iDetector: Automate Underground Forum Analysis Based on Heterogeneous Information Network

Yiming Zhang, Yujie Fan
Department of Computer Science and Electrical Engineering
West Virginia University, WV, USA
{ymzhang, yf0004}@mix.wvu.edu

Shifu Hou, Jian Liu
Department of Computer Science and Electrical Engineering
West Virginia University, WV, USA
{shhou, juliu}@mix.wvu.edu

Yanfang Ye ✉, Thirimachos Bourlai
Department of Computer Science and Electrical Engineering
West Virginia University, WV, USA
{yanfang.ye, thbourlai}@mail.wvu.edu

***Abstract*—Online underground forums have been widely used by cybercriminals to trade the illicit products, resources and services, which have played a central role in the cybercriminal ecosystem. Unfortunately, due to the number of forums, their size, and the expertise required, it's infeasible to perform manual exploration to understand their behavioral processes. In this paper, we propose a novel framework named *iDetector* to automate the analysis of underground forums for the detection of cybercrime-suspected threads. In *iDetector*, to detect whether the given threads are cybercrime-suspected threads, we not only analyze the content in the threads, but also utilize the relations among threads, users, replies, and topics. To model this kind of rich semantic relationships (i.e., thread-user, thread-reply, thread-topic, reply-user and reply-topic relations), we introduce a structured heterogeneous information network (HIN) for representation, which is capable to be composed of different types of entities and relations. To capture the complex relationships (e.g., two threads are relevant if they were posted by the same user and discussed the same topic), we use a meta-structure based approach to characterize the semantic relatedness over threads. As different meta-structures depict the relatedness over threads at different views, we then build a classifier using Laplacian scores to aggregate different similarities formulated by different meta-structures to make predictions. To the best of our knowledge, this is the first work to use structural HIN to automate underground forum analysis. Comprehensive experiments on real data collections from underground forums (e.g., Hack Forums) are conducted to validate the effectiveness of our developed system *iDetector* in cybercrime-suspected thread detection by comparisons with other alternative methods.**

***Index Terms*—Underground Forum Analysis, Cybercrime-suspected Thread Detection, Heterogeneous Information Network.**

I. Introduction

As Internet becomes increasingly ubiquitous, computing devices connected to Internet have permeated all facets of people's daily life (e.g., online shopping, social networking). Driven by the innovation centering on Internet ecosystem, the growth of e-commerce has been significantly increased in the overall sluggish economy [16]: worldwide e-commerce sales reached over $2.14 trillions in 2017 and are expected to increase to about $4 trillions in 2020 [4]. Though the Internet has become one of the most important drivers in the worldwide economy, it also provides an open and shared platform by dissolving the barriers so that everyone has opportunity to realize his/her innovations, which implies higher prospects for illicit profits at lower degrees of risk. That is, the Internet can virtually provide a natural and excellent platform for illegal Internet-based activities, commonly known as cybercrimes [21] (e.g., hacking, online scam, credit card fraud, etc.). Cybercrime has become increasingly dependent on the online underground markets, especially underground forums, through which cybercriminals can not only acquisitive the tools, methods and ideas to commit cybercrimes, but also trade the illicit products (e.g., malware [9], [40], [42]), resources (e.g., website traffics), and services (e.g., hacking services [7], [41]). The emerging underground markets, especially underground forums, have enabled cybercriminals to realize considerable profits. For example, the estimated annual revenue for an individual credit card steal organization was $300 millions [23]; it's also revealed that a group of cybercriminals profited $864 millions per year by renting out the DDoS attacks [8], [20].

As underground forums (e.g., *Blackhat World*, *Hack Forums*, *Nulled*, *Free-hack*, etc.) have widely used by cybercriminals to trade the illicit products, resources and services, they have played a central role in the cybercriminal ecosystem [35]. Therefore, analysis of underground forums can provide invaluable insight into cybercrime, and thus facilitate the law enforcement communities, security researchers and industry practitioners to devise effective interventions to disrupt the illicit activities [11], [14], [15], [27], [35]. For example, based on an influx of stolen credit card numbers being advertised for sale on an online forum, Brian Krebs has successfully alerted Target to an ongoing massive data breach [1]. Unfortunately, the underground forums are run in a covert and dynamic environment, where the nature of trading behaviors are concealed. For example, without further information, it's hard to determine whether the thread of "*Amazon account. Pricing: $19.95...*" was a legitimate thread posted by an Amazon user to transfer his/her account or a cybercrime-suspected thread posted by a cybercriminal to sell the compromised account. To uncover the burgeoning information of this underground trove, human analysts need to continually spend a multitude of time to keep the latest statuses and variances of all threads and topics under observation. This calls for novel tools and methodologies to automate the analysis of underground forums

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

to gain insights into their behavioral processes.

To address the above challenges, in this paper, we design and develop an intelligent system called *iDetector* to automate the analysis of underground forums for the detection of cybercrime-suspected threads. In *iDetector*, to detect whether the given threads are cybercrime-suspected threads, we not only analyze the content in the threads, but also utilize the relations among threads, users, replies, and topics. For example, as shown in Fig. 1, to decide whether *Thread-1* is a cybercrime-suspected thread, using its content "*Amazon account. Pricing: $19.95...*" is not sufficient; however, with the further information: (1) there's another thread (*Thread-2*) "*Random Cracked Amazon account. Pricing & Content: per account $19.95...* ", (2) both *Thread-1* and *Thread-2* were posted by the same user (i.e., a cybercriminal), and (3) both threads discussed the same topic (i.e., "Amazon account"), it can be inferred that *Thread-1* is highly possible a cybercrime-suspected thread.

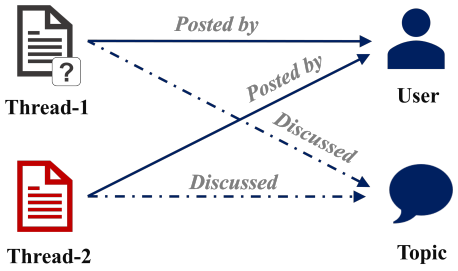

Fig. 1: Illustration of an HIN.

To model this kind of rich semantic relationships, in *iDetector*, we first introduce a structured heterogeneous information network (HIN) [12], [33] for representation, which is capable to be composed of different types of entities and relations. To capture the complex relationships (e.g., as illustrated in Fig. 1, two threads are relevant if they were posted by the same user and discussed the same topic), we use a meta-structure [19] based approach to characterize the semantic relatedness over threads. Then, we further integrate content-based similarity (i.e., similarity of posted threads) and relatedness depicted by each meta-structure to formulate a similarity measure over threads. Later, we build a classifier using Laplacian scores to aggregate different similarities formulated by different meta-structures to make predictions. In sum, our developed system *iDetector* which integrates the above proposed method has the following major traits:

- ***Novel feature representation to depict posted thread in underground forum:*** Instead of only using the content in the posted threads, we further utilize the rich relationships among threads, users, replies and topics (i.e., thread-user, thread-reply, thread-topic, reply-user and reply-topic relations) to represent the threads. Based on different kinds of relationships through different types of entities, the threads will be represented by a HIN, and a meta-structure based approach will be used to depict the relatedness between threads. To utilize both content- and relation-based information, we integrate similarity of the content information and relatedness depicted by each meta-structure to formulate a similarity measure over threads. The proposed solution provides a more feasible way to express the complex relationships among different types of entities (i.e., threads, users, replies, and topics) in underground forums than traditional approaches.
- ***Aggregation of different similarities for prediction:*** Different meta-structures capture the relatedness between threads at different views. The similarities over threads formulated by different meta-structures can be used to make decisions in an aggregated way. In this paper, we propose to aggregate different similarities using Laplacian scores to make predictions.
- ***A practical developed system to automate underground forum analysis:*** Based on the collected and annotated data from underground forums, we develop a practical system named *iDetector* to automate the analysis of underground forums for the detection of cybercrime-suspected threads. Comprehensive experimental studies are also provided to validate the performance of our developed system in comparisons with other alternative approaches.

The remainder of this paper is organized as follows: Section II introduces our system architecture. Section III presents our proposed method in detail. In Section IV, based on the real data collected and annotated from underground fourms, we systematically evaluate the performance of our proposed method in comparisons with other alternative approaches in cybercrime-suspected thread detection. Section V discusses the related work. Finally, Section VI concludes.

II. System Architecture

We develop a system called *iDetector* (shown in Fig. 2) to automate the analysis of underground forums for the detection of cybercrime-suspected threads.

For ***training***, it consists of four major components:

- ***Data Collector and Preprocessor.*** We first develop web crawling tools to collect the threads and their replies, as well as the users' profiles from underground forums. Note that the information of individual user is kept anonymous. For the collected threads and their replies, the preprocessor will further remove all the punctuations and stopwords, and then conduct lemmatization by using Stanford CoreNLP [24].
- ***Feature Extractor and HIN Constructor.*** A bag-of-words [39] feature vector will be extracted to represent each thread. Then, the relationships among threads, users, replies and topics will be further analyzed, such as, i) *user-post-thread* (i.e., a user posts a thread), ii) *user-write-reply* (i.e., a user writes a reply), iii) *reply-comment-thread* (i.e. a reply comments on a thread), iv) *thread-discuss-topic* (i.e. a thread discusses a specific topic) and v) *reply-hold-topic* (i.e. a reply holds (retains) a specific topic). Based on the extracted features, a structural HIN will be constructed. (See Section III-A for details.)
- ***Meta-structure Builder.*** In this module, different meta-structures are built from HIN to capture the relatedness between threads from different views. Then, we integrate

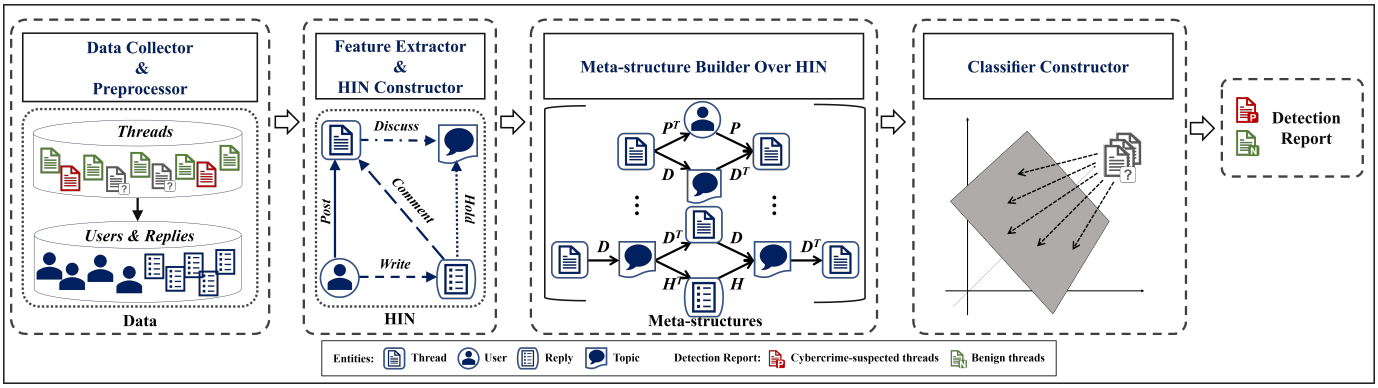

Fig. 2: System architecture of *iDetector*.

similarity of the posted threads and relatedness depicted by different meta-structures to formulate a set of similarity measures over threads. (See Section III-B for details.)

- ***Classifier Constructor.*** Given the similarity matrices over threads formulated by different meta-structures from the previous component, a classifier is build to aggregate different similarities using Laplacian scores for the detection of cybercrime-suspected threads in the underground forums. (See Section III-C for details.)

For ***prediction***, given an unlabeled thread and its replies, the content-based features will be extracted, and the above-mentioned relationships will be further analyzed; based on these extracted features and the constructed classification model, this thread will be labeled as either benign or cybercrime-suspected.

III. Proposed Method

In this section, we introduce the detailed approaches of how we represent underground forum threads utilizing both content- and relation-based features simultaneously, and how we solve the problem of cybercrime-suspected thread detection based on this representation.

A. HIN Construction

As the above discussion, to determine whether a thread is cybercrime-suspected thread, we not only use the content-based features, but also the complex relationships among threads, users, replies, and topics. To characterize the relatedness of two threads, we consider various kinds of relationships which include the followings.

R1: To describe the relation between a user and his/her posted thread, we generate the ***user-post-thread*** matrix $\mathbf{P}$ where each element $p_{i,j} \in \{0,1\}$ denotes if user i posts thread j.

R2: To denote the relation that a user writes a reply, we build the ***user-write-reply*** matrix $\mathbf{W}$ where each element $w_{i,j} \in \{0,1\}$ indicates whether user i writes reply j.

R3: To depict whether a reply comments on a specific thread, we build the ***reply-comment-thread*** matrix $\mathbf{C}$ where element $c_{i,j} \in \{0,1\}$ denotes if reply i comments on thread j.

R4: To represent the relation that a thread discusses a specific topic, we generate the ***thread-discuss-topic*** matrix $\mathbf{D}$ where each element $d_{i,j} \in \{0,1\}$ indicates whether thread i discusses topic j. In this application, we use Latent Dirichlet allocation (LDA) [6] for the topic extraction from the posted threads.

R5: To denote the relation that a reply retains (i.e., holds) a specific topic, we generate the ***reply-hold-topic*** matrix $\mathbf{H}$ where each element $h_{i,j} \in \{0,1\}$ indicates whether reply i holds topic j. Here, we also use LDA [6] for the topic extraction from users' replies.

A summary of the description of the above relations and elements in each relation matrix is shown in Table I.

TABLE I: The description of each matrix and its element.

Matrix	Element	Description
$\mathbf{P}$	$p_{i,j}$	If user i posts thread j, then $p_{i,j}=1$; otherwise, $p_{i,j}=0$.
$\mathbf{W}$	$w_{i,j}$	If user i writes reply j, then $w_{i,j}=1$; otherwise, $w_{i,j}=0$.
$\mathbf{C}$	$c_{i,j}$	If reply i comments on thread j, then $c_{i,j}=1$; otherwise, $c_{i,j}=0$.
$\mathbf{D}$	$d_{i,j}$	If thread i discusses topic j, then $d_{i,j}=1$; otherwise, $d_{i,j}=0$.
$\mathbf{H}$	$h_{i,j}$	If reply i holds topic j, then $h_{i,j}=1$; otherwise, $h_{i,j}=0$.

In order to depict threads, users, replies, topics and the rich relationships among them, it is important to model them in a proper way so that different kinds of relations can be better and easier handled. We introduce how to use HIN, which is capable to be composed of different types of entities and relations, to represent the threads by using the features described above. We first present the concept related to HIN.

Definition 1. Heterogeneous information network (HIN) [33]. A HIN is defined as a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$ with an entity type mapping ϕ: $\mathcal{V}\rightarrow\mathcal{A}$ and a relation type mapping ψ: $\mathcal{E}\rightarrow\mathcal{R}$, where $\mathcal{V}$ denotes the entity set and $\mathcal{E}$ is the relation set, $\mathcal{A}$ denotes the entity type set and $\mathcal{R}$ is the relation type set, and the number of entity types $|\mathcal{A}|>1$ or the number of relation types $|\mathcal{R}|>1$. The ***network schema*** [34] for a HIN $\mathcal{G}$, denoted as $\mathcal{T}_{\mathcal{G}}=(\mathcal{A},\mathcal{R})$, is a graph with nodes as entity types from $\mathcal{A}$ and edges as relation types from $\mathcal{R}$.

HIN not only provides the network structure of the data associations, but also provides a high-level abstraction of the categorical association. For the detection of cybercrime-

suspected threads, we have four entity types (i.e., thread, user, reply and topic) and five types of relations among them as described above. Based on the above definition, the network schema for HIN in our application is shown in Fig. 3.

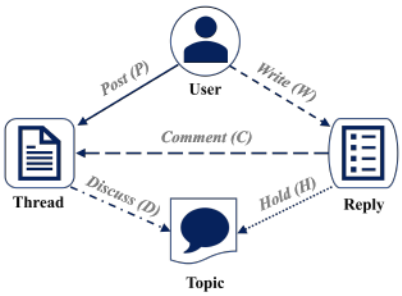

Fig. 3: Network schema for HIN.

B. Meta-structure Based Relatedness

The different types of entities and different relations between them motivate us to use a machine-readable representation to enrich the semantics of relatedness among threads. Meta-path [34] is used in the concept of HIN to formulate the semantics of higher-order relationships among entities. A ***meta-path*** [34] $\mathcal{P}$ is a path defined on the graph of network schema $\mathcal{T}_\mathcal{G} = (\mathcal{A}, \mathcal{R})$, and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_L} A_{L+1}$, which defines a composite relation $R = R_1 \cdot R_2 \cdot \ldots \cdot R_L$ between types A_1 and A_{L+1}, where $\cdot$ denotes relation composition operator, and L is the length of $\mathcal{P}$. An example of a meta-path for threads based on HIN schema shown in Fig. 3 is: $thread \xrightarrow{post^{-1}} user \xrightarrow{post} thread$, which states that two threads can be connected through the same user who posted them (i.e., P_1 in Fig. 4); another example is $thread \xrightarrow{comment^{-1}} reply \xrightarrow{write^{-1}} user \xrightarrow{write} reply \xrightarrow{comment} thread$, which denotes that two threads are related through their replies written by the same user (i.e., P_3 in Fig. 4) . Although meta-path can be used to depict the relatedness over threads in our application, it fails to capture a more complex relationship, such as two threads were posted by the same user and also discussed the same topic. This calls for a better characterization to handle such complex relationship.

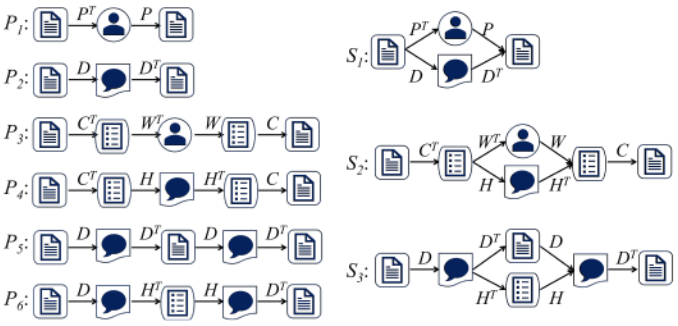

Fig. 4: Meta-paths (*left*) and meta-structures (*right*) on HIN. (The symbols in this figure are the abbreviations in Fig. 3.)

Meta-structure [19] is proposed to use a directed acyclic graph of entity and relation types to capture more complex relationship between two HIN entities. The concept of meta-structure is given as following [19].

Definition 4. Meta-structure [19]. A meta-structure $\mathcal{S}$ is a directed acyclic graph with a single source node n_s and a single target node n_t, defined on a HIN $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with schema $\mathcal{T}_\mathcal{G} = (\mathcal{A}, \mathcal{R})$. Formally, $\mathcal{S} = (N, M, n_s, n_t)$, where N is a set of nodes and M is a set of edges. For any node $x \in N, x \in \mathcal{A}$; for any link $(x, y) \in M, (x, y) \in \mathcal{R}$.

In our application, based on the HIN schema shown in Fig. 3, we generate three meaningful meta-structures to characterize the relatedness over threads (i.e., S_1-S_3 shown in Fig. 4): (1) S_1 depicts that two threads are related as they were posted by the same user and also discussed the same topic; (2) S_2 describes that two threads are connected since they had replies written by the same user and both their replies held (i.e., retained) the same topic; (3) S_3 denotes that two threads are relevant as they discussed the same topic which was also discussed in another thread and held (i.e., retained) in another reply. Actually, a meta-path is a special case of a meta-structure (e.g., P_1 and P_2 are particular cases of S_1). In Fig. 4, the meta-paths of P_1-P_6 (*left*) are the special cases of the constructed meta-structures of S_1-S_3 (*right*). But meta-structure is capable to express more complex relationship in a convenient way.

To compute the relatedness over threads using a particular meta-structure designed above, we use the commuting matrix [34], [44] to compute the counting-based similarity matrix for a meta-structure. Take S_2 as an example, the commuting matrix of threads computed using S_2 is $\mathbf{M}_{S_2} = \mathbf{C}^T[(\mathbf{W}^T\mathbf{W}) \circ (\mathbf{H}\mathbf{H}^T)]\mathbf{C}$, where **C, W, H** are the adjacency matrices between two corresponding entity types, $\circ$ denotes the Hadamard product [17] of two matrices. $\mathbf{M}_{S_2}(i, j)$ denotes the number of reply pairs commented on thread i and j that were written by the same users and held same topics. The commuting matrix of threads computed using S_1 is $\mathbf{M}_{S_1} = (\mathbf{P}^T\mathbf{P}) \circ (\mathbf{D}\mathbf{D}^T)$, whose element denotes the number of topic pairs discussed in thread i and j that were also posted by same users; while the commuting matrix of threads computed using S_3 is $\mathbf{M}_{S_3} = \mathbf{D}[(\mathbf{D}^T\mathbf{D}) \circ (\mathbf{H}^T\mathbf{H})]\mathbf{D}^T$, whose element denotes the number of topic pairs discussed in thread i and j that were also discussed in other threads and held in other replies.

After characterizing the relatedness of threads based on each meta-structure, we utilize both content- and relation-based information to measure the similarity over threads: we integrate similarity of threads' content and relatedness depicted by meta-structure to form a similarity measure matrix over threads. The similarity matrix over threads is denoted as **Z**, whose element is a combination of content-based similarity and meta-structure based relatedness. We define the similarity matrix $\mathbf{Z}_{\mathcal{S}_k}$ based on $\mathbf{M}_{\mathcal{S}_k}$ as:

$$\mathbf{Z}_{\mathcal{S}_k}(i, j) = [1 + ln(\mathbf{M}_{\mathcal{S}_k}(i, j) + 1)] \times Cos(t_i, t_j), \quad (1)$$

where $Cos(t_i, t_j)$ is the cosine similarity between thread i and j: for each thread, we convert them into a bag-of-words feature vector and then use cosine similarity measure [25] to estimate the closeness of two threads' content.

C. Classifier Combining Different Similarities

Different meta-structures capture the relatedness over threads at different views, i.e., $\boldsymbol{S}_1$-$\boldsymbol{S}_3$. Since HIN can naturally provide us different relatedness with different semantic meanings, instead of using a single meta-structure to depict the relatedness between threads, we propose to use Laplacian scores to weight the importance of different similarities based on different meta-structures for thread classification (i.e., whether a thread is cybercrime-suspected thread).

Suppose that there are K meta-structures $\mathcal{S}_k$ $(k = 1, 2, ..., K)$, we can calculate their corresponding commuting matrices $\mathbf{M}_{\mathcal{S}_k}$ $(k = 1, 2, ..., K)$. Then, we use Eq.(1) to compute the similarity matrix $\mathbf{Z}_{\mathcal{S}_k}$ $(k = 1, 2, ..., K)$ based on $\mathbf{M}_{\mathcal{S}_k}$. Following [36], [37], we combine different similarities to form a new similarity measure:

$$\mathbf{Z}'(i,j) = \frac{2 \times \sum_{k=1}^{K} w_k \mathbf{Z}_{\mathcal{S}_k}(i,j)}{\sum_{k=1}^{K} w_k \mathbf{Z}_{\mathcal{S}_k}(i,i) + \sum_{k=1}^{K} w_k \mathbf{Z}_{\mathcal{S}_k}(j,j)}, \quad (2)$$

where $\boldsymbol{w} = [w_1, w_2, ..., w_K]$ is the weighted vector of different similarities formulated by different meta-structures. In our application, we use Laplacian score [13] to learn the weight of each similarity, since it can be computed to reflect the locality preserving power of each feature. In this way, a new kernel is formed and we feed it to the Support Vector Machine (SVM) for classification. Note that if the matrix $\mathbf{Z}'$ is not a kernel (not a positive semi-definite matrix), we simply use the trick to remove the negative eigenvalues.

IV. Experimental Results And Analysis

In this section, we show three sets of experimental studies using real sample collections from Hack Forums (hackforums.net) to fully evaluate the performance of our developed system *iDetector* for automatic detection of cybercrime-suspected threads: (1) In the first set of experiments, based on HIN schema, we fully assess the performance of our proposed method; (2) In the second set of experiments, we evaluate our developed system *iDetector* which integrates our proposed method by comparisons with other alternative classification approaches; (3) Finally, we conduct case studies based on the detected cybercrime-suspected threads to gain deep insights into the behavioral processes of Hack Forums. The measures for evaluation of different methods are shown in Table II.

TABLE II: Performance indices for different methods.

Indices	Description
TP	# correctly classified as cybercrime-suspected threads
TN	# correctly classified as benign threads
FP	# mistakenly classified as cybercrime-suspected threads
FN	# mistakenly classified as benign threads
Precision	$TP/(TP+FP)$
Recall	$TP/(TP+FN)$
ACC	$(TP+TN)/(TP+TN+FP+FN)$
F1	$2*Precision*Recall/(Precision+Recall)$

A. Data Collection and Annotation

To obtain the data from Hack Forums, we develop a set of crawling tools to collect the posted threads and their replies as well as the users' profiles in a period of time. By the date, we have collected ***12,021 threads*** posted by ***5,571 users*** through March 2015 to December 2017. Note that the information for individual user is kept anonymous. After data collection and preprocessing, the five relationships (i.e., *R1-R5*) introduced in Section III-A are further extracted.

To obtain the prelabeled data for training, three groups of annotators (i.e., ***15 persons***) with knowledge from domain professional (i.e., cybersecurity researcher) ***spent 45 days to label*** whether the collected threads are cybercrime-suspected threads or not by cross-validations. The mutual agreement is above 95%, and only the ones with agreements are retained; that is, for the collected threads, 4,304 are labeled as cybercrime-suspected threads and 7,717 are labeled as benign.

B. Evaluation of the Proposed Method

In this set of experiments, based on the annotated dataset described in Section IV-A, we fully evaluate our proposed method by 10-fold cross-validations: (1) based on the HIN schema (as described in Section III-A), we first evaluate the performance of meta-structure based method in cybercrime-suspected thread detection by comparisons with meta-path based approach; (2) we then evaluate the proposed method using Laplacian scores for aggregation of different similarities formulated by different meta-structures.

We first construct three meta-structures (i.e., $\boldsymbol{S}_1$-$\boldsymbol{S}_3$ shown in Fig. 4: *right*) and generate the corresponding six meta-paths (i.e., $\boldsymbol{P}_1$-$\boldsymbol{P}_6$ shown in Fig. 4: *left*). To measure the similarity over threads, we integrate similarity of posted threads and relatedness depicted by each meta-structure or meta-path to form a similarity measure matrix. We evaluate their performances for cybercrime-suspected thread detection using SVM. For each meta-structure or meta-path, the generated similarity measure matrix is used as the kernel fed to SVM. For SVM, we use LibSVM in our experiments. The penalty is empirically set to be 1,000 and other parameters are set by default.

TABLE III: Evaluation of the proposed method.

ID	Kernel	Commuting Matrix	ACC	F1
$\boldsymbol{P}_1$	$\mathbf{Z}_{\mathcal{P}_1}$	$\mathbf{P}^T\mathbf{P}$	0.817	0.786
$\boldsymbol{P}_2$	$\mathbf{Z}_{\mathcal{P}_2}$	$\mathbf{D}\mathbf{D}^T$	0.814	0.785
$\boldsymbol{P}_3$	$\mathbf{Z}_{\mathcal{P}_3}$	$\mathbf{C}^T\mathbf{W}^T\mathbf{W}\mathbf{C}$	0.806	0.776
$\boldsymbol{P}_4$	$\mathbf{Z}_{\mathcal{P}_4}$	$\mathbf{C}^T\mathbf{H}\mathbf{H}^T\mathbf{C}$	0.791	0.757
$\boldsymbol{P}_5$	$\mathbf{Z}_{\mathcal{P}_5}$	$\mathbf{D}\mathbf{D}^T\mathbf{D}\mathbf{D}^T$	0.809	0.779
$\boldsymbol{P}_6$	$\mathbf{Z}_{\mathcal{P}_6}$	$\mathbf{D}\mathbf{H}^T\mathbf{H}\mathbf{D}^T$	0.801	0.768
$\boldsymbol{S}_1$	$\mathbf{Z}_{\mathcal{S}_1}$	$(\mathbf{P}^T\mathbf{P}) \circ (\mathbf{D}\mathbf{D}^T)$	0.843	0.816
$\boldsymbol{S}_2$	$\mathbf{Z}_{\mathcal{S}_2}$	$\mathbf{C}^T[(\mathbf{W}^T\mathbf{W}) \circ (\mathbf{H}\mathbf{H}^T)]\mathbf{C}$	0.825	0.801
$\boldsymbol{S}_3$	$\mathbf{Z}_{\mathcal{S}_3}$	$\mathbf{D}[(\mathbf{D}^T\mathbf{D}) \circ (\mathbf{H}^T\mathbf{H})]\mathbf{D}^T$	0.833	0.807
10	**Combined-kernel (3)**		**0.860**	**0.833**

The results in Table III show that each meta-structure does perform better than its corresponding meta-paths. For example, meta-paths of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are special cases of meta-structure

$\boldsymbol{S}_1$; but $\boldsymbol{S}_1$ works better than $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ in the problem of cybercrime-suspected thread detection. The reason behind this is that meta-structure is more expressive to characterize a complex relatedness over threads than meta-path. This also demonstrates that we can use meta-structure with subtle differences to significantly improve the quality of relation-based features and better express different relatedness over threads in our application.

We then combine all the generated similarity matrices formulated by the three different meta-structures (i.e., $\boldsymbol{S}_1$-$\boldsymbol{S}_3$) using Laplacian scores as the weights to construct a more powerful kernel (i.e., ID10) fed to SVM (as described in Section III-C). From the results shown in Table III, we can observe that Laplacian score indeed helps us select the important similarities. The "Combined-kernel (3)" is with 86.0% accuracy and 0.833 F1, which successfully outperforms any single similarities formulated by different meta-structures. This also shows Laplacian score can better reflect classification property and thus improve the cybercrime-suspected detection performance. To demonstrate the effectiveness of Laplacian score, we further study the correlation between each similarity and their related Laplacian score. From Fig. 5, we can see that the Laplacian score can successfully filter out the performance of each similarity. We also further evaluate the parameter sensitivity of our proposed method with different values of the penalty parameter C. From Fig. 6, we can see in a wide range of numbers, the performance of combined similarity is stable and not very sensitive to the penalty parameter. This indicates that for practical use, we can simply tune a parameter using some training data based on cross-validations, and apply that parameter to the test set without concerning the change of the parameter affecting the online performance.

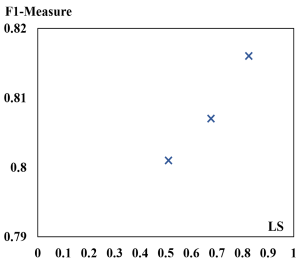

Fig. 5: Effectiveness of LS.

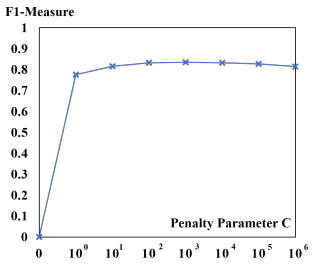

Fig. 6: Parameter sensitivity.

C. Comparisons with Other Alternative Methods

In this set of experiments, based on the dataset described in Section IV-A, we compare *iDetector* which integrate our proposed method described in Section III with other alternative methods by 10-fold cross-validations. For these methods, we construct four types of features:

- **Bag-of-words** [39]: Each thread is represented as a bag-of-words feature vector.
- **LDA** [6]: Base on the LDA model, each thread is represented as a LDA-based feature vector. In our case, a tool provided by Blei [5] is used to train the LDA model and the number of topics is empirically set to 50.
- **Word2vec** [26]: Each thread is presented as a weighted word2vec feature vector using the skip-gram model. In our application, we use the word2vec tool provided by Google [28] and the dimension of the word vectors is empirically set to 50 while other parameters are set by default.
- **Augmented**: This augments bag-of-words with relations of *R1-R5* described in Section III-A as flat features.

Based on these features, we consider two typical classification models, i.e., Naive Bayes (NB) and SVM. The experimental results are illustrated in Table IV. From the results, we can see that, compared with the other three features, feature engineering (i.e., denoted as "augmented") helps the performance of machine learning, since the rich semantics encoded in different types of relations can bring more information. However, the use of this information for traditional machine learning algorithms is simply flat features, i.e., concatenation of different features altogether. The results in Table IV also show that *iDetector* further outperforms all these alternative classification methods in automatic cybercrime-suspected thread detection. To check whether the overall improvement is significant, we also run 20 random trials of training and testing examples to compare *iDetector* and SVM with feature engineering, and the probability associated with a paired t-Test [3] with a two-tailed distribution is 3.17×10^{-14}. This shows that *iDetector* is significantly better than the best baseline method we compared. The reason behind this is that, in *iDetector*, we use more expressive representation for the data, and build the connection between the higher-level semantics of the data and the final results.

TABLE IV: Comparisons with other alternative methods.

ID	Method	Feature	ACC	F1
1	NB	Bag-of-words	0.726	0.671
2		LDA	0.737	0.684
3		Word2vec	0.770	0.723
4		Augmented	0.775	0.728
5	SVM	Bag-of-words	0.759	0.715
6		LDA	0.773	0.731
7		Word2vec	0.790	0.748
8		Augmented	0.796	0.756
9	**iDetector**		**0.860**	**0.833**

In this set of experiments, we further evaluate the scalability and stability of our developed system *iDetector*. For scalability evaluation, we evaluate the training time of our proposed method with different sizes of the training data sets. The scalability is shown in Fig. 7. It is illustrated that the running time is quadratic to the number of training samples. When dealing with more data, approximation or parallel algorithms should be developed. Therefore, for practical use, our approach is feasible for real application in automatic cybercrime-suspected thread detection. For stability evaluation, Fig. 8 shows the overall receiver operating characteristic (ROC) curves based on the 10-fold cross validations, from which we can see that

iDetector achieves an impressive 0.877 average TP rate for cybercrime-suspected threads detection.

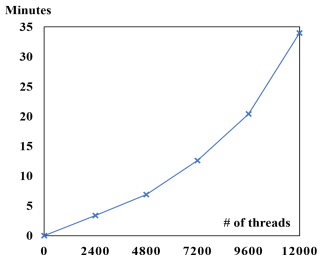

Fig. 7: Scalability evaluation.

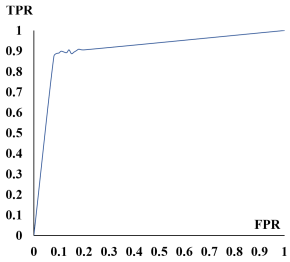

Fig. 8: Stability evaluation.

D. Case Studies

In this section, after the automatic detection of cybercrime-suspected threads from Hack forums using our developed system *iDetector*, to better understand and gain deep insights into its behavioral process, we further analyze and categorize those cybercrime-suspected threads and have several valuable findings. Table V shows the distribution of cybercrime-suspected threads in different sub-boards, from which we can observe that some sub-boards, such as "*Premium Sellers Section*" and "*Online Accounts*", play more significant roles than others in Hack forums. Table VI shows different topics discussed by these detected cybercrime-suspected threads, which indicates that "*Cracked Account and Phishing*" and "*Password hacking*" are the most prevalent illicit services provided on Hack Forums. One of the interesting findings is that cybercriminals sell the cracked individual credit card information by exploiting the vulnerabilities of online banking systems. This study also reveals that: (1) to protect legitimate users' privacy, there is an imminent need to improve the security of online banking services; (2) using the automatic tools to perform the surveillance of underground forums can be a valuable and supplementary way to facilitate the understanding of the behavioral processes of cybercrimes.

TABLE V: Distribution of cybercrime-suspected threads in different sub-boards. ("cs-threads" represent the detected cybercrime-suspected threads)

Sub-boards	cs-threads	Replies	Users
Marketplace Discussions	366	2,875	415
Premium Sellers Section	1146	9,002	1,298
Secondary Sellers Market	646	5,074	732
Online Accounts	897	7,046	1,016
Others	1249	9,811	1,415

V. Related Work

There have been many research efforts on automated analysis of online underground forums [2], [14], [22], [27], [29], [31], [32], [35], [38]. While existing research results are encouraging, many of these works systematized the analysis of the forums into a framework and only made an aggregate summary of forum activities without providing methodology [14], [35], or proposed some promising tools to study particular topics only utilizing the content information in the posted threads (e.g., SVM classifiers associated with LDA model based on the content in the posted threads were built to understand the functions and characteristics of assets in underground forums [30]), which still leave a large room for improvement. Different from these existing works, in this paper, we propose to utilize not only the content information in the posted threads, but also the relationships among threads, users, replies and topics (i.e., thread-user, thread-reply, thread-topic, reply-user and reply-topic relations) for cybercrime-suspected thread detection. Based on the extracted features, the threads are represented by a structured HIN.

TABLE VI: Different topics discussed in the detected cybercrime-suspected threads.

Topics of cybercrime-suspected threads	# threads	Percentage
Cracked Account and Phishing	1164	27.04%
Password hacking	613	14.24%
Web exploit and Vulnerabilities	531	12.34%
Software cracking and Crypters	523	12.15%
SQL injection	334	7.76%
Malware and Virus	312	7.25%
Others	827	19.21%

HIN is used to model different types of entities and relations [33]. It has been applied to various applications, such as scientific publication network analysis [34], biomedical knowledge mining [10], [43] and malware detection [18]. Several studies have already investigated the use of HIN information for relevance computation, however, most of them only use meta-path [34] to measure the similarity. Such simple path structure fails to capture a more complex relationship between two entities. To address this problem, Huang et al. [19] proposed to use meta-structure, which is a directed acyclic graph of entity and relation types to measure the proximity between two entities. Their work only considered one particular meta-structure to capture the relatedness over entities. Different from these works, in this paper, we consider different meta-structures which characterize the relatedness over threads at different views, and further propose a solution to aggregate different similarities formulated by different meta-structures. To the best of our knowledge, this is the first attempt to use structural HIN to automate the analysis of underground forums for cybercrime-suspected thread detection.

VI. Conclusion

In this paper, we design and develop an intelligent system named *iDetector* to automate the analysis of underground forums for cybercrime-suspected thread detection. In *iDetector*, we first construct a structural HIN to leverage the information of threads, users, replies and topics as well as the rich relationships among them, which gives the thread a higher-level semantic representation. Then, meta-structure based approach is used to characterize the semantic relatedness over threads. Afterwards, we integrate content-based similarity

and the relatedness depicted by each meta-structure to formulate a similarity measure over threads. We then use Laplacian scores to aggregate different similarities formulated by different meta-structures to build a classifier for cybercrime-suspected thread detection. The promising experimental results based on the collected and annotated data from Hack forums demonstrate that *iDetector* integrated our propose method outperforms other alternative approaches.

ACKNOWLEDGMENT

This work is partially supported by the U.S. National Science Foundation under grants CNS-1618629 and CNS-1814825, WV Higher Education Policy Commission Grant (HEPC.dsr.18.5), and WVU Research and Scholarship Advancement Grant (R-844).

REFERENCES

[1] L. Ablon, M. C. Libicki, and A. A. Golay, *Markets for cybercrime tools and stolen data: Hackers' bazaar*. Rand Corporation, 2014.

[2] E. Asiedu and T. Stengos, "An empirical estimation of the underground economy in ghana," *Economics Research International*, vol. 2014, 2014.

[3] P. Baldi and A. D. Long, "A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, 2001.

[4] K. T. Blagoeva and M. Mijoska, "Applying tam to study online shopping adoption among youth in the republic of macedonia."

[5] D. M. Blei and J. D. Lafferty, "Topic models," *Text mining: classification, clustering, and applications*, vol. 10, no. 71, p. 34, 2009.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, no. Jan, pp. 993–1022, 2003.

[7] L. Chen, S. Hou, and Y. Ye, "Securedroid: Enhancing security of machine learning-based detection against adversarial android malware attacks," in *ACSAC*. ACM, 2017, pp. 362–372.

[8] Damballa, "Want to rent an 80-120k ddos botnet?" in *https://www.damballa.com/want-to-rent-an-80-120k-ddos-botnet/*, 2014.

[9] Y. Fan, Y. Ye, and L. Chen, "Malicious sequential pattern mining for automatic malware detection," *Expert Systems with Applications*, vol. 52, pp. 16–25, 2016.

[10] Y. Fan, Y. Zhang, Y. Ye, X. Li, and W. Zheng, "Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from twitter and case studies," in *CIKM*. ACM, 2017, pp. 1259–1267.

[11] J. Franklin, A. Perrig, V. Paxson, and S. Savage, "An inquiry into the nature and causes of the wealth of internet miscreants." in *CCS*, 2007, pp. 375–388.

[12] J. Han, Y. Sun, X. Yan, and P. S. Yu, "Mining knowledge from databases: an information network analysis approach," in *SIGMOD*. ACM, 2010, pp. 1251–1252.

[13] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *NIPS*, 2006, pp. 507–514.

[14] T. J. Holt, "Examining the forces shaping cybercrime markets online," *Social Science Computer Review*, vol. 31, no. 2, pp. 165–177, 2013.

[15] T. J. Holt and E. Lampke, "Exploring stolen data markets online: products and market forces," *Criminal Justice Studies*, vol. 23, no. 1, pp. 33–50, 2010.

[16] Y. Hong, "Pivot to internet plus: Molding chinas digital economy for economic restructuring?" *International Journal of Communication*, vol. 11, pp. 1486–1506, 2017.

[17] R. A. Horn, "The hadamard product," in *Proc. Symp. Appl. Math*, vol. 40, 1990, pp. 87–169.

[18] S. Hou, Y. Ye, Y. Song, and M. Abdulhayoglu, "Hindroid: An intelligent android malware detection system based on structured heterogeneous information network," in *KDD*. ACM, 2017, pp. 1507–1515.

[19] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li, "Meta structure: Computing relevance in large heterogeneous information networks," in *KDD*. ACM, 2016, pp. 1595–1604.

[20] M. Karami and D. McCoy, "Understanding the emerging threat of ddos-as-a-service," in *The 6th USENIX Workshop on Large-Scale Exploits and Emergent Threats*, 2013.

[21] E. Kraemer-Mbula, P. Tang, and H. Rush, "The cybercrime ecosystem: Online innovation in the shadows?" *Technological Forecasting and Social Change*, vol. 80, no. 3, pp. 541–555, 2013.

[22] X. Liao, K. Yuan, X. Wang, Z. Pei, H. Yang, J. Chen, H. Duan, K. Du, E. Alowaisheq, S. Alrwais *et al.*, "Seeking nonsense, looking for trouble: Efficient promotional-infection detection through semantic inconsistency search," in *S&P*, 2016, pp. 707–723.

[23] G. D. Maio, A. Kapravelos, Y. Shoshitaishvili, C. Kruegel, and G. Vigna, "Pexy: The other side of exploit kits," in *Proceedings of the International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*, 2014, pp. 132–151.

[24] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[25] R. Mihalcea, C. Corley, C. Strapparava *et al.*, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, vol. 6, 2006, pp. 775–780.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[27] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *Proceedings of the ACM Internet Measurement Conference*, 2011, pp. 71–80.

[28] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[29] R. S. Portnoff, S. Afroz, G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, D. McCoy, K. Levchenko, and V. Paxson, "Tools for automated analysis of cybercriminal markets," in *WWW*, 2017, pp. 657–666.

[30] S. Samtani, R. Chinn, and H. Chen, "Exploring hacker assets in underground forums," in *ISI*. IEEE, 2015, pp. 31–36.

[31] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem." in *USENIX Security Symposium*, 2015, pp. 33–48.

[32] B. Stone-Gross, R. Abman, R. A. Kemmerer, C. Kruegel, D. G. Steigerwald, and G. Vigna, "The underground economy of fake antivirus software," *Economics of information security and privacy III*, pp. 55–78, 2013.

[33] Y. Sun and J. Han, "Mining heterogeneous information networks: principles and methodologies," *DMKD*, vol. 3, no. 2, pp. 1–159, 2012.

[34] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *VLDB*, vol. 4, no. 11, pp. 992–1003, 2011.

[35] K. Thomas, D. Yuxing, H. David, W. Elie, B. C. Grier, T. J. Holt, C. Kruegel, D. McCoy, S. Savage, and G. Vigna, "Framing dependencies introduced by underground commoditization," in *WEIS*, 2015.

[36] C. Wang, Y. Song, H. Li, and J. Zhang, "Text classification with heterogeneous information network kernels," in *AAAI*, 2016, pp. 2130–2136.

[37] C. Wang, Y. Song, H. Li, M. Zhang, and J. Han, "Knowsim: A document similarity measure on structured heterogeneous information networks," in *ICDM*. IEEE, 2015, pp. 1015–1020.

[38] H. Yang, X. Ma, K. Du, Z. Li, H. Duan, X. Su, G. Liu, Z. Geng, and J. Wu, "How to learn klingon without a dictionary: Detection and measurement of black keywords used by the underground economy," in *S&P*, 2017, pp. 751–769.

[39] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *MIR*. ACM, 2007, pp. 197–206.

[40] Y. Ye, L. Chen, S. Hou, W. Hardy, and X. Li, "Deepam: a heterogeneous deep learning framework for intelligent malware detection," *Knowledge and Information Systems*, vol. 54, no. 2, pp. 265–285, 2018.

[41] Y. Ye, T. Li, D. Adjeroh, and S. S. Iyengar, "A survey on malware detection using data mining techniques," *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, p. 41, 2017.

[42] Y. Ye, T. Li, S. Zhu, W. Zhuang, E. Tas, U. Gupta, and M. Abdulhayoglu, "Combining file content and file relations for cloud based malware detection," in *KDD*. ACM, 2011, pp. 222–230.

[43] Y. Zhang, Y. Fan, Y. Ye, X. Li, and W. Zheng, "Detecting opioid users from twitter and understanding their perceptions toward mat," in *ICDMW*. IEEE, 2017, pp. 502–509.

[44] H. Zhao, Q. Yao, J. Li, Y. Song, and D. L. Lee, "Meta-graph based recommendation fusion over heterogeneous information networks," in *KDD*. ACM, 2017, pp. 635–644.