Latent Text Mining for Cybercrime Forensics

Raymond Y. K. Lau and Yunqing Xia

***Abstract*—Recent research reveals that the number of cyber-attacks has been doubled in the past three years. This is a devastating growth of the number of cyber-attacks, and it reveals a serious business problem around the world. Existing intrusion detection systems (IDSs), intrusion prevention systems (IPSs), and anti-malware systems mainly rely on low-level network traffic features or program code signatures to detect cyber-attacks. However, since hackers can constantly change their attack tactics by, it is extremely difficult for existing security solutions to detect cyber-attacks. There are increasing more evidences showing that cybercriminals tend to exchange cybercrime knowledge and transact via online social media. Accordingly, it presents unprecedented opportunities for security intelligence experts to tap into online social media to extract the vital security intelligence for cyber-attack forensics. The main contributions of this paper are the design, development, and evaluation of a Latent Dirichlet Allocation (LDA)-based latent text mining model for cyber-attack forensics. Our preliminary evaluation of the proposed latent text mining model based on a real-world data set crawled from Twitter and Blog sites shows that it significantly outperforms the probabilistic latent semantic indexing (pLSI) method in terms of extracting more relevant and richer concepts describing real-world cyber-attack incidents.**

***Index Terms*—Text mining, latent dirichlet allocation, cyber-attacks, cyber forensics.**

I. Introduction

THE recent cybercrime study performed by Hewlett-Packard shows that there is a 42% increase in the number of cyber-attacks when compared to the previous year, with organizations experiencing an average of 102 successful attacks per week. The average annualized cost of cybercrime climbed to $8.9 million per attacked organization. A previous study also showed that security breaches may cost a corporation as much as 1% of its market value [1]. The recent cyber-attacks against HSBC, the major commercial banks such as Bank of America, JPMorgan Chase, Citi Bank, and the New York Stock Exchange reveal that this is a very serious business problem that threatens the daily operation and the shareholder value of any click-and-mortar organizations. The devastating cyber-attack figures reported by HP in 2012 reveal that existing cyber security technologies such as intrusion detection systems (IDSs), intrusion prevention systems (IPSs), and anti-malware systems are not effective to protect organizations from various cybercrimes, particularly distributed denial of service (DDoS) attacks. One of the main reasons is that existing cyber security solutions solely rely on low-level network traffic features or software coding signatures to identify potential cybercrimes or cyber-attacks. Since cybercriminals can constantly change their tactics (e.g., by using an entirely different set of botnets), and hence the constantly changing low-level features, it is extremely difficult for existing cyber security solutions to detect cybercrimes. Moreover, existing IDS, IPSs, and other malware detection software cannot predict cyber-attacks before these attacks are launched, nor are they equipped with the capability to analyze who launches the attacks and why these attacks are initiated, that is, cybercrime forensics.

There are increasingly more evidences to show that cybercriminals tend to exchange cybercrime knowledge, or even transacting cyber-attack tools such as Botnets through the "dark markets" established in social media on the Internet [2]-[4]. Such a trend gives rise to unprecedented opportunities for cyber security analysts and researchers to tap into the security intelligence embedded in online social media for better understanding the activities of cybercriminals and for semi-automated cyber-attack forensics. Fig. 1 shows a screen shot of the announcement released by the hacker group "Anonymous" at Twitter while the servers of HSBC were being attacked in October 2012. Social media analytics is defined as the design, development, and evaluation of informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application [5]-[6]. Social media analytics has been explored in the areas of terrorist intelligence mining [7]-[9], financial investment prediction [10]-[12], and market intelligence mining [13]-[14], and so on. However, social media analytics and text mining for cyber-attack forensic has received little attention by researchers to date despite its great potentials of mining high-level features to augment existing cyber security intelligence technologies such as IDSs and IPSs for more effective cyber-attack detection, analysis, and prevention.

The main contributions of this paper is the design, development, and evaluation of a novel LDA-based text mining model for mining latent security intelligence from online social media to facilitate cyber-attack forensics. More specifically, the novelty and the technological innovations of our research work reported in this paper are as follows:

The proposed latent text mining model enhances existing IPSs in terms of mining latent features and relationships from social media to perform cyber forensics such as, automated construction and visualization of cybercriminal communities and the extraction of causal relationships of cybercrimes;

By leveraging high-level features and relationships mined from online social media, the proposed latent text mining model can predict a variety of cyber-attacks such as DDoS,

Manuscript received October 10, 2012; revised November 10, 2012.

Raymond Y. K. Lau is with the City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong SAR (e-mail: raylau@cityu.edu.hk).

Yunqing Xia is with Centre for Speech and Language Technologies, Tsinghua University, Beijing 100084, China (e-mail: yqxia@tsinghua.edu.cn).

DOI: 10.7763/IJFCC.2013.V2.187

malware, phishing, worms, and so on rather than only a specific kind of cyber-attack;

By mining the security intelligence embedded in online social media, not only cyber-attack forensics is facilitated but also the intelligence has a potential to be used to predict cyber-attacks before they are actually launched.

The remainder of the paper is organized as follows. Section II highlights the system architecture of the proposed service. Section III illustrates the proposed computational model of latent text mining for cybercrime forensics. Section IV reports our preliminary experiment and the result of the proposed latent text mining model. Finally, we offer concluding remarks and describe future directions of our research work.

Fig. 1. Announcement of a series of cyber-attacks Against HSBC at twitter on 18 OCTOBER 2012

II. General System Architecture

The proposed service consists of five layers. For the Data Collection Layer, autonomous information agents simulate human users and tap into private (requiring user authentication) multilingual online social media to observe and collect cyber-attack related messages. These agents leverage a machine translation module to process multilingual messages. In addition, dedicated crawlers can visit various web sites (through search engine indexes) and public multilingual social media to retrieve cyber-attack messages. For the Data Cleaning Layer, a text pre-processor is applied to pre-processing multilingual textual data such as word segmentation, stop word removal, stemming, and part-of-speech (POS) tagging, etc. For the Data Mining Layer, a linguistic feature extractor makes use of a set of pre-defined feature patterns and name entity recognition (NER) rules to extract explicit features from pre-processed messages. Moreover, an LDA-based latent feature miner is applied to mine cyber-attack related concepts from texts. One novelty of the our research is the design of a novel LDA-based latent relationship miner which employs a bootstrapping approach to discover domain-specific relationship indicators for hacker relationships discovery and cyber-attack causal relationship mining. Finally, the cyber-attack analyzer utilizes an information theoretic approach to identify the inherently associated cyber-attack events or people for the generation cyber-attack communities and causal relationships of attacks.

For the Learning and Classification Layer, an ensemble of state-of-the-art machine learning classifiers are deployed to produce cyber-attack predictions based on an optimal set of explicit and latent features. Furthermore, another novelty of our research is that the GA-based adaptor takes into account the feedbacks about the classifiers' previous predictions, and periodically tunes the "weights" of the classifiers and the classification features for more effective cyber-attack predictions in the following prediction cycles. Moreover, the adaptor evaluates which online social media sources tend to produce the classification features with higher weighting, and then direct the information agents and crawlers to traverse to those sources with a higher probability in the following data collection cycles. As a result, the proposed prototype service can maintain its forensic and prediction performance even if cyber attackers modify their attack tactics, and so the changes of explicit or latent traces left in online social media. For the Presentation Layer, the presentation manager employs a combination of texts, tables, and graphs to present the cyber-attack prediction results, and visualize the causal relationships of cyber-attacks.

III. The Computational Model for Cyber-Attack Analysis

For the proposed text mining model for cyber-attack forensics, we apply Latent Dirichlet Allocation (LDA) [15] as the computational foundation to develop such a model to extract high-level concepts from hacker forums or other online social media to augment existing IDSs and IPSs. We do not intend to make a claim that the proposed model can replace existing IDSs and IPSs; instead the proposed model is a vital complimentary tool to existing security intelligence solutions so that both high-level and low-level features can be used to better combat cybercrimes. Essentially, the LDA model can be seen as a conceptual clustering method which automatically groups semantically related terms together to form high-level cybercrime related features (concepts) according to the nature of the given data set.

For the LDA model, a set of latent features A is assumed to be associated with a collection of cybercrime related messages (documents) D. Each document (message) $d \in D$ is characterized by a multinomial distribution θ, which is controlled by a hyper-parameter α [15]. A latent feature a (i.e., a cybercrime related high-level concept) is then selected according to the multinomial distribution θ controlled by another hyper-parameter β. Given a feature a, a term t (e.g., word, phrase, bigram, trigram, etc.) is then generated according to the multinomial distribution ϕ. Plate notation [15] is used to describe this probabilistic generation process in Fig. 2. For the plate notation, shaded and unshaved circles indicate observed and latent variables, respectively. Plates (i.e., the rectangles) indicate repeated sampling, with the number of iterations defined by the variable shown at the bottom of each plate.

According to LDA, the likelihood of generating a cybercrime message collection D can be estimated as follows.

$$P(D \mid \alpha, \beta) = \iint \prod_{a \in A} P(\phi_a \mid \beta) \prod_{d \in D} P(\theta_d \mid \alpha) \left[\prod_{t \in d} \sum_{a \in A} P(a \mid \theta_d) P(t \mid a, \phi_a) \right] d\theta_d d\phi_a \quad (1)$$

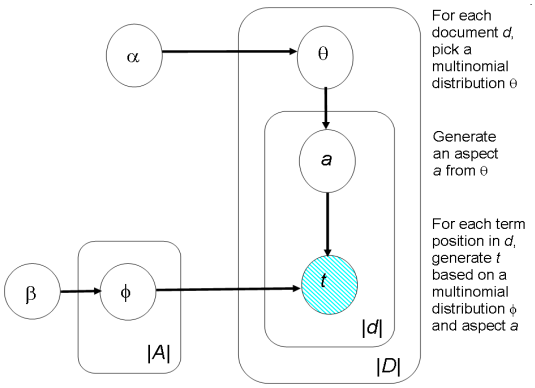

Fig. 2. The LDA generation process of cyber-attack messages

Our ultimate goal is to obtain the conditional $P(a_i \mid t_i, d)$, that is, given a term t_i in a message d, which latent feature (i.e., a cybercrime concept) generates the term in that position. The proposed latent feature mining method is context-sensitive**,** as the conditional $P(a_i \mid t_i, d)$ is derived from a training message corpus D that varies across contexts (e.g., messages pertaining to different kinds of cybercrimes). As it is computationally expensive to directly estimate the conditional $P(a_i \mid t_i, d)$, an approximation method of Gibbs sampling [16], a Markov chain Monte Carlo algorithm, has been applied to estimate the conditional probabilities of LDA-based models [17]. For Gibbs sampling, a Markov chain (a model of transitions among successive states) is established by repeatedly drawing a latent feature for each observable term, based on its conditional probability over other variables [17]. According to Gibbs approximation, the conditional probability $P(a_i \mid t_i, d)$ is approximated by the following:

$$\begin{aligned} &P(a_{\ i} = k \mid t_i = m, d) \propto \\ &P(t_i = m \mid a_i = k) \bullet P(a_{\ i} = k \mid d) \propto \\ &\frac{C_{mk}^{VA} + \beta}{\sum_{m' \in V} C_{m'k}^{VA} + |V|\beta} \bullet \frac{C_k^d + \alpha}{\sum_{k' \in A} C_{k'}^d + |A|\alpha} \end{aligned} \quad (2)$$

where C_{mk}^{VA} is the count of the number of times that term $t_i = m$ is assigned to feature $a_i = k$, excluding the current term; V is the vocabulary set that is used to compose the collection D, and C_k^d is the count of the number of terms in d associated with feature $a_i = k$, excluding the current term.

During model learning, the count matrix $M^{VA} : V \times A$ (term by feature) is maintained and updated in each Gibbs iteration. Based on the count matrix M^{VA}, it is feasible to derive the approximations $\bar{\phi}$ and $\bar{\theta}$ of the multinomial distributions ϕ and θ. The computational complexity of the standard Gibbs sampling algorithm is characterized by $O(I \bullet |A| \bullet |V| \bullet |D|)$, where I is the number of Gibbs iterations; V is the vocabulary set of a corpus D, and $|A|$ is the pre-defined number of high-level features.

IV. Experiment and Result

To evaluate the effectiveness of the proposed latent text mining model, we crawled the cybercrime related messages from Twitter and blogs. We manually identified a group of 20 popular hackers and then identified the twitter feeds (e.g., the "Anonymous", also known as "Fawkes Security" who claimed to be responsible for the series of cyber-attacks against HSBC in October 2012 at Twitter) and the corresponding web sites (e.g., pastebin.com) for later crawling by our dedicated crawler programs. A total of 12,136 cybercrime related messages covering the period from October 2011 to October 2012 were crawled in October 2012. A subset of 4,015 messages from our total downloaded messages was manually inspected by two human experts and the cyber-attack related high-level concepts were annotated by them. This subset of human-annotated messages was applied to evaluate the effectiveness of the proposed latent text mining model. We extend the open source Core Ling Pipe API to implement our revised Gibbs sampling algorithm for the implementation of the proposed LDA model.

TABLE I. Top two Cyber-Attack Related Concepts

*T*1	Probability	*T*2	Probability
collaborative	0.0274	worm	0.0192
attacks	0.0255	virus	0.0187
denial	0.0241	malicious	0.0181
service	0.0235	malware	0.0165
networks	0.0233	attacks	0.0159
banks	0.0227	spread	0.0143
finances	0.0205	pc	0.0141
hate	0.0196	infected	0.0139
muslim	0.0153	replicate	0.0126
insults	0.0142	assets	0.0115

Table I depicts two representatives of high-level features (topics) mined by applying the proposed latent text mining to our experimental data set. For readability reason, we demonstrate the non-stemmed version of the results even though both stemmed and non-stemmed versions of feature mining are supported by our prototype system. It should be noted that these features are semantically coherent and they are all related to prominent cyber-attack related issues such as the reason behind the series of DDoS attacks against the major commercial banks and stock exchanges in the U.S. in September 2012 (Topic 1).

To assess the effectiveness of the proposed LDA-based latent text mining method, we invited two security intelligence experts to examine the high-level features generated from the proposed text mining model. In addition, we implemented a baseline system using the pLSI model that we have successfully applied to intelligent query expansion before [18]. Then, we requested the two experts to examine the latent concepts mined by both systems. In particular, a

questionnaire capturing six dimensions including "relevance", "semantic richness", "semantic coherence", "completeness", "clarity", and "interestingness" are answered by both experts. For each dimension, a score in a semantic scale between 1 and 10 was assigned by each expert. The average scores across the six dimensions achieved by the experimental system and the baseline system are depicted in Table II.

TABLE II: COMPARATIVE PERFORMANCE OF CONCEPT MINING

Dimension	LDA	pLSI
relevance	8.8	8.3
semantic richness	8.2	8.1
semantic coherence	8.5	8.2
completeness	7.9	7.8
clarity	8.1	7.6
interestingness	9.1	8.3

It is clear that the cyber-attack concepts mined by the LDA-based system consistently attract higher scores across the six dimensions when compared to the concepts produced by the pLSI system. According to our paired one-tail *t*-test ($p < .01$), the LDA-based experimental system is significantly better than the pLSI baseline system. The main reason of such a performance improvement achieved by the LDA-based system is that it can more effectively cluster semantically related terms under each latent concept. As a result, the human experts find the outputs produced by the LDA-based system easier to follow and more relevant with respect to the cyber-attack domain. Accordingly, our preliminary experiment confirms the effectiveness of the proposed latent text mining method for cyber-attack forensics.

V. Conclusions and Future Work

Recent research shows that there is a rapid increase of the number of cyber-attacks and the financial loss brought to the click-and-mortar organizations. The main contributions of the work reported in this paper are the design, development, and evaluation of a novel LDA-based text mining model to combat cybercrimes and facilitate cybercrime forensics. The proposed model can be seen as a complimentary tool to existing intrusion detection systems, intrusion prevention systems, and anti-malware systems to improve the overall cyber security by leveraging both low-level and high-level detection and forensics features. Our preliminary experimental result shows that the proposed model can discover semantically rich and relevant latent concepts related to cyber-attacks, and it significantly outperform the pLSI based baseline method based on a real-world data set we crawled from Twitter and websites capturing cybercrime related messages.

Future work involves a much larger scale of experiments to quantitatively evaluate the effectiveness and efficiency of the proposed latent text mining model. Moreover, alternative latent text mining methods such as supervised LDA and sequential pattern mining will be examined and compared with the performance of the current un-supervised LDA-based method. Also, the application of the mined cyber-attack related high-level features to cyber-attack prediction and prevention will be studied. Finally, how to integrate the proposed model into existing IPSs to enhance the overall effectiveness of the whole system will be examined.

References

[1] S. Goel and H. A. Shawky, "Estimating the market impact of security breach announcements on firm values," *Information and Management*, vol. 46, pp. 404-410, 2009.

[2] P. J. Denning and D. E. Denning, "The profession of IT: discussing cyber attack," *Communications of the ACM*, vol. 53, no. 9, pp. 29-31, 2010.

[3] J. Franklin, A. Perrig, and V. Paxson, S. Savage, "An inquiry into the nature and causes of the wealth of internet miscreants," in *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 375-388, 2007.

[4] S. Goel, "Cyberwarfare: connecting the dots in cyber intelligence," *Communications of the ACM*, vol. 54, no. 8, pp.132-140, 2011.

[5] D. Zeng and H. Chen, R. Lusch, and S. H. Li, "Social Media Analytics and Intelligence," *IEEE Intelligent Systems*, vol. 25, no. 6, pp. 13-16, 2010.

[6] J. Leskovec, "Social media analytics: tracking, modeling and predicting the flow of information through networks," in *Proceedings of the World Wide Web Conference*, pp. 277-278, 2011.

[7] A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect Analysis of Web Forums and Blogs Using Correlation Ensembles," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 9, pp. 1168-1180, 2008.

[8] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums," *ACM Transactions on Information Systems*, vol. 26, no. 3, 2008.

[9] Y. Zhou, E. Reid, J. Qin, and H. Chen, and G. Lai, "US Domestic Extremist Groups on the Web: Link and Content Analysis," *IEEE Intelligent Systems*, vol. 20, no. 5, pp. 44-51, 2005.

[10] J. Bolle, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1-8, 2011.

[11] S. Das, "The Finance Web: Internet Information and Markets," *IEEE Intelligent Systems*, vol. 25, no. 2, pp. 74-78, 2010.

[12] R. Y. K. Lau, S. S. Liao, K. F. Wong, and C. W. Chiu, "Web 2.0 Environmental Scanning and Adaptive Decision Support for Business Merger and Acquisition," *MIS Quarterly*, vol. 36, no. 4, pp. 1239-1268, 2012.

[13] N. Archak, A. Ghose, and P. Ipeirotis, "Deriving the pricing power of product features by mining consumer reviews," *Management Science* vol. 57, no. 8, pp. 1485-1509, 2011.

[14] M. Trusov, A. Bodapati, and R. Bucklin, "Determining influential users in internet social networks," *Journal of Marketing Research*, vol. 47, no. 4, pp. 643-658, 2010.

[15] D. M. Blei, A.Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[16] S. Geman, and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Relation of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721-741, 1984.

[17] M. Steyvers, P. Smyth, M. R. Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.306-315, 2004.

[18] Song, D., Huang, Q., Bruza, and P. D. Lau, "An Aspect Query Language Model Based on Query Decomposition and High-Order Contextual Term Association," *Computational Intelligence*, vol. 28, no. 1, pp. 1-23.

Raymond Y. K. Lau is an Assistant Professor in the Department of Information Systems at City University of Hong Kong. He has worked at the academia and the ICT industry for over twenty years. He is the author of more than 100 refereed international journals and conference papers. His research work has been published in renowned journals such as ACM Transactions on Information Systems, IEEE Transactions on Knowledge and Data Engineering, IEEE Internet Computing, Journal of MIS, Decision Support Systems, etc. His research interests include Information Retrieval, Text Mining, and Agent-Mediated e-Commerce. He is the associate editor of the International Journal of Systems and Service-Oriented Engineering. He is a senior member of the IEEE and the ACM respectively.