A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media

CRIME SCENE CRIME SCENE CRIME SCENE
CRIME SCENE CRIME SCENE CRIME SCENE

© IMAGE 100 LTD.

Raymond Y.K. Lau
Department of Information Systems,
City University of Hong Kong, HONG KONG SAR

Yunqing Xia
Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, CHINA

Yunming Ye
Shenzhen Key Laboratory of Internet Information
Collaboration, Shenzhen Graduate School,
Harbin Institute of Technology, Shenzhen 518055, CHINA

Digital Object Identifier 10.1109/MCI.2013.2291689
Date of publication: 14 January 2014

Abstract—There has been a rapid growth in the number of cybercrimes that cause tremendous financial loss to organizations. Recent studies reveal that cybercriminals tend to collaborate or even transact cyber-attack tools via the "dark markets" established in online social media. Accordingly, it presents unprecedented opportunities for researchers to tap into these underground cybercriminal communities to develop better insights about collaborative cybercrime activities so as to combat the ever increasing number of cybercrimes. The main contribution of this paper is the development of a novel weakly supervised cybercriminal network mining method to facilitate cybercrime forensics. In particular, the proposed method is underpinned by a probabilistic generative model enhanced by a novel context-sensitive Gibbs sampling algorithm. Evaluated based on two social media corpora, our experimental results reveal that the proposed method significantly outperforms the

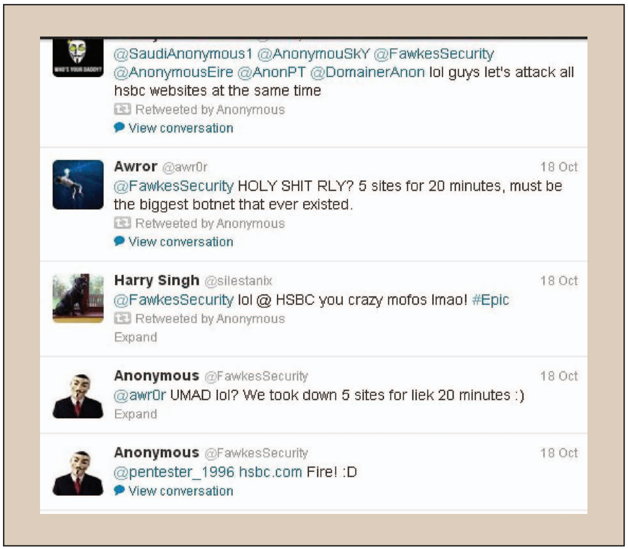
1556-603X/14/$31.00©2014IEEE

Latent Dirichlet Allocation (LDA) based method and the Support Vector Machine (SVM) based method by 5.23% and 16.62% in terms of Area Under the ROC Curve (AUC), respectively. It also achieves comparable performance as the state-of-the-art Partially Labeled Dirichlet Allocation (PLDA) method. To the best of our knowledge, this is the first successful research of applying a probabilistic generative model to mine cybercriminal networks from online social media.

I. Introduction

According to Hewlett-Packard's cybercrime report released in 2012, there was a 42% growth in the number of cybercrimes when compared to the figure of the previous year, with organizations experiencing an average of 102 successful attacks per week.[1] The average annualized cost of cybercrime also surged to 8.9 million per attacked organization. Existing cyber-security technologies such as intrusion prevention systems (IPSs) and anti-malware software are not effective enough to protect organizations from various cybercrimes, particularly the distributed denial of service (DDoS) attacks. One of the reasons is that existing cyber-security solutions are weak in cybercrime forensics and predictions. We believe that applying advanced computational intelligence methods to uncover the underground cybercriminal networks is the first step toward comprehensive cybercrime forensics, which contributes to combat the rapidly growing trend of cybercrimes.

Increasingly more evidences have shown that cybercriminals tend to exchange cyber-attack knowledge, or even transact cyber-attack tools such as Botnets through the *dark markets* established in online social media (Denning and Denning, 2010; Franklin et al., 2007; Goel, 2011). Such a trend offers unprecedented opportunities for cyber-security analysts and researchers to tap into the security intelligence embedded in the conversational messages posted to online social media so as to develop better insights about the collaborative cybercrimes and the communities of cybercriminals. By means of automated mining of collaborative cybercriminal networks, the effectiveness and efficiency of cybercrime forensics can be greatly enhanced. Figures 1 and 2 highlight examples of cybercriminals' dialogs captured in online social media. Figure 1 shows the hacker group called *Anonymous* (also known as *FawkesSecurity*) who performed a live broadcast on Twitter and claimed responsibility for the series of DDoS attacks against Hong Kong and Shanghai Banking Corporation (HSBC) in October 2012. Figure 2 depicts a dialog about the sales of cyber-attack tools on a popular Internet forum.



@SaudiAnonymous1 @AnonymouSKY @FawkesSecurity @AnonymousEire @AnonPT @DomainerAnon lol guys let's attack all hsbc websites at the same time
Retweeted by Anonymous
View conversation

Awror @awr0r 18 Oct
@FawkesSecurity HOLY SHIT RLY? 5 sites for 20 minutes, must be the biggest botnet that ever existed.
Retweeted by Anonymous
View conversation

Harry Singh @silestanix 18 Oct
@FawkesSecurity lol @ HSBC you crazy mofos lmao! #Epic
Retweeted by Anonymous
Expand

Anonymous @FawkesSecurity 18 Oct
@awr0r UMAD lol? We took down 5 sites for liek 20 minutes :)
Expand

Anonymous @FawkesSecurity 18 Oct
@pentester_1996 hsbc.com Fire! :D
View conversation

FIGURE 1 The exchange of knowledge between cybercriminals on Twitter.

While much research efforts have been devoted to social community mining and social network analysis in the past two decades, little work is performed for the automated discovery and analysis of cybercriminal networks, a special type of social network. Given the tremendous financial losses incurred due to cybercrimes, there is a pressing need to exploit advanced computational intelligence approaches for the development of automated cybercriminal network mining method to facilitate cybercrime forensics. Existing network mining methods mainly utilize manually constructed relationship lexicons (Xia et al., 2013), or manually defined lexico-syntactic patterns (Bao et al., 2008; Li et al., 2006) to uncover the implicit relationships of entities (e.g., companies) from free texts. However, since natural languages are ambiguous and very flexible, pre-defined lexicons or lexico-syntactic patterns can only identify a limited number of explicit relationships embedded in texts. As a result, the recall achieved by lexicon-based network mining methods tends to be low.

Since concept-level approaches to natural language processing (Cambria et al., 2013) can better grasp the implicit semantics associated with text, they often outperform lexicon-based methods by leveraging on semantic networks or knowledge bases. However, so far concept-level approaches have only been applied to tasks that are loosely related to cybercrimes, e.g., troll filtering (Cambria et al., 2010) and cyberbulling (Dinakar et al., 2012). A possible reason for this might be that it is difficult to mine concept-level knowledge for the cybercrime domain.

Supervised machine learning methods do not rely heavily on external knowledge and, hence, could represent a viable solution for cybercriminal network mining. However, such methods often require a lot of time and resources to build a dataset that is good enough to effectively train a machine learning classifier. Moreover, it is quite difficult to label messages that capture *implicit* and *hidden* cybercriminal relationships even if the annotators are security experts. The following two examples reveal an *explicit collaborative* and an *implicit transactional* cybercriminal relationships embedded in two messages extracted from online social media. The former is relatively easy to identify by computers using pre-defined relationship indicators such as *join* captured in a relationship lexicon, whereas the latter is more difficult for computers to detect automatically.

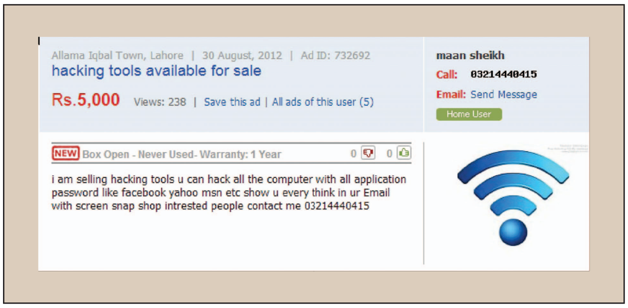
[1] http://www.hpenterprisesecurity.com/ponemon-2012-cost-of-cyber-crime-study-reports

- Explicit Collaborative Relationship: "Wanna hack remote machines & turn them to U botnets? **Join** me in hackertalk. U've got to have fun!"
- Implicit Transactional Relationship: "Tools're really cool, but need U support for continuous upgrading; pls. pm me if interested."

The main contributions of our research reported in this paper are the development and evaluation of a novel weakly supervised cybercriminal network mining method which can uncover both explicit and implicit relationships among cybercriminals based on their conversational messages posted to online social media. To the best of our knowledge, this is the first successful research for developing a probabilistic generative model to mine cybercriminal networks from online social media. In this paper, we focus on mining two types of semantics: *transactional* and *collaborative* relationships among cybercriminals. The basic intuition behind the proposed computational method is that a probabilistic generative model is applied to extract multi-word expressions describing two types of cybercriminal relationships in unlabeled messages. These dynamically discovered concept descriptions are then applied to identify both explicit and implicit cybercriminal relationships embedded in online messages.

Our concept-based cybercriminal relationship classification method is more promising than keyword-based methods which relies on semantics rather than syntax. Concept-based approaches have already been shown to perform better than word-based methods for tasks such as topic modeling (Rajagopal et al., 2013), domain adaptation (Xia et al., 2013) and opinion mining (Cambria et al., 2013). For example, a keyword-based method may mistakenly classify the message "just give the port side a hack with the axe" as a cybercrime message if the keyword "hack" is solely used to identify online hacking activities. In contrast, concept-based methods leverage on semantic sets to classify messages. For instance, the concept "hacking in the cyberspace" is represented by a set of semantically related terms such as {"computer", "hack", "online", "network", ...,}. Accordingly, our proposed concept-based method will not classify the aforementioned message as a cybercrime related message because the sample message bears little semantic similarity with the given concept.

The remainder of the paper is organized as follows: Section 2 provides an overview of existing research related to our study and compares the existing methods with ours; Section 3 illustrates the computational details of the proposed weakly supervised cybercriminal network mining method; Section 4 describes the evaluation procedures and discusses our experimental results; finally, Section 5 offers concluding remarks and describes future directions of our research work.



Allama Iqbal Town, Lahore | 30 August, 2012 | Ad ID: 732692

hacking tools available for sale

Rs.5,000 Views: 238 | Save this ad | All ads of this user (5)

maan sheikh

Call: 03214440415

Email: Send Message

Home User

NEW Box Open - Never Used- Warranty: 1 Year 0 0

i am selling hacking tools u can hack all the computer with all application password like facebook yahoo msn etc show u every think in ur Email with screen snap shop intrested people contact me 03214440415

FIGURE 2 A transactional dialog between cybercriminals at www.hafeezcentre.pk.

2. Related Work

The work closest to ours is the application of a probabilistic latent semantic analysis (pLSA) model to mine latent topics describing cybercrimes from blog messages (Tsai and Chan, 2007). However, only a qualitative evaluation about the quality of the extracted cybercrime concepts was performed. Moreover, the pLSA model suffers from the so-called problem of over-fitting and the extraordinary computational costs of learning a large number of model parameters (e.g., learning model parameters per document) (Blei et al., 2003; Rosen-Zvi et al., 2010). Our proposed method not only discovers latent topics (concepts) about cybercrimes from online social media, but it also leverages these latent topics to uncover a cybercriminal network. Du and Yang (2011) applied social network analysis (SNA) method to construct cyber-attack graphs based on the source and destination IP addresses of cyber-attacks extracted from the Internet. Our proposed approach can tap into online social media and utilize high-level features (e.g., concepts embedded in conversational messages) to uncover the collaborative patterns of cybercriminals.

Hu et al. (2009) applied dynamic social network analysis methods to identify facilitators of co-offending relationships in a large-scale narcotics network consisting of individuals and vehicles. Wu and Banzhaf (2010) examined various computational intelligence methods (e.g., artificial neural networks, fuzzy systems, evolutionary computation methods, artificial immune systems, and swarm intelligence) for intrusion detection using low-level network-based features. Ting et al. (2010) applied the Apriori association rule mining method to linkage identification which attempted to identify the tightly linked genes and bound them together to form building blocks to alleviate the disruptiveness induced by crossover operations. Abbass et al. (2011) proposed the computational red team (CRT) method that leverages the relationships between both blue team and red team of agents to improve agents' decision making processes. Martino and Sperduti (2010) evaluated two classes of methods, namely neural networks and kernel methods for modeling entities and their relationships in the forms of trees and graphs.

The CoMiner system applied Natural Language Processing (NLP) techniques and predefined lexico-syntactic patterns to identify competitive relationships and competitive domains among companies (Bao et al., 2008; Li et al., 2006). In particular, the Point-wise Mutual Information (PMI) measure was applied to estimate the strength of a competitive relationship between two companies. Xia et al. (2013) employed a predefined relationship lexicon and shallow NLP techniques to develop the CoNet system for the discovery of potential business relationships from online financial news articles. Two different types of business relationships such as cooperative and competitive relationships were identified according to a set of pre-defined relationship indicators captured in a relationship lexicon. Since the CoNet system mainly relied on a limited set of seeding relationship indicators to identify business relationships, the recall of such a system may be low. Our research differs from the aforementioned studies in that we examine the mining of cybercriminal networks instead of business networks.

3. A Methodology of Collaborative Cybercriminal Network Discovery

The basic intuition behind the proposed cybercriminal network discovery method is that latent concepts describing specific types of cybercriminal relationships (e.g., transacting cyber-attack tools) are extracted by a probabilistic generative model to bootstrap the performance of cybercriminal relationship identification. Figure 3 illustrates the main steps of the proposed cybercriminal network mining methodology. First, conversational messages US_i that refer to at least two users are extracted from a collection of unlabeled documents (e.g., online messages posted by hackers). In addition, generic seeding relationship indicators are applied to label a set of messages LS_i describing transactional activities, or collaborative cyber-attack activities among cybercriminals.

The entities (i.e., individuals or groups of cybercriminals) being referred to within the conversational messages are identified using an extended named entity recognition (NER) module of GATE (Maynard et al., 2001). An initial list of well-known cybercriminals is provided by cyber-security experts to enrich the ordinary entity dictionary of GATE. In addition, the user identities associated with these messages are extracted based on the publicly available user profiles on online social media. The extracted messages are then fed into an LDA-based (Blei et al., 2003; Rosen-Zvi et al., 2010; Steyvers et al., 2004) topic modeling module to extract relevant concepts (i.e., the topics describing various cybercriminal relationships) to alleviate the low-recall problem of a purely lexicon-based relationship identification approach. In particular, we developed a novel context-sensitive (CS) Gibbs sampling algorithm to implement the LDA-based probabilistic generative model.

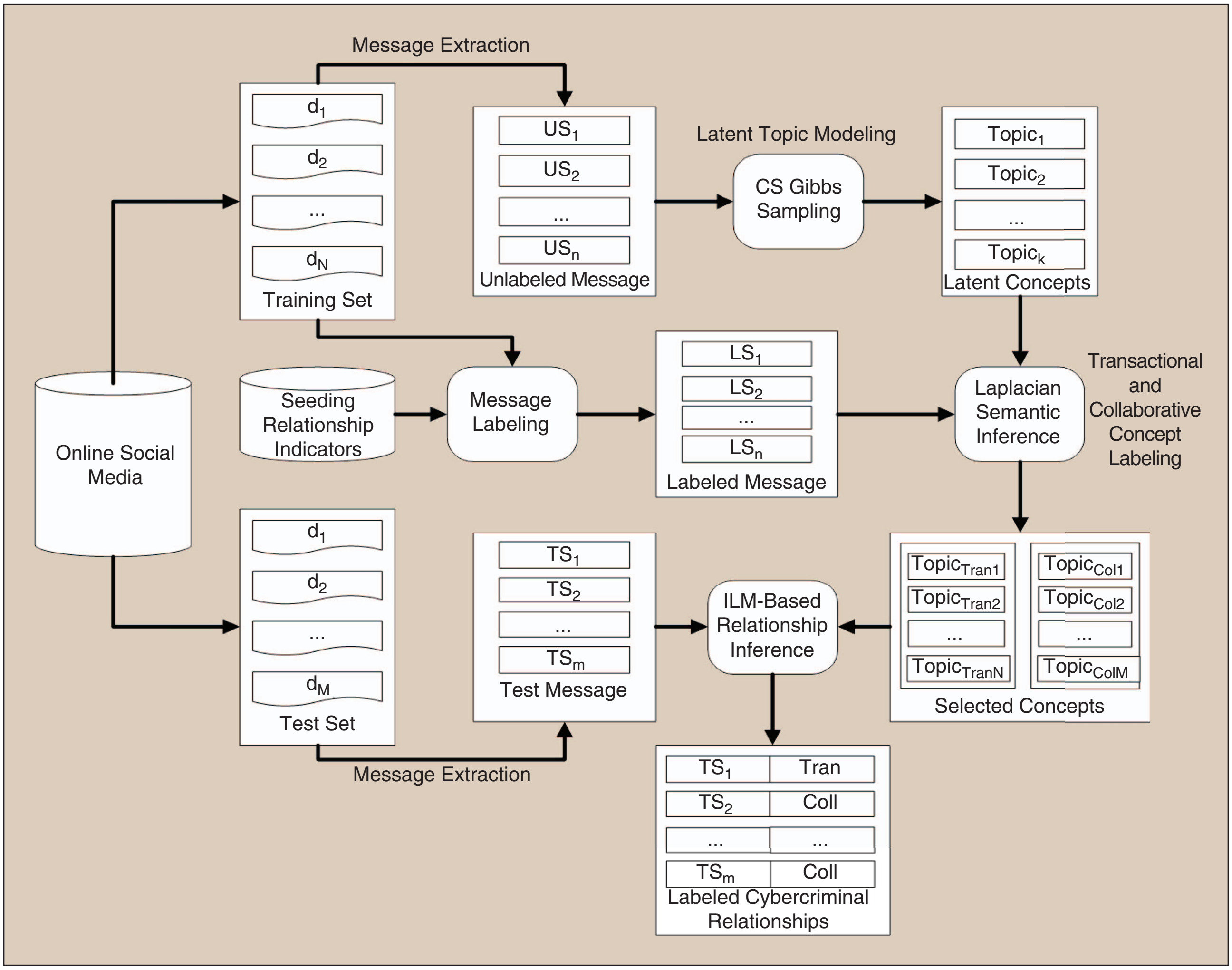


FIGURE 3 A probabilistic generative model for collaborative cybercriminal network mining.

By applying a Laplacian-based semantic ranking method (Song et al., 2009), the messages with relationship labels LS_i are used to infer the semantic labels (e.g., *transactional* or *collaborative*) of the mined latent concepts. Then, these labeled concepts are applied to determine the implicit cybercriminal relationship embedded in an arbitrary message TS_i that refers to at least two parties (i.e., cybercriminals). More specifically, we develop a novel inferential language modeling (ILM) method to infer the hidden cybercriminal relationship in a message. Finally, based on the identified cybercriminal relationship between each pair of users, a cybercriminal network is generated for a given period. The proposed computational methodology is general enough to support the discovery of any types of cybercrime related relationships if the corresponding sets of seeding relationship indicators are provided. Since our methodology requires a small set of seeding relationship indicators as input, it is a weakly supervised relationship mining method. However, the distinct advantage of the proposed method is that there is no need for the expensive manual labeling of a large number of training messages.

3.1. Learning Latent Concepts

For LDA-based latent concept learning, each conversational message $d \in D$ of an unlabeled social media corpus D is considered to be characterized by a multinomial distribution θ, which is in turn controlled by a hyper-parameter, a Dirichlet priori α (Blei et al., 2003). Z represents the set of latent concepts characterizing D. A latent concept $z_i \in Z$ (e.g., a type of cybercriminal relationship) is selected according to the multinomial distribution θ. Given the concept z_i, a term t_k is then generated according to the multinomial distribution ϕ, controlled by another hyper-parameter, a Dirichlet priori β. Our ultimate goal is to infer the conditional probability $\phi = Pr(t_k | z_i)$ which represents a latent concept about a specific type of cybercriminal relationship. However, directly computing the multinomial distribution ϕ or θ is computationally very expensive. Accordingly, Gibbs sampling (Geman and Geman, 1984), a Markov Chain Monte Carlo algorithm, is extended to estimate the multinomial distributions of the LDA model. For Gibbs sampling, a Markov chain is established by repeatedly drawing a latent topic for each observable term, based on its conditional probability over other variables (Steyvers et al., 2004). The approximations $\overline{\phi}$ and $\overline{\theta}$ of the multinomial distributions ϕ and θ are defined as follows.

$$\overline{\theta} = \frac{N_{np}^{ZD} + \alpha}{\sum_{n' \in Z} N_{n'p}^{ZD} + |Z|\alpha} \tag{1}$$

$$\overline{\phi} = \frac{N_{mn}^{VZ} + \beta}{\sum_{m' \in V} N_{m'n}^{VZ} + |V|\beta}. \tag{2}$$

The term N_{mn}^{VZ} represents a count matrix that captures the number of times that term $t_k = m$ is assigned to the latent concept $z_i = n$, excluding the current word position. The terms m and n represent the indices of the count matrix N_{mn}^{VZ}. The term V is the set of vocabulary that is used to compose the collection D, and the set Z represents the set of latent concepts characterizing the collection D. The count matrix N_{np}^{ZD} records the number of times that the latent concept $z_i = n$ is assigned to a document $d_i = p$, excluding the current document under consideration. The terms n and p are the indices of the count matrix N_{np}^{ZD}. In Eq. 2, the term $(N_{mn}^{VZ} + \beta)/\left(\sum_{m' \in V} N_{m'n}^{VZ} + |V|\beta\right)$ is used to estimate the probability of the term t_k given the latent concept z_i, whereas $(N_{np}^{ZD} + \alpha)/\left(\sum_{n' \in Z} N_{n'p}^{ZD} + |Z|\alpha\right)$ is applied to approximate the probability of z_i given the document d_i in Eq. 1.

Our previous success in applying association rule mining to improve both the quality of the mined latent aspects and the time of convergence of an expectation-maximization (EM) algorithm for a probabilistic latent semantic indexing (pLSI) model (Song et al., 2012) motivates us to explore context-sensitive text mining to extend the Gibbs sampler for LDA-based latent topic modeling. This is indeed one of the main research contributions of our research work. The details of our novel context-sensitive text mining (Lau et al., 2008) enhanced Gibbs sampling algorithm CS-GibbsSampling are depicted in Figure 4. The algorithm begins with the initialization of several control variables and the count matrices N_{mn}^{VZ} and N_{np}^{ZD} which are the basis to estimate $\overline{\theta}$ and $\overline{\phi}$. Then, step 6 of the CS-GibbsSampling algorithm invokes context-sensitive text mining to discover term association knowledge (i.e. information flows) to guide term-to-topic assignment in the count matrix N_{mn}^{VZ}. In context-sensitive text mining (Lau et al., 2008), a virtual text window of ω_{win} terms is composed, and it moves from left to right one term each time in each document. The co-occurrence information among terms within each virtual text window is collected and analyzed to produce term associations of the form $(t_1, t_2, \ldots, t_{k-1}) \rightarrow t_k$, where t_k is a term of the corpus D (Lau et al., 2008). More specifically, steps 7 to 11 of the algorithm states that if the terms appearing in the left hand side of a context-sensitive term association (i.e., information flow) are assigned to a topic (column) of N_{mn}^{VZ}, the term that appears in the right-hand side of the association should also be assigned to the same topic (column) because terms $(t_1, t_2, \ldots, t_{k-1})$ logically implies t_k. When compared to the original Gibbs sampling algorithm which randomly assigns terms to topics (Geman and Geman, 1984), our proposed algorithm can better leverage contextual knowledge (i.e., information flows) extracted from a domain-specific corpus to make a more informed term-to-topic assignment.

The main loop of the CS-GibbsSampling algorithm is controlled by testing change of perplexity and the maximum number of iterations (step 12). If the change of perplexity in recent two iterations ξ is less than or equal to the threshold ζ or the number of iterations I reaches the maximum Max_I, the Gibbs sampling algorithm will be terminated. Basically, steps 13 to 30 of the algorithm illustrates the iteration process of updating the count matrices N_{mn}^{VZ} and N_{np}^{ZD}. The proposed Gibbs sampling algorithm computes the entropy of a term t_m to filter noisy terms (steps 31 to 35). These noisy terms are associated with many topics, and are unlikely to be representative terms to uniquely describe a topic. The entropy of a term is defined as

```
Algorithm CS-GibbsSampling($D, \alpha, \beta, \zeta, \delta, \kappa, \omega_{\text{win}}, \text{Max}_I$)
Inputs: an unlabeled corpus $D$, hyper-parameters $\alpha$ and $\beta$ controlling Dirichlet prior distributions, termination
threshold $\zeta$, entropy threshold $\delta$ for noisy term removal, the number of latent topics $\kappa$, the text window size $\omega_{win}$
for context-sensitive text mining, maximum number of iterations $\text{Max}_I$
Main Procedure:
NoisyTerms $\longleftarrow \phi, I \longleftarrow 0$
$\xi \longleftarrow \zeta + 1$   // initialize the perplexity change variable $\xi$
$|Z| \longleftarrow \kappa$   // specified no. of topics
initialize count matrix $N_{np}^{ZD}$ with random non-negative integers
initialize count matrix $N_{mn}^{VZ}$ with random non-negative integers
apply context-sensitive text mining to $D$ using $\omega_{\text{win}}$ to obtain rule set $R$ of the form $(t_1, t_2, ..., t_{k-1}) \longrightarrow t_k$
for each column $n$ in $N_{mn}^{VZ}$ do   // adjust each column $n$ of count matrix $N_{mn}^{VZ}$ according to $R$
    if $(t_1, t_2, ..., t_{k-1}) \longrightarrow t_k \in R$ and min$(\{c_{1n}, c_{2n}, ..., c_{(k-1)n}\}) > 0$
        $C_{kn} \longleftarrow$ min$(\{c_{1n}, c_{2n}, ..., c_{(k-1)n}\})$   // $C_{kn}$ is an element in $N_{mn}^{VZ}$
    end
end
while $\xi > \zeta$ and $I < \text{Max}_I$ do
    for each $d_i \in D$ do
        for each $t_i \in d_i$ do
            if $t_i \notin$ NoisyTerms   // skipping noisy terms
                $m \longleftarrow$ row$(N_{mn}^{VZ}, t_i)$   // obtain term index in $N_{mn}^{VZ}$
                $n \longleftarrow$ col$(N_{mn}^{VZ}, t_i)$   // retrieve current topic assignment
                $c_{mn} \longleftarrow c_{mn} - 1$   // decrement count in $N_{mn}^{VZ}$
                $p \longleftarrow$ doc$(d_i)$   // obtain document index in $N_{np}^{ZD}$
                $c_{np} \longleftarrow c_{np} - 1$   // decrement count in $N_{np}^{ZD}$
                for $k = 1$ to $|Z|$ do
                $Pr(z_k|t_i, d_i) \longleftarrow \frac{N_{np}^{ZD} + \alpha}{\Sigma_{n' \in Z} N_{n'p}^{ZD} + |Z|\alpha} \cdot \frac{N_{mn}^{VZ} + \beta}{\Sigma_{m' \in V} N_{m'n}^{VZ} + |V|\beta}$
                end
                sample to obtain $z$ using the new distribution $Pr(z_k|t_i, d_i)$
                $n \longleftarrow$ row$(z)$   // obtain topic index from $N_{np}^{ZD}$
                $c_{np} \longleftarrow c_{np} + 1$   // increment count in $N_{np}^{ZD}$
                $c_{mn} \longleftarrow c_{mn} + 1$   // increment count in $N_{mn}^{VZ}$
            end
        end
    end
    for each row $m$ in $N_{mn}^{VZ}$ do
        if entropy$(t_m) > \delta$
            NoisyTerms $\longleftarrow$ NoisayTerms $\cup\, t_m$   // skipping noisy terms in Gibbs sampling
        end
    end
    compute the perplexity perp of a held-out sample $D_{\text{test}}$ using the trained model
    compute the perplexity change $\xi$ between the recent two iterations
    $I \longleftarrow I + 1$
end
return $N_{mn}^{VZ}, N_{np}^{ZD}$
```

FIGURE 4 Context-sensitive gibbs sampling.

follows: $\text{entropy}(t_m) = \left(\sum_{i=1}^{|Z|} \text{ass}(z_i, t_m)\right) / |Z|$ where the assignment function $\text{ass}(z_i, t_m)$ returns 1 if the term t_m is assigned to the topic $z_i \in Z$. The computational complexity of the context-sensitive Gibbs sampling algorithm is characterized by $O(I \cdot |Z| \cdot |V|^2 \cdot |D|)$, where I is the number of Gibbs iterations; V is the vocabulary set of a corpus D, and $|Z|$ is the pre-defined number of latent concepts. The additional computational cost of our CS-GibbsSampling algorithm when compared to the classical Gibbs Sampling is characterized by $|V|^2$ which is devoted to perform context-sensitive text mining over a corpus. Our previous empirical results have shown that the text mining process is relatively efficient even for a medium to large textual corpus (Lau et al., 2008, 2009).

Determining the optimal number of latent concepts of an LDA-based model is always a challenge. We adopted a perplexity-based approach to estimate a reasonable number of latent concepts characterizing an unlabeled training corpus D and filter out noisy topics (Steyvers et al., 2004). More specifically, the context-sensitive Gibbs sampling algorithm is invoked with different number of concepts κ. Then, we select the smallest number of concepts that achieved a good (i.e., small) perplexity score. We extended the open source Core LingPipe API[2] to implement the novel context-sensitive Gibbs sampling algorithm that operationalizes the aforementioned probabilistic generative model.

3.2. Laplacian Scoring for Latent Concept Labeling

Since the latent concepts discovered by LDA-based latent semantic mining are unlabeled, we must extend the classical LDA method such that we can infer the semantic label (e.g., collaborative or transactional) of a mined latent concept. Song et al. (2009) have applied Laplacian scoring, which is originally developed for feature selection, to latent topic selection. The intuition is that if a candidate latent concept is semantically close to some messages characterized by a specific semantic label, the latent concept is likely to have such a semantic label as well. A message with relationship label LS_i was extracted by applying some generic seeding relationship indicators to an unlabeled corpus D. For example, by applying the seeding indicator "sales" that usually signifies a *transactional* cybercriminal relationship, the following message LS_i is labeled based on our social media corpus: "@iqbal hacking tools 4 **sales**; U'd pm me if like purchasing."

Since we focus on collaborative and transactional relationships in this paper, two different Laplacian rankings are constructed to identify two sets of latent concepts with the respective semantics. Let $L_{k,r}, r \in \{\text{collaborative, transactional}\}$ denotes the Laplacian score of the k*th* concept for the semantic label r. Moreover, let $d_{i,k}$ represents the i*th* message corresponding to the k*th* concept. Based on the outputs from the Laplacian scoring method (Song et al., 2009), we can then select the highly ranked and *consistent* latent concepts $z_i \in \text{dom}(L_r)$ to represent the specific type of cybercriminal relationship r. For each relationship type r, an aggregated concept $C_j = \{z_i \in \text{dom}(L_r) : j = r\}$ is then constructed to infer the relationship label of an arbitrary message TS_i.

3.3. Semantic Inference of Cybercriminal Relationships

After Laplacian-based latent concept labeling (Song et al., 2009), the aggregated concepts C_{tran} (i.e., sets of transactional concepts) and C_{coll} (i.e., sets of collaborative concepts) are applied to infer the relationship label of an arbitrary message $d = TS_i$ that refers to at least two cybercriminals. The basic intuition is that if the contents of the message d are more likely to generate the transactional (collaborative) concept, the message will be considered to describe a transactional (collaborative) cybercriminal relationship. We develop a novel inferential language modeling method to estimate the probability of d generating a specific cybercriminal relationship label.

$$Pr(C_j|M_d) = \prod_{t_i \in C_j} Pr(t_i|M_d) \tag{3}$$

$$Pr(t_i|M_d) = (1-\lambda)\, Pr_{\text{INF}}(t_i|M_d) + \lambda Pr_{\text{ML}}(t_i|M_D) \tag{4}$$

$$Pr_{\text{INF}}(t_i|M_d) = (1-\gamma)\, Pr_{\text{ML}}(t_i|M_d) + \gamma Pr_{\text{TM}}(t_i|M_R) \tag{5}$$

$$Pr_{\text{TM}}(t_i|M_R) = \tanh\left(\sum_{(t_j \rightarrow t_i) \in R} Pr(t_j \rightarrow t_i) \cdot Pr_{\text{ML}}(t_j|M_d) \right). \tag{6}$$

Semantic-based language modeling $Pr(C_j|M_d)$ is applied to estimate the probability that the message $d = TS_i$ generates the aggregated concept C_j, where M_d is a document language model (Ponte and Croft, 1998). However, to cope with the challenge that the mined concepts are incomplete, each term $t_i \in C_j$ should be smoothed with respect to the entire cybercrime message corpus D by means of the maximum likelihood collection language model M_D in Eq. 4. In this paper, we develop a novel inferential language model $Pr_{\text{INF}}(t_i|M_d)$ that consists of the maximum likelihood estimation $Pr_{\text{ML}}(t_i|M_d)$ of the term t_i with respect to d, and the context-sensitive text mining (Lau et al., 2008) based smoothing $Pr_{\text{TM}}(t_i|M_R)$. The maximum likelihood document language model $Pr_{\text{ML}}(t_i|M_d) = \text{freq}(t_i, d)/|d|$ is defined based on the frequency of t_i appearing in the message d. The Jelinek-Mercer smoothing parameters $\lambda = 0.35$ and $\gamma = 0.31$ are empirically established based on a subset of our cybercrime message corpus. Previous research also shows that the Jelinek-Mercer smoothing parameter usually falls in the range of $[0.1, 0.7]$ (Zhai and Lafferty, 2004).

Finally, context-sensitive text mining based language model M_R is defined according to Eq. 6. The set of term associations (i.e., information flows) R is mined by applying the context-sensitive text mining method (Lau et al., 2008) to the unlabeled cybercrime corpus D. A context-sensitive term association of the form $t_j \rightarrow t_i$ is applied to text mining based language modeling $Pr_{\text{TM}}(t_i|M_R)$ to estimate the probability that the message generates a term t_j which is contextually associated with a relationship indicator t_i captured in C_j. In particular, the strength of term association $Pr(t_j \rightarrow t_i)$ is derived based on the context-sensitive text mining method. Pragmatically, we only consider the top $\chi = 3$ term associations for each relationship indicator t_i captured in R. Since the inference that d generating t_j which implies the relationship indicator t_i involves uncertainty, the maximum likelihood estimation of $P_{\text{ML}}(t_j|M_d)$ is discounted by a factor $Pr(t_j \rightarrow t_i)$. The hyperbolic tangent function ensures that $Pr_{\text{TM}}(t_i|M_R)$ is a valid probability. The main differences between the proposed language model and other exiting inference-based language models (Nie et al., 2006) are that we apply context-sensitive text mining to extract term associations and we use a summation instead of multiplication operator to combine the probabilities of deduced terms to smoothen the maximum likelihood document language model.

For a binary classification, if $(Pr(C_{\text{tran}}|M_d) - Pr(C_{\text{coll}}|M_d)) > \omega_{\text{rel}}$ or $(Pr(C_{\text{coll}}|M_d) - Pr(C_{\text{tran}}|M_d)) > \omega_{\text{rel}}$ is established, the message d is classified to have the same cybercriminal relationship label of the aggregated concept C_j which is generated with a higher probability according to the proposed inferential language model. Otherwise, the

[2] http://alias-i.com/lingpipe/web/download.html

TABLE 1 Details of the cybercrime corpora and latent topics.

	TWITTER	ONLINE FORUMS	TOTAL
# CYBERCRIME MESSAGES	28,114	25,695	53,809
# CYBERCRIME SENTENCES	87,153	131,045	218,198
# MESSAGES WITH TWO USERS	5,112	3,294	8,406
# ANNOTATED MESSAGES	4,283	2,830	7,113
# COLLABORATIVE MESSAGES	1,045	972	2,017
# TRANSACTIONAL MESSAGES	314	677	991
# AMBIGUOUS MESSAGES	2,924	1,181	4,105
# SELECTED TOPICS	65	70	–
# TRANSACTIONAL CONCEPTS	5	7	–
# COLLABORATIVE CONCEPTS	11	11	–

cybercriminal relationship is undefined. The classification threshold $\omega_{\text{rel}} = 0.2E-8$ was empirically established based on a subset of our cybercrime message corpus. Alternatively, the proposed inferential language model can perform a ranking-based classification of the relationship labels of messages. In that case, messages are ranked according to the their probabilities of having a specific relationship label. Accordingly, a binary classification threshold is not needed.

After relationship classification, a frequency count $\text{Rel}_j(v_x, v_y)$ for a specific type of relationship $j \in \{\text{transactional, collaborative}\}$ is developed for each pair of cybercriminals (v_x, v_y). These frequency values are then subject to a linear normalization $\text{Rel}_{\text{nor}} = (\text{Rel} - \text{Rel}_{\text{min}}) / (\text{Rel}_{\text{max}} - \text{Rel}_{\text{min}})$ to develop the final relationship scores for all the pairs. If a pair (v_x, v_y) has both a transactional relationship and a collaborative relationship at the same time (i.e., $\text{Rel}_{\text{tran}}(v_x, v_y) > 0$ and $\text{Rel}_{\text{coll}}(v_x, v_y) > 0$), our system simply assigns the more specific transactional relationship to the pair. Finally, a cybercriminal network is composed based on the identified relationships among all the valid pairs pertaining to a specific period.

4. System Evaluation

To the best of our knowledge, a benchmark data set for the evaluation of cybercriminal network mining algorithms is not yet available at the time of this writing. To evaluate the effectiveness of the proposed cybercriminal network mining method, we first needed to retrieve cybercrime related messages from online social media. For the construction of our evaluation corpora, we made use of two kinds of social media sources, namely micro blogs and online forums. We accessed to the largest micro blog service, Twitter, and a dozen of online forums (e.g., hacktalk.net, blackhatworld.com, pastebin.com, etc.) to develop two cybercrime related corpora. We manually identified a list of 35 well-known cybercriminals as the seeding users. Then, we used these seeding users as the starting points to perform breadth-first crawling to retrieve the messages posted by other suspects of cybercrimes. As for Twitter, we retrieved the relevant tweets via a publicly available API called Topsy.[3] For instance, for the well-known cybercriminal group *Anonymous* (also known as *FawkesSecurity* on Twitter) who claimed to be responsible for a series of cyber-attacks against HSBC in October 2012, our crawler first identified all the followers of this account, and then invoked an internal filter to extract all cybercrime related tweets from different followers or friends. For online forums, our crawler identified the related user accounts that appeared in the same thread of messages or directly embedded in the message contents. A total of 28,114 cybercrime related messages covering the period from January 2009 to December 2012 were retrieved from Twitter in January 2013. In addition, a total of 25,695 cybercrime related messages of the same period were retrieved from various Internet forums. To distinguish cybercrime related messages from ordinary online chatting, our internal filter simply utilized a list of 21 common cybercrime keywords to determine the nature of each conversational message. These keywords were provided by a group of six cyber-security experts who were the employees of a cyber-security consulting firm.

For each cybercrime message corpus, a subset of messages with at least two cybercriminals mentioned in each message was manually inspected and annotated by a group of three cybersecurity experts so as to determine the specific cybercriminal relationship captured in the message. For the experiments reported in this paper, we only focus on two types of cybercriminal relationships namely *transactional* relationship and *collaborative* relationship. A transactional relationship refers to buying or selling cyber-attack tools between two parties, whereas a collaborative relationship simply implies the sharing of information or tools between cybercriminals and it does not involve any monetary exchange between the two parties. Only if all three experts agreed on a specific type of relationship captured in a message, would that message be annotated with the corresponding relationship label. The average inter-rater agreement of six annotators as measured by Cohen's Kappa is $\mathcal{K} = 0.75$ which indicates a relatively consistent and reliable expert judgment for the construction of our evaluation corpora. The details of our cybercrime message corpora applied to our experiments are given in Table 1. Common performance evaluation measures such as Precision (P), Recall (R), F-measure (F), and Accuracy (A) were applied to our experiments. Moreover, the Receiver Operating Characteristic (ROC) curve (Hand and Till, 2001) was also adopted to assess the performance of all systems such that the results are independent of any particular classification threshold value chosen.

Apart from the context-sensitive LDA (CSLDA) experimental system that is underpinned by context-sensitive Gibbs sampling for latent cybercriminal relationship mining, we also implemented several baseline systems to perform a comparative evaluation. Other probabilistic generative models such as Partially Labeled Dirichlet Allocation (Ramage et al., 2011) and Latent Dirichlet Allocation (Blei et al., 2003) were also adopted as baseline systems. For both of these baseline systems, a classical Gibbs sampler (Geman and Geman, 1984) were employed. PLDA requires human assigned tags for documents. Unfortunately, unlike Web pages, these tags are not normally

[3] http://topsy.com/

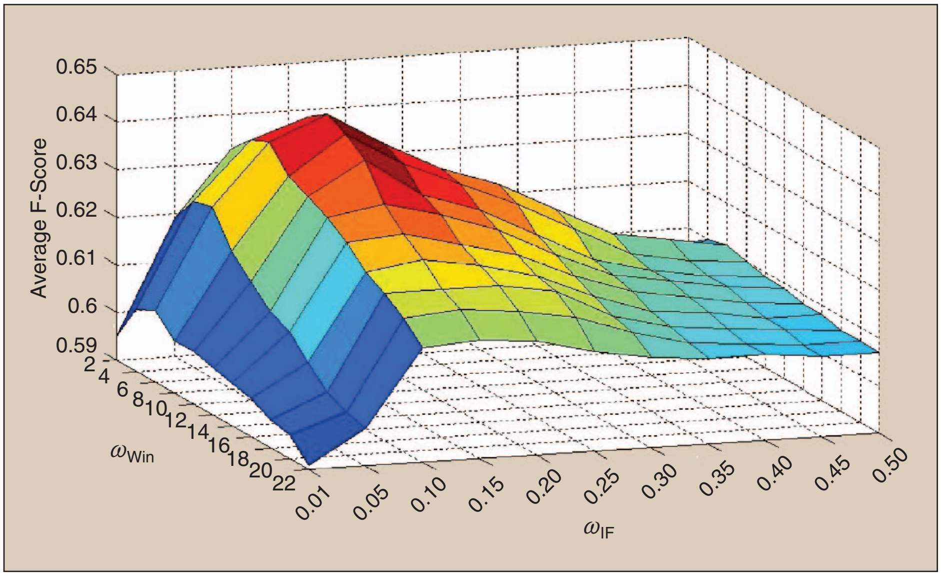


FIGURE 5 Empirical parameter setting for context-sensitive term associations.

available for cybercrime related messages. Given the large number of un-tagged messages of our corpora, we employed a semi-automated method for tag generation. Firstly, six annotators were responsible for tagging around 10% of our messages. Secondly, the cosine similarities between tagged messages and un-tagged messages were computed. Finally, an un-tagged message was assigned the tag of its most similar tagged message if the cosine similarity score was above a predefined threshold (i.e., 0.5). For the remaining messages without any tag assigned, we annotated them with the generic tag "cybercrime". Similar to the approach adopted by (Ramage et al., 2011), we specified that each tag was associated with 5 topics for the PLDA baseline system.

Another baseline systems (SEED) employed 11 seeding transactional indicators, 16 seeding collaborative indicators, and the corresponding synonyms (top 3 synonyms of each seeding indicator) extracted from WordNet (Fellbaum, 1998) to identify cybercriminal relationships from messages. These relationship indicators were also used by the experimental system to infer the labels of mined latent concepts. The seeding relationship indicators and the cybercrime corpora were stemmed using the same Porter stemming algorithm (Porter, 1980). The SEED baseline system simply uses a transactional (collaborative) strength measure $\mathrm{tran}(CP_i) = (|\mathrm{TranInd}| - |\mathrm{CollInd}|)/(|\mathrm{TranInd}| + |\mathrm{CollInd}|)$ $(\mathrm{coll}(CP_i) = (|\mathrm{CollInd}| - |\mathrm{TranInd}|)/(|\mathrm{TranInd}| + |\mathrm{CollInd}|))$ to determine the relationship label of a test message CP_i, where TranInd and CollInd are the sets of transactional and collaborative indicators found in the message. For example, if $\mathrm{tran}(CP_i) > \omega_{\mathrm{tran}}$ $(\mathrm{coll}(CP_i) > \omega_{\mathrm{coll}})$ is true, the message is considered to be transactional (collaborative). The threshold $\omega_{\mathrm{tran}}(\omega_{\mathrm{coll}})$ was empirically established for our experiments. In addition, classical supervised machine learning classifiers such as Support Vector machine (SVM) with a RBF kernel[4], and Conditional Random Fields (CRF)[5] were also used. Stop word removal, case transformation, and stemming were applied to the cybercrime corpora before they were processed by the experimental and the baseline systems. For the SVM and CRF baseline systems, word-based features and TFIDF term weighting were applied. In addition, part-of-speech, number of seeding transactional indicators, number of seeding collaborative indicators, and lexical features such as sentence length and lexical diversity were applied to the baseline systems. For the CRF baseline system, contextual information such as words preceding and after seeding relationship indicators or user names, was also used. A standard three-fold cross validation was applied to our experiments.

For the experimental system CSLDA, concepts representing transactional and collaborative cybercriminal relationships were first acquired via LDA-based latent topic modeling. The number of latent concepts $|Z|$ was estimated according to the perplexity measure (Steyvers et al., 2004). Following the empirical finding of Griffiths and Steyvers (Steyvers et al., 2004), the hyper-parameters α and β of the context-sensitive Gibbs sampling algorithm was set to $50/|Z|$ and 0.1, respectively. Two independent Gibbs samples were used and the first three hundreds Gibbs samples produced by our algorithm were ignored to ensure that the *burn-in* period had been by-passed. The maximum loop control of Gibbs sampling $\mathrm{Max}_I = 1{,}000$ was set to ensure a proper convergence. The statistics of latent concepts learning and Laplacian concept labeling of the experimental system are summarized in the second half of Table 1. For context-sensitive Gibbs sampling, the virtual text window size ω_{win} and the information flow quality threshold ω_{IF} for context-sensitive term associations extraction (Lau et al., 2008) were empirically established based on the Twitter corpus. We tried different combinations of ω_{win} and ω_{IF} as shown in Figure 5 while fixing the values of other system parameters. We found that $\omega_{\mathrm{win}} = 6$, and $\omega_{\mathrm{IF}} = 0.15$ led to the best F-measure, and then we applied these parameter values to the experiments based on the forum corpus as well. Based on the selection parameter $\omega_{\mathrm{IF}} = 0.15$, there were $1{,}512$ and $1{,}306$ context-sensitive term associations extracted from the Twitter and the forum corpora, respectively.

It should be noted that our empirical parameter setting method may not be able to identify the global optimum for ω_{win}, ω_{IF}, ω_{rel}, and other parameters. A more sophisticated parameter tuning method will only further improve the performance of our proposed computational method reported in this

[4] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

[5] http://crfpp.googlecode.com/svn/trunk/doc/index.html

TABLE 2 The comparative relationship classification performance of various systems.

CORPUS	SYSTEM	TRANSACTIONAL RELATIONSHIP					COLLABORATIVE RELATIONSHIP				
		THR.	R	P	F	A	THR.	R	P	F	A
TWITTER	CSLDA	0.2E-8	0.589	0.670	0.627	0.949	0.2E-8	0.636	0.687	0.661	0.841
	PLDA	0.2E-8	0.580	0.662	0.618	0.947	0.2E-8	0.640	0.690	0.664	0.842
	LDA	0.2E-8	0.557	0.632	0.592	0.944	0.2E-8	0.599	0.646	0.622	0.822
	SVM	0.025	0.532	0.599	0.563	0.940	0.020	0.575	0.618	0.596	0.810
	CRF	0.005	0.506	0.562	0.533	0.935	0.008	0.564	0.600	0.581	0.802
	SEED	0.010	0.296	0.596	0.396	0.934	0.015	0.326	0.575	0.416	0.777
ONLINE FORUMS	CSLDA	0.2E-8	0.567	0.663	0.611	0.828	0.2E-8	0.618	0.665	0.641	0.762
	PLDA	0.2E-8	0.563	0.664	0.609	0.827	0.2E-8	0.616	0.663	0.639	0.760
	LDA	0.2E-8	0.548	0.637	0.589	0.817	0.2E-8	0.598	0.638	0.617	0.746
	SVM	0.025	0.518	0.602	0.557	0.803	0.020	0.577	0.602	0.589	0.724
	CRF	0.005	0.510	0.591	0.547	0.798	0.008	0.557	0.585	0.570	0.712
	SEED	0.010	0.288	0.598	0.389	0.783	0.015	0.303	0.595	0.402	0.690
AVERAGE	CSLDA	–	0.578	0.667	0.619	0.888	–	0.627	0.676	0.651	0.801
	PLDA	–	0.571	0.663	0.614	0.887	–	0.628	0.677	0.651	0.801
	LDA	–	0.553	0.635	0.591	0.881	–	0.598	0.642	0.620	0.784
	SVM	–	0.525	0.600	0.560	0.871	–	0.576	0.610	0.593	0.767
	CRF	–	0.508	0.576	0.540	0.867	–	0.560	0.592	0.576	0.757
	SEED	–	0.292	0.597	0.392	0.859	–	0.315	0.585	0.409	0.733

TABLE 3 Examples of mined concepts for cybercriminal relationship classification.

CSLDA				LDA			
TERM	TRANSACTIONAL ($Pr(t_k \mid z_i)$)	TERM	COLLABORATIVE ($Pr(t_k \mid z_i)$)	TERM	TRANSACTIONAL ($Pr(t_k \mid z_i)$)	TERM	COLLABORATIVE ($Pr(t_k \mid z_i)$)
SELL	0.137	CHAT	0.283	BUY	0.116	COLLABORATE	0.165
MONEY	0.106	JOIN	0.267	ACCOUNT	0.101	WORK	0.142
ACCOUNT	0.102	COLLABORATE	0.229	INTEREST	0.087	CHAT	0.128
TRANSFER	0.095	JOINT	0.164	SELL	0.075	JOIN	0.103
BUY	0.091	LEARN	0.103	SEND	0.069	DOWNLOAD	0.098
BILL	0.083	ACQUIRE	0.098	MONEY	0.062	ASK	0.087
CREDIT	0.075	DOWNLOAD	0.071	EXCHANGE	0.055	TALK	0.082
SETTLE	0.073	EXCHANGE	0.052	PAY	0.048	DIRECT	0.077
BID	0.065	ADVICE	0.033	BILL	0.041	LEARN	0.063
PAY	0.052	HELP	0.026	ASK	0.032	FUN	0.054

paper. Applying soft-computing methods such as genetic algorithms (Goldberg, 1989; Lau et al., 2006) to bootstrap the performance of our probabilistic generative model for cybercriminal network mining will be left as part of our future work.

The results achieved by the experimental and the baseline systems are summarized in Table 2. The column with label "Thr." refers to the classification threshold applied to a specific system to produce the corresponding performance scores. Table 2 shows that the CSLDA experimental system outperforms most baseline systems for both transactional and collaborative cybercriminal relationship classification. In terms of F-measure, the experimental system significantly outperforms the LDA baseline system by 4.82% ($t(2) = 3.62, p < .01$ of paired one-tail t-test) and 5.03% ($t(2) = 3.78, p < .01$ of paired one-tail t-test) for transactional and collaborative cybercriminal relationship classification, respectively. Moreover, the experimental system significantly outperforms the best supervised classifier (i.e., SVM) by 10.55% ($t(2) = 4.87, p < .01$ of paired one-tail t-test) and 9.80% ($t(2) = 4.12, p < .01$ of paired one-tail t-test) for transactional and collaborative cybercriminal relationship classification, respectively. The CSLDA system performs slightly better than the PLDA system for transactional relationship classification, and it achieves comparable performance as the PLDA system for collaborative relationship classification. Through weakly supervised mining of an unlabeled corpus, the CSLDA system produces concept descriptions corresponding to two types of cybercriminal relationships; these concepts are then applied to bootstrap the performance of cybercriminal relationship classification. Table 3 highlights two representative concepts (i.e., relationship descriptions) mined by the CSLDA method and two concepts mined by

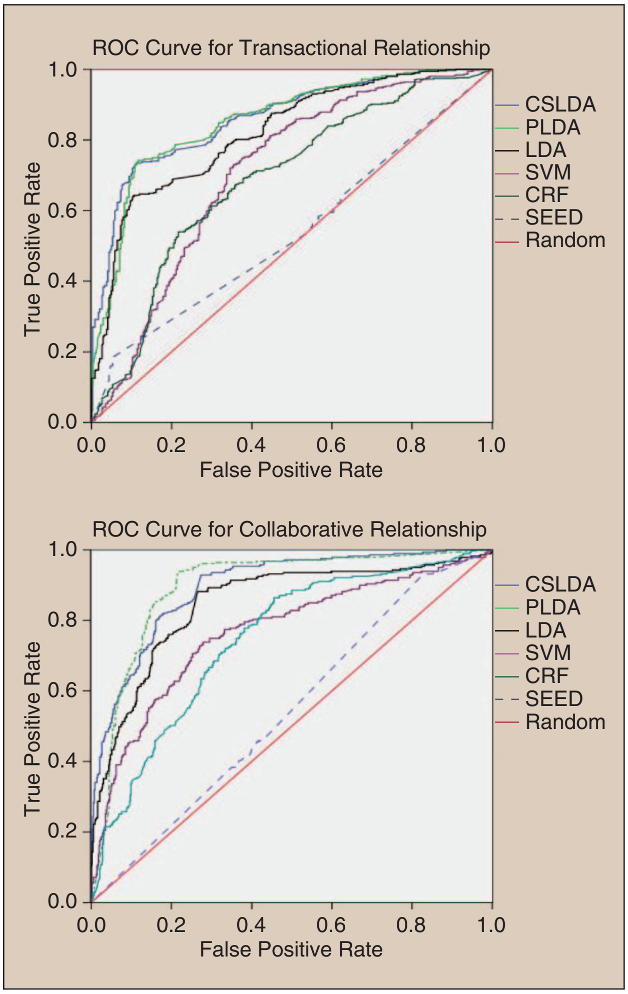


FIGURE 6 The ROC curve based on the Twitter corpus.

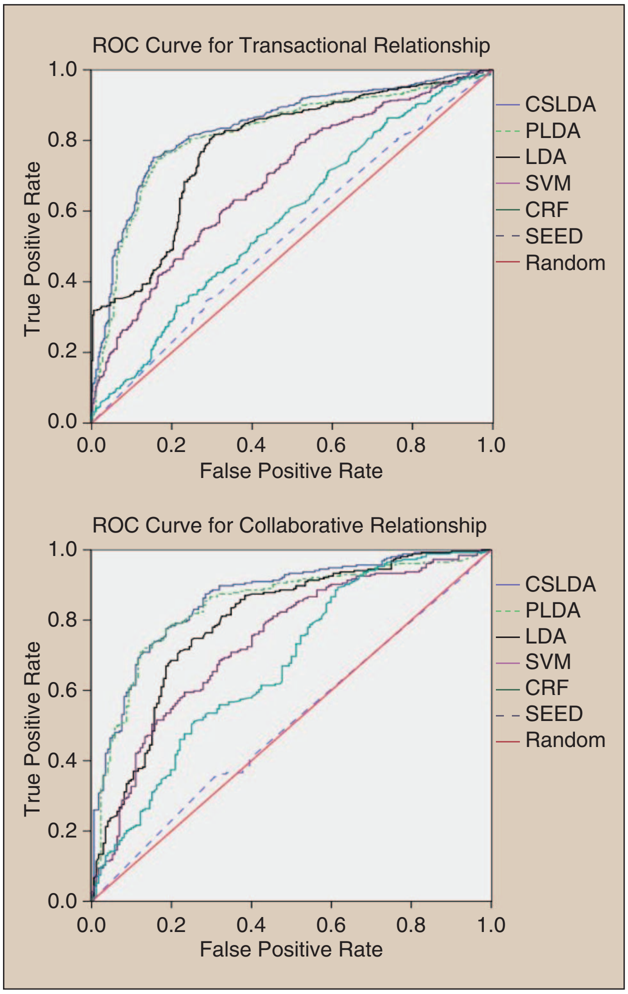


FIGURE 7 The ROC curve based on the forum corpus.

the LDA method, respectively. For brevity reason, only the top 10 terms for each concept are shown in Table 3.

In addition, Figures 6 and 7 show the ROC curves of various systems for transactional and collaborative relationship classification under the Twitter corpus and the online forums corpus, respectively. These ROC curves were plotted using SPSS with a confidence level of $p < .05$. The area under a ROC curve (AUC), that is, the probability of a classifier correctly identifies a true-positive case, was also used for a comparative evaluation. Table 4 summarizes the AUC values achieved by the experimental and baseline systems. Again, it is shown that the CSLDA experimental system performs better than most of the baseline systems do. In terms of average AUC values, the experimental system significantly outperforms the LDA baseline system by 5.23% ($t(2) = 3.81, p < .01$ of paired one-tail t-test) and the best supervised classifier (i.e., SVM) by 16.62% ($t(2) = 6.25$, $p < .01$ of paired one-tail t-test), respectively. The main reason of such a remarkable performance improvement is that the CSLDA method can leverage on a large number of unlabeled messages to learn domain-specific concepts about cybercriminal relationships, and hence to enhance cybercriminal relationship classification.

The CSLDA system performs slightly better than the state-of-the-art PLDA system although the improvement is not statistically significant. However, the distinct advantage of the CSLDA system over the PLDA system is that only a small number of seeding relationship indicators are required to guide latent topic modeling and concept-based cybercriminal relationship classification. In contrast, the PLDA system requires a manually assigned tag for each document to guide topic modeling. Unfortunately, these human assigned tags are not readily available for cybercrime related messages; this makes it extremely difficult to apply the PLDA method to mine cybercriminal networks.

According to our experimental results, we can conclude that the proposed weakly supervised concept-based cybercriminal relationship classification method significantly outperforms most of the baseline methods. It also achieves comparable performance as the state-of-the-art PLDA method. The proposed system is more effective than the SEED baseline system that only utilizes a limited number of pre-defined relationship keywords. Moreover, the proposed CSLDA system significantly outperforms the supervised machine learning classifiers such as SVM and CRF. Surprisingly, the state-of-the-art classifier CRF did not perform well in our experiments. A possible reason is that the number of labeled

TABLE 4 The comparative AUC values obtained by various systems.

CORPUS	SYSTEM	TRANSACTIONAL RELATIONSHIP		COLLABORATIVE RELATIONSHIP	
		AUC	STD. ERR.	AUC	STD. ERR.
TWITTER	CSLDA	0.848	0.013	0.857	0.014
	PLDA	0.839	0.014	0.860	0.017
	LDA	0.811	0.015	0.821	0.016
	SVM	0.703	0.017	0.754	0.018
	CRF	0.685	0.022	0.742	0.019
	SEED	0.541	0.022	0.552	0.020
ONLINE FORUMS	CSLDA	0.823	0.015	0.835	0.018
	PLDA	0.819	0.017	0.828	0.021
	LDA	0.775	0.021	0.789	0.017
	SVM	0.695	0.022	0.734	0.024
	CRF	0.551	0.012	0.603	0.021
	SEED	0.531	0.013	0.511	0.018
AVERAGE	CSLDA	0.836	0.014	0.846	0.016
	PLDA	0.829	0.016	0.844	0.019
	LDA	0.793	0.018	0.805	0.017
	SVM	0.699	0.020	0.744	0.021
	CRF	0.618	0.017	0.673	0.020
	SEED	0.536	0.018	0.532	0.019

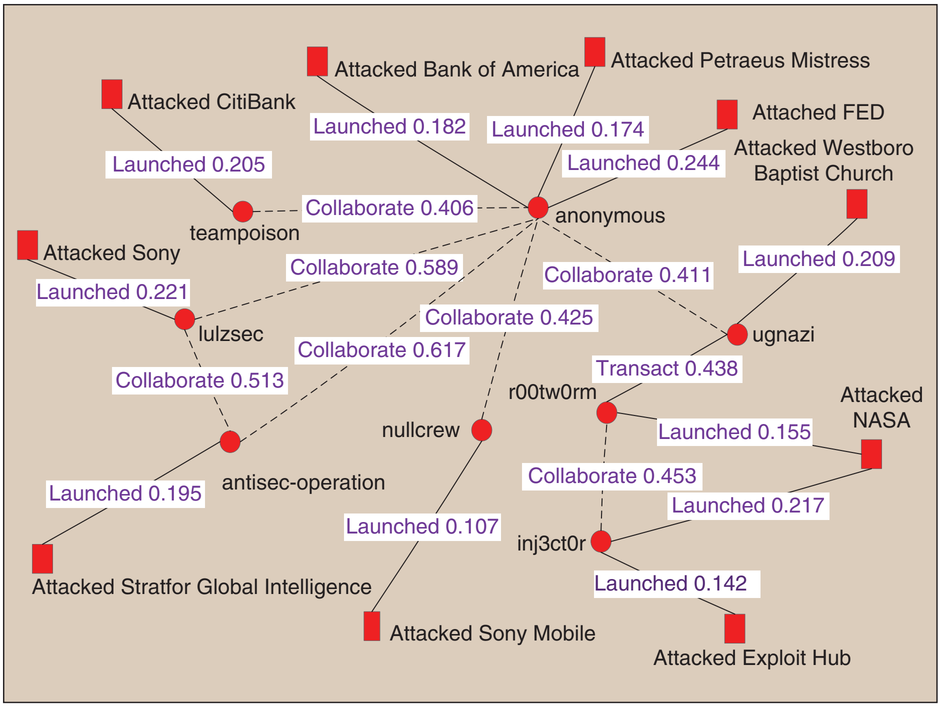


FIGURE 8 A sample segment of the mined cybercriminal network.

training examples is not sufficient to effectively train the CRF classifier to learn discriminative label sequences. Since it is extremely costly and time-consuming to label a large number of cybercriminal relationships, it may not be practical to apply supervised machine learning classifiers to cybercriminal relationship mining from online social media. With the absence of a large number of labeled cybercrime messages, the proposed CSLDA method is just able to leverage unlabeled cybercrime messages to learn latent concepts which are semantically rich representations of different types of cybercriminal relationships. Our CSLDA method outperforms the classical LDA method because the proposed context-sensitive Gibbs sampling algorithm can discover higher quality latent concepts (as shown in Table 3) when compared to that produced by the standard Gibbs sampler. These high-quality latent concepts can then be applied to bootstrap cybercriminal relationship classification.

Figure 8 shows a sample segment of the cybercriminal network mined based on our cybercrime corpora. The cybercriminal network is plotted using the open source graph display program called Pajek.[6] Each circle represents a cybercriminal or cybercriminal group (e.g., the *Anonymous* group) and a square box represents an attack incident (e.g., "Bank of America") which is most likely associated with the corresponding cybercriminal. We employed a PMI-based method to estimate the strength of association between a cybercriminal and an attack mentioned in social media messages. However, the computational details about cybercrime forensics will not be covered in this paper. Dash lines between cybercriminals represent collaborative cybercriminal relationships (e.g., *Anonymous* and *nullcrew*), whereas solid lines between cybercriminals indicate transactional relationships (e.g., *ugnazi* and *r00tw0rm*). The strength of a cybercriminal relationship is shown along an edge which is labeled with the type of relationship. This network segment was verified as a correct sub-network by our cyber-security experts. According to our experts' qualitative feedback, this kind of automatically generated cybercriminal network can considerably enhance cyber-security forensics because a large amount of intuitive and high-quality cyber-security intelligence can be generated instantly with minimal human intervention.

5. Conclusions

Latest cyber-security studies show that there is a rapid growth in the number of cybercrimes which cause tremendous financial losses to click-and-mortar organizations in recent years. The main contribution of the research work reported in this paper is the design of a novel, weakly supervised cybercriminal network mining method that is underpinned by a

[6] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

context-sensitive text mining enhanced probabilistic generative model. The proposed computational algorithm can effectively extract semantically rich representations of latent concepts describing transactional and collaborative relationships among cybercriminals based on publicly accessible messages posted to online social media. These latent concepts are then applied to bootstrap the performance of inferential language modeling-based relationship classification in texts. Based on two textual corpora extracted from online social media, our experimental results reveal that the proposed weakly supervised cybercriminal network mining method significantly outperforms the classical unsupervised LDA method by 5.23% and the well-known supervised SVM classifier by 16.62% in terms of AUC, respectively. In addition, the proposed method achieves comparable performance as the state-of-the-art PLDA method. However, the distinct advantage of the proposed method over the PLDA method is that manually tagged messages are not required for cybercrime concept learning, hence making it feasible to apply the proposed method to cybercriminal network mining.

Future work will examine soft-computing methods such as genetic algorithms to search for optimal or near-optimal system parameter values to further enhance the effectiveness of the proposed method. The issue of evolutionary cybercriminal networks will also be examined. In particular, a dynamic topic model will be examined to mine evolving concepts about cybercriminal relationships. A larger scale of empirical experiment with more corpora extracted from online social media will be performed to further evaluate both the effectiveness and the scalability of the proposed computational method. Finally, the analysis and forensics of cyber-attack behavior based on the mined cybercriminal network will be carried out.

Acknowledgment

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 145712), SRG grants from City University of Hong Kong (Project No. 7003002, 7004120, 7008138), and a grant from the Shenzhen Municipal Science and Technology R&D Funding—Basic Research Program (Project No. JCYJ20130401145617281). Xia's research was supported in part by Natural Science Foundation of China (Project No. 61272233). Ye's research was supported in part by Natural Science Foundation of China (Project No. 61272538), and Shenzhen Strategic Emerging Industries Program (Project No. JCYJ20120613135329670).

References

[1] H. Abbass, A. Bender, S. Gaidow, and P. Whitbread, "Computational red teaming: Past, present and future," *IEEE Comput. Intell. Mag.*, vol. 6, no. 1, pp. 30–42, 2011.

[2] S. Bao, R. Li, Y. Yu, and Y. Cao, "Competitor mining with the web," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 10, pp. 1297–1310, 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[4] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, 2013.

[5] E. Cambria, T. Mazzocco, and A. Hussain, "Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining," *Biol. Insp. Cogn. Arch.*, vol. 4, pp. 41–53, Apr. 2013.

[6] E. Cambria, P. Chandra, A. Sharma, and A. Hussain, "Do not feel the trolls," in *Proc. Int. Semantic Web Conf.*, Shanghai, China, 2010.

[7] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 29–31, 2012.

[8] P. J. Denning and D. E. Denning, "The profession of IT: Discussing cyber attack," *Commun. ACM*, vol. 53, no. 9, pp. 29–31, 2010.

[9] H. Du and S. J. Yang, "Discovering collaborative cyber attack patterns using social network analysis," in *Proc. 4th Int. Conf. Social Computing, Behavioral-Cultural Modeling Prediction*, 2011, vol. 6589, pp. 129–136.

[10] C. Fellbaum, *Wordnet: A electronic lexical database*. Cambridge, MA: MIT Press, 1998.

[11] J. Franklin, A. Perrig, V. Paxson, and S. Savage, "An inquiry into the nature and causes of the wealth of internet miscreants," in *Proc. ACM Conf. Computer Communications Security*, Alexandria, VA, Oct. 28-31, 2007, pp. 375–388.

[12] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian relation of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, 1984.

[13] S. Goel, "Cyberwarfare: Connecting the dots in cyber intelligence," *Commun. ACM*, vol. 54, no. 8, pp. 132–140, 2011.

[14] D. Goldberg, *Genetic Algorithms in Search, Optimisation and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[15] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Mach. Learn.*, vol. 45, no. 2, pp. 171–186, 2001.

[16] D. Hu, S. Kaza, and H. Chen, "Identifying significant facilitators of dark network evolution," *J. Amer. Soc. Inform. Sci. Technol.*, vol. 60, no. 4, pp. 655–665, 2009.

[17] R. Y. K. Lau, P. Bruza, and D. Song, "Towards a belief revision based adaptive and context-sensitive information retrieval system," *ACM Trans. Inform. Syst.*, vol. 26, no. 2, 2008.

[18] R. Y. K. Lau, D. Song, Y. Li, C. H. Cheung, and J. X. Hao, "Towards a fuzzy domain ontology extraction method for adaptive e-learning," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 6, pp. 800–813, 2009.

[19] R. Y. K. Lau, M. Tang, O. Wong, S. Milliner, and Y. Chen, "An evolutionary learning approach for adaptive negotiation agents," *Int. J. Intell. Syst.*, vol. 21, no. 1, pp. 41–72, 2006.

[20] R. Li, S. Bao, J. Wang, Y. Yu, and Y. Cao, "Cominer: An effective algorithm for mining competitors from the Web," Data Mining, in *Proc. Int. Conf. Data Mining*, 2006, pp. 948–952.

[21] G. Martino and A. Sperduti, "Mining structured data," *IEEE Comput. Intell. Mag.*, vol. 5, no. 1, pp. 42–49, 2010.

[22] D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks, "Named entity recognition from diverse text types," in *Proc. Conf. Recent Advances Natural Language Processing*, 2001.

[23] J.-Y. Nie, G. Cao, and J. Bai, "Inferential language models for information retrieval," *ACM Trans. Asian Lang. Inf. Process.*, vol. 5, no. 4, pp. 296–322, 2006.

[24] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Research Development Information Retrieval*, Melbourne, Australia, 1998, pp. 275–281.

[25] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[26] D. Rajagopal, D. Olsher, E. Cambria, and K. Kwok, "Commonsense-based topic modeling," in *Proc. ACM Int. Conf. Knowledge Discovery Data Mining*, Chicago, IL, 2013.

[27] D. Ramage, C. D. Manning, and S. T. Dumais, "Partially labeled topic models for interpretable text mining," in *Proc. 17th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, San Diego, CA, 2011, pp. 457–465.

[28] M. Rosen-Zvi, C. Chemudugunta, T. L. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," *ACM Trans. Inform. Syst.*, vol. 28, no. 1, pp. 1–38, Article 4, 2010.

[29] D. Song, Q. Huang, P. D. Bruza, and R. Y. K. Lau, "An aspect query language model based on query decomposition and high-order contextual term association," *Comput. Intell.*, vol. 28, no. 1, pp. 1–23, 2012.

[30] Y. Song, S. Pan, S. Liu, M. X. Zhou, and W. Qian, "Topic and keyword re-ranking for LDA-based topic modeling," in *Proc. 18th ACM Conf. Information Knowledge Management*, 2009, pp. 1757–1760.

[31] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths, "Probabilistic author-topic models for information discovery," in *Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, Seattle, Washington, 2004, pp. 306–315.

[32] C.-K. Ting, W.-M. Zeng, and T.-C. Lin, "Linkage discovery through data mining [research frontier]," *IEEE Comput. Intell. Mag.*, vol. 5, no. 1, pp. 10–13, 2010.

[33] F. S. Tsai and K. L. Chan, "Detecting cyber security threats in weblogs using probabilistic models," in *Proc. Pacific Asia Workshop Intelligence Security Informatics* (Lecture Notes in Computer Science), 2007, vol. 4430, pp. 46–57.

[34] S. X. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: A review," *Appl. Soft Comput.*, vol. 10, no. 1, pp. 1–35, 2010.

[35] Y. Xia, W. Su, R. Y. K. Lau, and Y. Liu, "Discovering latent commercial networks from online financial news articles," *Enterprise Inform. Syst.*, vol. 7, no. 3, pp. 303–331, 2013.

[36] R. Xia, C. Zong, X. Hu, and E. Cambria, "Feature ensemble plus sample selection: A comprehensive approach to domain adaptation for sentiment classification," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 10–18, 2013.

[37] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. Inform. Syst.*, vol. 22, no. 2, pp. 179–214, 2004.