

References

[1] Motoyama, M., McCoy, D., Levchenko, K., Savage, S., & Voelker, G. M. (2011, November). An analysis of underground forums. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference (pp. 71-80). ACM.

[2] Herley, C., & Florêncio, D. (2010). Nobody sells gold for the price of silver: Dishonesty, uncertainty and the underground economy. In Economics of Information Security and Privacy (pp. 33-53). Springer US.

[3] Fossi, M., Johnson, E., Turner, D., Mack, T., Blackbird, J., McKinney, D., ... & Gough, J. (2008). Symantec report on the underground economy. Symantec Corporation.

[4] Franklin, J., Perrig, A., Paxson, V., & Savage, S. (2007, October). An inquiry into the nature and causes of the wealth of internet miscreants. In ACM conference on Computer and communications security (pp. 375-388).

[5] Benjamin, V., Li, W., Holt, T., & Chen, H. (2015, May). Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on (pp. 85-90). IEEE.

[6] Jianwei, Z., Liang, G., & Haixin, D. (2012, July). Investigating China's online underground economy. In Conference on the Political Economy of Information Security in China.

[7] Thomas, R., & Martin, J. (2006). The underground economy: priceless. ; login:: the magazine of USENIX & SAGE, 31 (6), 7-16.

[8] Moshchuk, A., Bragin, T., Gribble, S. D., & Levy, H. M. (2006, February). A Crawler-based Study of Spyware in the Web. In NDSS (Vol. 1, p. 2).

[9] Wang, Y. M., Beck, D., Jiang, X., Roussev, R., Verbowski, C., Chen, S., & King, S. (2006, February). Automated web patrol with strider honeymonkeys. In Proceedings of the 2006 Network and Distributed System Security Symposium (pp. 35-49).

[10] Yip, M., Shadbolt, N., & Webber, C. (2013, May). Why forums?: an empirical analysis into the facilitating factors of carding forums. In Proceedings of the 5th Annual ACM Web Science Conference (pp. 453-462). ACM.

[11] Odabas, M. (2015, July). Toward an Economic Sociology of Online Hacker Communities. In 27th Annual Meeting. Sase.

[12] Li, W., & Chen, H. (2014, September). Identifying top sellers in underground economy using deep learning-based sentiment analysis. In Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint (pp. 64-67). IEEE.

[13] Lau, R.Y., Xia, Y. and Li, C., 2012. Social Media Analytics for Cyber Attack Forensic. International Journal of Research in Engineering and Technology (IJRET), 1 (4), pp.217-220.

[14] Holt, T. J., Strumsky, D., Smirnova, O., & Kilger, M. (2012). Examining the social networks of malware writers and hackers. International Journal of Cyber Criminology, 6 (1), 891.

[15] Hong-cheng D. Nature of Cyber Language and Its Semantically Irrational Tendency [J]. Journal of Hefei University of Technology (Social Sciences), 2008, 3: 029.

[16] Castellví, M. T. C., Bagot, R. E., & Palatresi, J. V. (2001). Automatic term detection: A review of current systems. Recent advances in computational terminology, 2, 53-88.

[17] Bengio, Y. (2008). Neural net language models. Scholarpedia, 3 (1), 3881.

[18] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[19] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781, 2013.

[20] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. the Journal of machine Learning research, 3, pp.993-1022.

[21] Fallmann, H., Wondracek, G., & Platzer, C. (2010). Covertly probing underground economy marketplaces (pp. 101-110). Springer Berlin Heidelberg.

[22] Agichtein, E., Brill, E., & Dumais, S. (2006, August). Improving web search ranking by incorporating user behavior information. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 19-26). ACM.

[23] Big Data Search and Mining Lab, BIT.: Natural Language Processing and Information Retrieval Sharing Platform. http://ictclas.nlpir.org/