Key Player Identification in Underground Forums over Attributed Heterogeneous Information Network Embedding Framework

Yiming Zhang, Yujie Fan, Yanfang Ye*
Department of CDS, Case Western Reserve University, OH, USA

Liang Zhao
Department of IST
George Mason University
VA, USA

Chuan Shi
School of CS
Beijing University of Posts and Telecommunications, Beijing, China

ABSTRACT

Online underground forums have been widely used by cybercriminals to exchange knowledge and trade in illicit products or services, which have played a central role in the cybercriminal ecosystem. In order to combat the evolving cybercrimes, in this paper, we propose and develop an intelligent system named *iDetective* to automate the analysis of underground forums for the identification of key players (i.e., users who play the vital role in the value chain). In *iDetective*, we first introduce an attributed heterogeneous information network (AHIN) for user representation and use a meta-path based approach to incorporate higher-level semantics to build up relatedness over users in underground forums; then we propose *Player2Vec* to efficiently learn node (i.e., user) representations in AHIN for key player identification. In *Player2Vec*, we first map the constructed AHIN to a multi-view network which consists of multiple single-view attributed graphs encoding the relatedness over users depicted by different designed meta-paths; then we employ graph convolutional network (GCN) to learn embeddings of each single-view attributed graph; later, an attention mechanism is designed to fuse different embeddings learned based on different single-view attributed graphs for final representations. Comprehensive experiments on the data collections from different underground forums (i.e., Hack Forums, Nulled) are conducted to validate the effectiveness of *iDetective* in key player identification by comparisons with alternative approaches.

CCS CONCEPTS

• **Security and privacy** → **Web application security**; • **Networks** → *Online social networks*; • **Computing methodologies** → **Machine learning algorithms**.

KEYWORDS

Underground forums; key player identification; attributed heterogeneous information network; network embedding

*Corresponding author: yanfang.ye@case.edu

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357876

ACM Reference Format:
Yiming Zhang, Yujie Fan, Yanfang Ye, Liang Zhao, and Chuan Shi. 2019. Key Player Identification in Underground Forums over Attributed Heterogeneous Information Network Embedding Framework. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3357876

1 INTRODUCTION

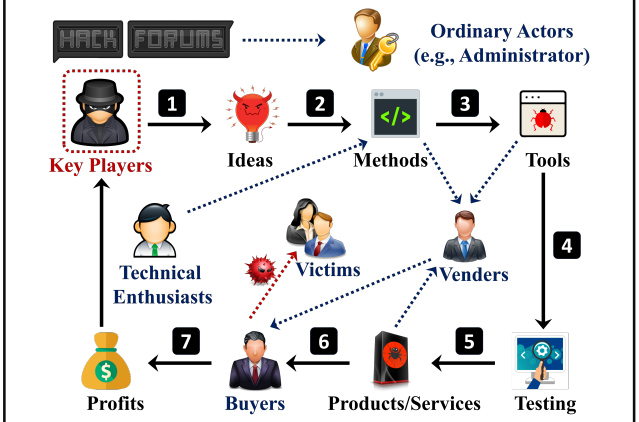
As the Internet has become one of the most important drivers in the global economy (e.g., worldwide e-commerce sales reached over $2.3 trillion dollars in 2017 and its revenues are projected to grow to $4.88 trillion dollars in 2021 [31]), it also provides an open and shared platform by dissolving the 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 (e.g., hacking, online scam, credit card fraud). Cybercrimes have become increasingly dependent on the online underground forums, through which cybercriminals can exchange knowledge (including ideas, methods and tools) and trade in illicit products (e.g., malware, stolen credit cards) or services (e.g., hacking services, bogus Amazon reviews). The emerging underground forums, such as Hack Forums [14], Nulled [28], and Black Hat World [4], have enabled cybercriminals to realize considerable profits. For example, the estimated annual revenue for an individual credit card steal organization was $300 millions [26]; it's also revealed that a group of cybercriminals could profit $864 millions per year by renting out the DDoS attacks [20].

Underground forums, which provide the platforms for cybercriminals to exchange knowledge and trade in illicit products or services that facilitate all stages of cybercrimes, have played a central role in the cybercriminal ecosystem. As one of the most prevalent underground forums, Hack Forums consists of 640,820 registered users with 3,621,989 posted threads containing 790,286 illicit products or services. We here use Hack Forums as a showcase to investigate the profit model and monetization process. As shown in Figure 1, the participants in the value chain can be categorized into different groups according to the roles they play: (1) ***Key players***: This group of users play a vital role in the value chain, as they are the "decathlon" who are capable of exploiting and disseminating vulnerabilities, developing and testing malicious tools, selling and monetizing illicit products or services. For example, as shown in Figure 1, we found that a Hack Forums user "Ban***s" (we here anonymize his user name) first ① analyzed the market demand

for social media account hacking services, then ② - ③ devised a method to develop cracking tools exploiting a cookie vulnerability to perform brute-force attacks on social media accounts, later ④ purchased some compromised Facebook accounts to test his developed tools, and thus ⑤ - ⑦ sold and monetized his hacking services. We also found that "Ban***s" shared his developed cracking tools with other Hack Forums users for free, who could further use it for profits. The skilled individuals like "Ban***s" are heavily relied upon by peers within the communities (e.g., novice hackers or newcomers can gain knowledge from them to cultivate their own specialties). (2) ***Non-key players***: Other users including vendors, buyers, victims, technical enthusiasts, administrators and moderators are categorized as non-key player. Considering the vital role of key players play in the value chain (i.e., their vulnerability exploiting capability, technical skills, cashing channel, and influence on others), it is important to identify key players in underground forums to facilitate the deployment of effective countermeasures to disrupt the illicit activities. Toward this goal, human analysts need to continually spend a multitude of time to keep the latest statuses and variances of the activities in underground forums under observation. This calls for novel tools and methodologies to automate the identification of key players to enable the law enforcement and security practitioners to devise effective interventions.

To address the above challenges, in this paper, we design and develop an intelligent system called *iDetective* to automate the analysis of underground forums for key player identification. In *iDetective*, we first introduce an attributed heterogeneous information network (AHIN) [32] to represent the rich relationship among users, threads, replies, comments and sections, and then use a meta-path based approach [33] to incorporate higher-level semantics to build up relatedness over users in underground forums. Then, we propose *Player2Vec* to efficiently learn node (i.e., underground forum user) representations in AHIN for key player identification. In *Player2Vec*, (1) we first map the constructed AHIN to a multi-view network which consists of K single-view attributed graphs encoding the relatedness over users depicted by K designed meta-paths; and then (2) we employ graph convolutional network (GCN) to learn embeddings of each single-view attributed graph; later, (3) an attention mechanism is designed to fuse different embeddings learned based on different single-view attributed graphs for final representations. The major contributions of our work are summarized as follows:

- We propose *a novel yet natural feature representation to depict underground forum users*. In our application for key player identification in underground forums, an AHIN is introduced to represent the rich semantic relationships between users and other entities (i.e., threads, comments, replies, and sections), and then a meta-path based approach is presented to characterize the relatedness over users.
- We propose *an efficient yet elegant AHIN representation learning model* (i.e., *Player2Vec*) to learn low-dimensional representations for nodes in AHIN. The proposed model leverage structural relations represented by AHIN and attributed information attached on nodes (i.e., users) to learn latent representations of nodes.
- We develop *an automatic system* called *iDetective* to identify key players in underground forums to facilitate law enforcement and cybersecurity practitioners to devise effective interventions to combat the evolving cybercrimes. Based on the large-scale data collected from different underground forums (i.e., Hack Forums, Nulled) and the pre-labeled ground-truth, comprehensive experimental studies are performed to validate the effectiveness of *iDetective* in key player identification by comparisons with baselines. Although we use underground forums as a showcase, the proposed method and the developed system can be easily expanded to other social media platforms.



Role of key players in the value chain in Hack Forums

1. Key players generate ideas by analysis of market demands
2. Exploit the vulnerabilities and devise the methods
3. Develop the tools using the devised methods
4. Testing the tools by purchasing some accounts and bots
5. Produce and advertise illicit products or services
6. Sell the illicit products or services
7. Monetize the illicit products or services

Figure 1: Participants in the value chain in Hack Forums.

The rest of the paper is organized as follows. Section 2 presents the system overview of *iDetective*. Section 3 introduces our proposed method in detail. Based on the large-scale data collections from different underground forums and the pre-labeled ground-truth, Section 4 systematically evaluates the performance of our developed system *iDetective*. Section 5 discusses the related work. Finally, Section 6 concludes.

2 SYSTEM OVERVIEW

The system overview of *iDetective* is shown in Figure 2, which consists of the following components:

- ***Data Collector and Preprocessor.*** A set of web crawling tools are developed to collect the data from Hack Forums and Nulled, including users' profiles (which are fully anonymized) and posted content (i.e., threads, replies and comments). For users' posted content, the preprocessor will further remove all the punctuations and stopwords, and then conduct lemmatization by using Stanford CoreNLP [27].

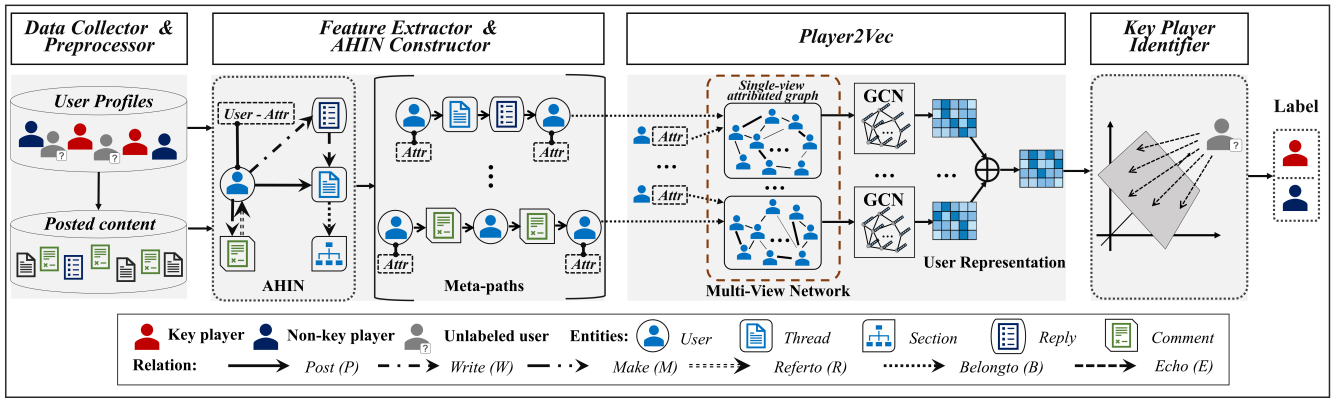


Figure 2: System overview of *iDetective*.

- ***Feature Extractor.*** Based on the data collected and preprocessed from the previous module, it extracts various kinds of relations, i.e., *user-thread*, *user-reply*, *user-comment*, *thread-section*, *reply-thread* and *comment-user* relations.
- ***AHIN Constructor.*** Based on the above extracted features, an AHIN is first presented to model the relations among different types of entities; and then different mesta-paths are built from the AHIN to capture the relatedness over users in underground forums from different views.
- ***Player2Vec.*** It first maps the constructed AHIN to a multi-view network which consists of K single-view attributed graphs encoding the relatedness over users depicted by K designed meta-paths; and then it employs GCN model to learn embeddings of each single-view attributed graph; later it exploits an attention mechanism to fuse different embeddings learned based on different single-view attributed graphs for final representations.
- ***Key Player Identifier.*** After employing *Player2Vec*, the learned low-dimensional vectors of nodes (i.e., users) will be fed to train the classifier, based on which the unlabeled users can be predicted as either *key player* or *non-key player*.

3 PROPOSED METHOD

In this section, we introduce the detailed approach of how we represent underground forum users, and how we solve the problem of key player identification based on the representation.

3.1 Feature Extraction

To comprehensively depict underground forum users, we not only consider users' attributed information but also various kinds of relationships as described below.

User's attributed information. In underground forums, users' profile information plays an important role for the identification. Therefore, we extract *username* and *contact information* from their profiles. Note that, for username, we first apply standard string matching techniques to measure the similarity of two usernames: if the similarity is greater than a user-specific threshold, we regard these two usernames as the same (e.g., "KD***Donald" and "KD***D0nald"). Then, we apply one-hot encoding to convert the extracted features to a binary feature vector. To obtain better user representations, we also consider each user's posted text content including threads, replies and comments. For text content, we propose to exploit *doc2vec* [22] to convert each text of variant size into a fixed length feature vector. Then we concatenate these two feature vectors to form an attributed feature vector for each user.

Relation-based Features. To comprehensively depict underground forum users, we consider various kinds of relationships including:

- ***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 represent whether a user makes a comment, we generate the ***user-make-comment*** matrix $\mathbf{M}$ where each element $m_{i,j} \in \{0, 1\}$ indicates whether user i makes comment j.
- ***R4***: To depict if a thread belongs to a section (i.e., there are many different sections in underground forums, such as "Hacking Tools and Programs", "Monetizing Techniques", "Premium Sellers Section", etc.), we generate the ***thread-belongto-section*** matrix $\mathbf{B}$ where each element $b_{i,j} \in \{0, 1\}$ denotes if thread i belongs to section j.
- ***R5***: To describe whether a reply echoes (i.e., responds to) a thread, we build the ***reply-echo-thread*** matrix $\mathbf{E}$ where element $e_{i,j} \in \{0, 1\}$ denotes if reply i echoes thread j.
- ***R6***: To denote the relation that a comment refers to (i.e., is made to) a user, we generate the ***comment-referto-user*** matrix $\mathbf{R}$ where each element $r_{i,j} \in \{0, 1\}$ indicates if comment i refers to user j.

3.2 AHIN Construction

For the above extracted features, though heterogeneous information network (HIN) [33] has shown the success of modeling different types of entities and relations, it has limited capability of modeling additional attributes attached to entities. To address this challenge, we propose to use attributed heterogeneous information network (AHIN) for representation.

Definition 1. ***Attributed Heterogeneous Information Network (AHIN)*** *[25]. Let $\mathcal{T} = \{T_1, ..., T_m\}$ be a set of m entity types. For each entity type T_i, let $\mathcal{X}_i$ be the set of entities of type T_i and A_i be the set of attributes defined for entities of type T_i. An entity x_j of type T_i is associated with an attribute vector $\mathbf{f}_j = (f_{j1}, f_{j2}, ..., f_{j|A_i|})$. An AHIN is defined by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ with an* entity *type mapping $\phi: \mathcal{V} \rightarrow \mathcal{T}$ and a* relation *type mapping $\psi: \mathcal{E} \rightarrow \mathcal{R}$, where $\mathcal{V} = \bigcup_{i=1}^{m} \mathcal{X}_i$ denotes the entity set and $\mathcal{E}$ is the relation set, $\mathcal{T}$ denotes the entity type set and $\mathcal{R}$ is the relation type set, $\mathcal{A} = \bigcup_{i=1}^{m} A_i$, and the number of entity types $|\mathcal{T}| > 1$ or the number of relation types $|\mathcal{R}| > 1$. The* ***network schema*** *[25] for an AHIN $\mathcal{G}$, denoted by $\mathcal{T}_\mathcal{G} = (\mathcal{T}, \mathcal{R})$, is a graph with nodes as entity types from $\mathcal{T}$ and edges as relation types from $\mathcal{R}$.*

For our case, i.e., the identification of key players in underground forums, we have five entity types (i.e., user, thread, reply, comment and section) and six types of relations among them (i.e., ***R1-R6***); furthermore, nodes with entity type of user is also attached with attributed feature vectors described in Section 3.1. Based on the definition above, the network schema for AHIN in our application is shown in Figure 3, which enables the underground forum users to be represented in an expressive way.

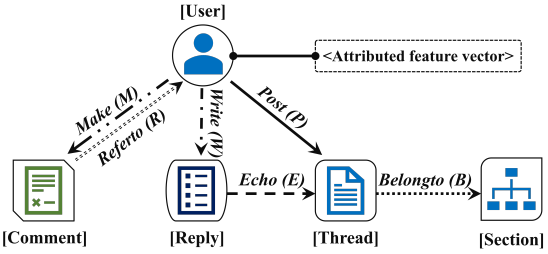


Figure 3: Network schema for AHIN.

The different types of entities and relations motivate us to use a machine-readable representation to enrich the semantics of relatedness among underground forum users. To formulate the higher-order relationships among entities in AHIN, the concept of meta-path has been proposed [33]: a ***meta-path*** $\mathcal{P}$ is a path defined on the network schema $\mathcal{T}_\mathcal{G} = (\mathcal{A}, \mathcal{R})$, and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_L} A_{L+1}$, which defines a composite relation $R = R_1 \cdot R_2 \cdot \dots \cdot R_L$ between types A_1 and A_{L+1}, where $\cdot$ denotes relation composition operator, and L is the length of $\mathcal{P}$. Given a network schema with different types of entities and relations, we can enumerate a lot of meta-paths. In our application, based on the collected data, we design five meaningful meta-paths (i.e., ***PID1-PID5*** shown in Figure 4) to characterize relatedness over underground forum users. Different meta-paths depict the relatedness between two users at different views. For example, a typical one to formulate the relatedness over underground forum users is ***PID2***: $user \xrightarrow{post} thread \xrightarrow{belongto} section \xrightarrow{belongto^{-1}} thread \xrightarrow{post^{-1}} user$ which means that two users can be connected through the path that their posted threads belong to the same section (e.g., they posted threads introducing how to write a hacking tool in the section of "Hacking Tools and Programs"); while another meta-path ***PID3***: $user \xrightarrow{write} reply \xrightarrow{echo} thread \xrightarrow{echo^{-1}} reply \xrightarrow{write^{-1}} user$ denotes that two users are related if they both reply to a thread (e.g., they both left reviews for a thread advertising the illicit Facebook bots). In our application, meta-path is a straightforward method to connect users via different relationships among different entities in AHIN, and enables us to portray the relatedness over underground forum users in a comprehensive way.

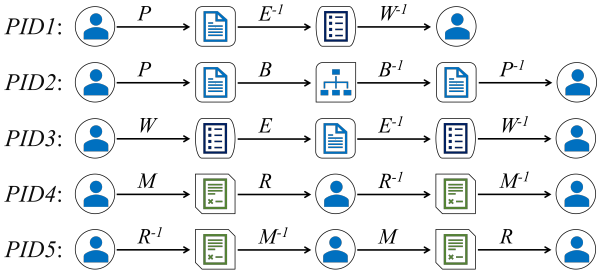


Figure 4: Meta-paths built for key player identification. (The symbols are the abbreviations shown in Figure 3.)

3.3 Player2Vec

In order to efficiently solve the node classification problem (i.e., key player identification) in AHIN, scalable representation learning method for AHIN is in need. To tackle this problem, we first formalize the problem of AHIN representation learning as below.

Definition 2. ***AHIN Representation Learning*** *[7, 11]. Given an AHIN $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, the representation learning task is to learn a function $f: \mathcal{V} \rightarrow \mathbb{R}^D$ that maps each node $v \in \mathcal{V}$ to a vector in a D-dimensional space $\mathbb{R}^D$, $D \ll |\mathcal{V}|$ that are capable of preserving both structural and semantic relations among them.*

Although many network embedding methods [7, 11, 30, 34] have been proposed recently, few of them addressed node embeddings in AHIN. To address this challenge, we propose to map the constructed AHIN to a multi-view network that consists of k single-view attributed graphs encoding the relatedness over users depicted by k designed meta-paths. We first define a multi-view network as: a ***multi-view network*** $\mathcal{G}^\mathcal{K} = (\mathcal{V}, {\mathcal{E}^k}_{k \in K}, \mathcal{A})$ is a network consisting of a set $\mathcal{V}$ of nodes and a set K of views, where $\mathcal{E}^k$ consists of all edges in view $k \in K$. If a multi-view network is weighted, then there exists a weight mapping $w: {\mathcal{E}^k}_{k \in K} \rightarrow \mathbb{R}$ such that $w^k_{vv'} := w(e^k_{vv'})$ is the weight of the edge $e^k_{vv'} \in \mathcal{E}^k$, which joints nodes $v \in \mathcal{V}$ and $v' \in \mathcal{V}$ in view $k \in K$.

Based on the above definition, given an AHIN $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ and K meta-paths, we build a multi-view network with K single-view attributed graph $\mathcal{G}^k = (\mathcal{V}^k, \mathcal{E}^k, \mathcal{A})$ for the k-th view guided by the meta-path $\mathcal{P}_k$ ($k = \{1, ..., K\}$). In our case, in each single-view attributed graph $\mathcal{G}^k$, each node denotes a user and an edge between two users denotes if these two users can be connected under meta-path *PIDi*. Figure. 5 illustrates the mapped multi-view network from the constructed AHIN in our application. These single-view attributed graphs depict different interaction relations among users, which can reflect different-views of user latent representations.

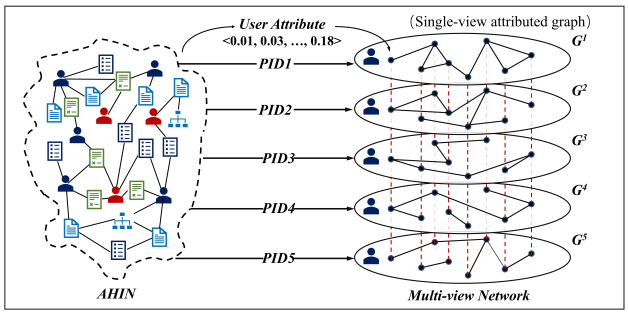


Figure 5: Multi-view network built from AHIN.

To this end, let N be the number of users and $\mathbf{X}$ be a matrix of user feature vectors. The user feature matrix in each single-view attributed graph $\mathcal{G}^k$ is represented as:

$$\mathbf{X} = \mathbf{u}_1 \oplus \mathbf{u}_2 \oplus \ \dots \ \oplus \mathbf{u}_N, \tag{1}$$

where $\oplus$ is the concatenation operator, $\mathbf{u}_i \in \mathbb{R}^f$ is the f-dimensional attributed feature vector for the i-th node (i.e., user) in each single-view attributed graph $\mathcal{G}^k$.

After mapping the constructed AHIN to a multi-view network which consists of K single-view attributed graphs encoding the relatedness over users depicted by K designed meta-paths, we then employ graph convolutional network (GCN) to learn embeddings of each single-view attributed graph and later exploit an attention mechanism to fuse different embeddings learned based on different single-view attributed graphs for final representations.

3.3.1 Single-View Attributed Graph Embedding with GCN. Since GCN have shown its power to learn node representation [3, 5, 15, 21], we here employ GCN for each single-view attributed graph to learn its embeddings. Assume that $\mathbf{f}_i^k \in \mathbb{R}^d$ denotes the embedding of node $i \in \mathcal{V}^k$ in the k-th view graph, where d is the dimension of node embedding. The adjacency matrix $A^k \in \mathbb{R}^{N*N}$ denotes the connection relations among users in the k-th view graph. That is, $A_{ij}^k = 1$ denotes user i and j are connected by the meta path $\mathcal{P}_k$.

Our goal is to learn the node embedding $\mathbf{f}_i^k$ on single-view attributed graph $\mathcal{G}^k = (\mathcal{V}^k, \mathcal{E}^k)$. Following the basic idea of GCN, the convolutional layer is devised as:

$$H^{k,l+1} = \sigma(\widetilde{A^k} H^{k,l} W^{k,l}), \tag{2}$$

where $\widetilde{A^k}$ is a symmetric normalization of A^k with self-loop, i.e. $\widetilde{A^k} = \hat{D^k}^{-\frac{1}{2}} \hat{A^k} \hat{D^k}^{-\frac{1}{2}}$ with $\hat{A^k} = A^k + I_N$. Here I_N is the identity matrix and $\hat{D^k}$ is the diagonal node degree matrix of $\hat{A^k}$. The $H^{k,l}$ and $W^{k,l}$ denote the l-th hidden layer and the layer-specific parameters respectively, and σ denotes an activation function, such as the ReLU$(\cdot) = \max(0, \cdot)$. Along the convolutional network, we obtain the final node embedding as follows:

$$\mathbf{f}^k = \widetilde{A^k}(\text{ReLU}\dots\text{ReLU}(\widetilde{A^k} X W^{k,0})\dots W^{k,L-2}) W^{k,L-1}), \tag{3}$$

where L denotes the number of convolutional layers. The L-layer GCN effectively convolves the L_{th}-order neighborhood of every node. To make the training process stable and reliable, we employ the same initialization process in [12].

Previous studies [23] have shown that the graph convolution is a type of Laplacian smoothing. The Laplacian smoothing computes the new features of a node based on itself and its neighbors [21]. However, if we apply Laplacian smoothing many times in a deep neural network, the feature vector of each node will converge to the similar values with its neighbors'[23]. So we consider a two-layer GCN for key player identification in our model. And thus we denote the user embedding as:

$$\mathbf{f}^k = \text{GCN}(\mathbf{X}, A^k) = \widetilde{A^k}\,\text{ReLU}(\widetilde{A^k}\,\mathbf{X}\,W^{k,0})\,W^{k,1}, \tag{4}$$

where $W^{k,0} \in \mathbb{R}^{F*h}$ and $W^{k,1} \in \mathbb{R}^{h*d}$ denote the weight matrix with h feature maps for input layer and the weight matrix for output layer, respectively.

3.3.2 Multi-View Network Embedding with Attention. We have introduced the learning process of user embeddings in individual attributed graph from a single view in the above section. Then we need to fuse the user embeddings from multi-view network to generate a uniform user embeddings. Intuitively, users are likely to have different preferences over the embeddings generated from different meta-paths. Due to the effectiveness of attention mechanism in various machine learning tasks[6, 17], we design an attention mechanism to learn robust user embeddings through automatically learning the attention weights of different views, instead of simply averaging those user embeddings (i.e., $\mathbf{f}_i^k$ for $k \in K$ in Eq.(4)).

Specifically, we define the attention weight of view k for node i using a softmax unit as follows:

$$\alpha_{i,k} = \frac{\exp({\mathbf{z}^k}^T \cdot \mathbf{f}_i^C)}{\sum_{k' \in K} \exp({\mathbf{z}^{k'}}^T \cdot \mathbf{f}_i^C)}, \tag{5}$$

where $\mathbf{z}^k \in \mathbb{R}^{|K|*d}$ is the attention vector for view k and $\mathbf{f}_i^C$ is the concatenation of node i's embedding with respect to all views. A higher $\alpha_{i,k}$ means that view k is more informative for node i. After obtaining the path attention scores $\alpha_{i,k}$, the final embedding is given as the following weighted sum form:

$$\mathbf{e}_i = \sum_{k \in K} \alpha_{i,k} \cdot \mathbf{f}_i^k, \tag{6}$$

where $\mathbf{f}_i^k$ is the embedding for node i based on view k.

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we fully evaluate the performance of our developed system *iDetective* for key player identification by comparisons with other baselines.

4.1 Data Collection and Annotation

To fully evaluate the system of *iDetective*, we develop a set of crawling tools to collect the users' profiles, posted threads, replies and comments as well as the sections in underground forums through September 2017 to September 2018. We briefly introduce these two forums and summarize the data we collected.

- **Hack Forums** [14] is currently one of the largest online hacking forums. The threads on the forum cover a wide range of cybercrime related topics, such as web hacking, crypters, keyloggers,

hacking tools and monetizing. By the date, we have collected 238,212 threads posted by 74,909 users in Hack Forums.
- **Nulled** [28] is a large ongoing cracking community primarily focused on leaks and tools for data breach. By the date, we have collected 356,605 threads posted by 118,738 users in Nulled.

In order to obtain the ground truth, we select a portion of users from Hack Forums (i.e., users posted threads in the sub-forum of "Hack" and the section of "Premium Sellers Section") and Nulled (i.e., users posted threads in the sub-forum of "Cracking" and the section of "Premium Sellers Section") for labeling. Then, we (i.e., six annotators in three groups) have spent three months to manually label whether these users (i.e., 5,500 users in Hack Forums and 5,500 users in Nulled) are key players or not following the criteria of: (1) the key players should be active in the forum (post rate per day is greater than 1 and time spent online per day is greater than 0.5 hours); (2) they should be capable of exploiting and disseminating vulnerabilities and develop malicious tools (number of posted products or services containing vulnerabilities or malicious tools is great than 3), and (3) they are also active in selling and monetizing illicit products or services (reputation is greater than 500, number of total reviews for sold products or service is great than 200, each sold product or service is great than 2 stars). The mutual agreement is above 95%, and only the ones with agreements are retained. Based on these criteria, (i) for Hack Forums, 933 users are labeled as key players and 4,347 are non-key players; and (ii) for Nulled, 861 users are labeled as key players and 4,371 are non-key players. We use accuracy (ACC) and F1 measure (F1) as the evaluation metric.

4.2 Evaluation of Multi-view Network

In this set of experiments, based on the dataset described in Section 4.1, we evaluate the performance of multi-view network for key player identification by comparisons with single-view attributed graphs guided by different meta-paths. For each single-view graph, we apply GCN to learn the low-dimensional user embeddings. To compare with multiple-view network, we merge the links of different single-views into a unified view and then learn user embeddings via GCN. Finally, we feed user embeddings to Support Vector Machine (SVM) to build the classification model for key player identification. In the experiments, we randomly select 90% of the data for training, while the remaining 10% is used for testing. For SVM, we use LibSVM and the penalty is empirically set to be 10 while other parameters are set by default. The experimental results are shown in Table 1, from which we can see that different single-view graphs show different performances, since each of them represents a specific semantic in key player identification task. Further, multi-view method successfully outperforms each single-view on both Hack Forums and Nulled in terms of ACC and F1, which reveals that the incorporation of relatedness depicted by different single views can provide much higher-level semantics and thus improve the key player identification performance.

Table 1: Identification Results of different views.

ID	Related Meta-path	Hack Forums		Nulled	
		ACC	F1	ACC	F1
G^1	PID1	0.803	0.590	0.798	0.565
G^2	PID2	0.787	0.566	0.771	0.526
G^3	PID3	0.806	0.595	0.812	0.587
G^4	PID4	0.792	0.574	0.791	0.555
G^5	PID5	0.811	0.603	0.819	0.598
Multi-view (G^1- G^5)		0.857	0.679	0.861	0.671

4.3 Network Embedding Comparisons

In this set of experiments, we compare *Player2Vec* with several state-of-the-art network embedding methods. To further examine the attention mechanism, we also prepare two variants of *Player2Vec*. The baselines are given below:

- **DeepWalk** [30] learns node vectors by capturing node pairs within w-hop neighborhood via uniform random walks in the network. In our experiment settings, we ignore the heterogeneous and and attribute information, and directly feed AHIN for embedding.
- **LINE** [34] learns embeddings by preserving first-order and second-order proximities between nodes. Same as DeepWalk, we directly feed AHIN for embedding.
- **metapath2vec** [7] is a HIN embedding method which embeds the semantic information of a single meta-path. Here, similar to our proposed method, we also fuse different embeddings guided by different meta-paths via the attention mechanism.
- **HIN2Vec** [11] learns node embeddings to capture rich relation semantics in HIN via a neural network model.
- **Player2Vec**$-cat$ is a variant of our method which concatenates all learned representations based on each single-view attributed graph to form a new embedding.
- **Player2Vec**$-noAtt$ is another variant of our method which allocates equal weight to each single-view graph without learning the weights of views by the attention mechanism.

For the former four embedding methods, we use the same parameters: vector dimension $d = 100$ (LINE: 50 for each order (1st- and 2nd-order)), walks per node $r = 10$, walk length $l = 80$ and window size $w = 10$. For **Player2Vec** and its variants, we set the same parameters as [21]. To facilitate the comparisons, we use the experimental procedure as in [7, 30, 34]: we randomly select a portion of data (ranging from 10% to 90%) for training and the remaining ones for testing. SVM is used as the classification model for all the methods. Table 2 illustrates the results of different network embedding methods in key player identification. From Table 2, we can see that *Player2Vec* consistently and significantly outperforms all baselines for key player identification. That is to say, *Player2Vec* learns significantly better user representations than current state-of-the-art methods. The success of *Player2Vec* lies in: (1) the proper consideration and accommodation of the heterogeneous property of AHIN, (2) the advantage of GCN which can explore network characteristics at a spectrum of frequency bands for learning higher-level semantics, and (3) the attention mechanism for aggregating node embeddings learned based on different single-view graphs.

To demonstrate the effectiveness of the attention mechanism in our method, we further analyze the correlation between the learned weights and the actual performance of each view. As shown in Figure 6, we can see that the performances of different views are

Table 2: Comparisons with other network embedding methods in key player identification.

Dataset	Indices	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%
Hack Forums	ACC	DeepWalk	0.740	0.750	0.760	0.781	0.791	0.800	0.812	0.817	0.828
		LINE	0.748	0.759	0.768	0.788	0.799	0.812	0.821	0.831	0.843
		metapath2vec	0.781	0.787	0.800	0.818	0.827	0.842	0.850	0.861	0.872
		Hin2Vec	0.779	0.789	0.797	0.821	0.833	0.842	0.849	0.860	0.870
		Player2Vec–*cat*	0.792	0.807	0.814	0.838	0.847	0.852	0.868	0.873	0.884
		Player2Vec–*noAtt*	0.800	0.808	0.823	0.837	0.849	0.859	0.872	0.881	0.893
		Player2Vec	0.817	0.832	0.842	0.861	0.871	0.880	0.888	0.897	0.909
	F1	DeepWalk	0.475	0.488	0.502	0.531	0.546	0.561	0.578	0.590	0.607
		LINE	0.488	0.501	0.515	0.544	0.560	0.578	0.593	0.610	0.629
		metapath2vec	0.531	0.543	0.561	0.591	0.606	0.629	0.644	0.664	0.684
		Hin2Vec	0.530	0.545	0.558	0.594	0.612	0.629	0.643	0.662	0.682
		Player2Vec–*cat*	0.551	0.570	0.583	0.621	0.637	0.649	0.675	0.688	0.710
		Player2Vec–*noAtt*	0.561	0.574	0.595	0.624	0.643	0.661	0.684	0.702	0.725
		Player2Vec	0.590	0.611	0.629	0.664	0.682	0.701	0.718	0.738	0.762
Nulled	ACC	DeepWalk	0.744	0.749	0.758	0.782	0.793	0.797	0.807	0.824	0.827
		LINE	0.752	0.761	0.769	0.792	0.800	0.811	0.821	0.827	0.839
		metapath2vec	0.782	0.789	0.802	0.818	0.828	0.843	0.847	0.861	0.870
		Hin2Vec	0.779	0.789	0.801	0.817	0.830	0.844	0.853	0.861	0.871
		Player2Vec–*cat*	0.797	0.807	0.813	0.836	0.845	0.859	0.864	0.873	0.883
		Player2Vec–*noAtt*	0.800	0.812	0.820	0.838	0.849	0.860	0.873	0.883	0.889
		Player2Vec	0.820	0.830	0.838	0.861	0.870	0.884	0.891	0.898	0.913
	F1	DeepWalk	0.494	0.506	0.519	0.548	0.565	0.576	0.591	0.613	0.623
		LINE	0.507	0.520	0.533	0.564	0.578	0.594	0.610	0.624	0.642
		metapath2vec	0.549	0.562	0.580	0.608	0.624	0.646	0.657	0.679	0.697
		Hin2Vec	0.547	0.562	0.579	0.607	0.626	0.647	0.664	0.679	0.698
		Player2Vec–*cat*	0.572	0.587	0.599	0.635	0.651	0.673	0.686	0.704	0.723
		Player2Vec–*noAtt*	0.578	0.595	0.609	0.641	0.659	0.678	0.700	0.720	0.734
		Player2Vec	0.610	0.626	0.641	0.679	0.696	0.720	0.737	0.752	0.782

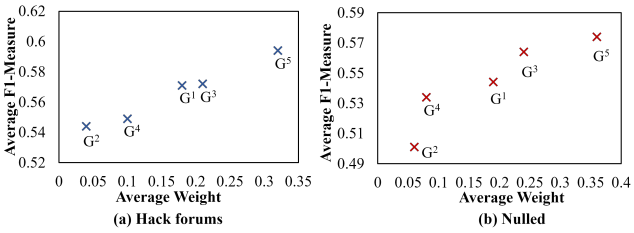


Figure 6: Weight and F1 correlation.

positively correlated with the average weights learned by attention mechanism. In other words, the weights over different views learned by our attention mechanism are very intuitive, which enable different nodes to focus on those most informative views.

4.4 Comparisons with Alternative Approaches

In this set of experiments, we compare *iDetective* with other alternative machine learning methods for key player identification. Here, we construct three types of features:

- **BoW *(f-1)***: Each user is represented by a bag-of-words vector based on the posted threads, replies and comments.
- **Relations *(f-2)***: Four relation features are extracted to represent each underground forum user, i.e., whether two users i) are in the same section, ii) reply to the same thread, iii) comment on the same user; and iv) are commented by the same user.
- **BoW+Relations *(f-3)***: This augments bag-of-words with relations as flat features.

Based on these features, we consider two typical classification models, i.e., Naive Bayes (NB) and SVM. We also compare *iDetective* with other proposed methods for influential users identification (i.e., RRI [2], DA [1], and KADetector [40]): (i) In RRI, the radicalness of each user and association among users are measured and then customized PageRank algorithm is exploited to rank the influential users. (ii) In DA, each user is represented by content-based features (i.e., lexicon matching) and structural features via interaction coherence analysis, based on which X-means algorithm is performed to identify activists. (iii) In KADetector, a HIN embedding model (without considering attribute information attached to the nodes) is proposed to learn the desirable node representations for key player identification. In this set of experiments, we perform ten-fold cross validations. The experimental results are illustrated in Table 3. From the results, we observe that feature engineering (***f-3***) 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 is simply flat features, i.e., concatenation

of different features altogether, which is less informative than the features extracted from HIN. In contrast, *iDetective* and KADetector add the knowledge represented as HIN significantly works better than other baseline methods. Further, our developed system *iDetective* successfully outperforms KADetector. The reason behind this is that *iDetective* is able to fully utilize user attributed features to build the higher-level semantic and structural connection between users, and thus achieves better performance in key player identification.

Table 3: Comparisons with other methods.

Method	Settings	Hack Forums		Nulled	
		ACC	F1	ACC	F1
NB	***f-1***	0.689	0.439	0.690	0.423
	f-2	0.675	0.423	0.673	0.404
	f-3	0.728	0.486	0.728	0.468
SVM	***f-1***	0.724	0.481	0.723	0.462
	f-2	0.711	0.465	0.710	0.446
	f-3	0.769	0.541	0.773	0.528
RRI[2]	-	0.722	0.479	0.735	0.477
DA[1]	-	0.772	0.545	0.776	0.533
KADetector[40]	-	0.884	0.729	0.887	0.721
iDetective	-	**0.907**	**0.775**	**0.912**	**0.773**

4.5 Evaluations of Parameter Sensitivity, Scalability and Stability

In this set of experiments, we systematically evaluate the parameter sensitivity, scalability and stability of *iDetective*. The experimental studies are conducted under the environment of ubuntu 16.04 operating system, plus two Intel Xeon E5-2620 v4 CPU, 4-way SLI GeForce GTX 1080 Ti Graphics Cards and 80 GB of RAM.

We first conduct the **sensitivity** analysis of how different choices of dimension d will affect the performance of *iDetective* in key player identification. As shown in Figure 7.(a), the performance tends to be stable once d reaches around 150. Overall, *iDetective* is not strictly sensitive to this parameter and is able to reach high performance under a cost-effective parameter choice. We then further evaluate the **scalability** and **stability** of *iDetective*. For scalability, we evaluate the running time of *iDetective* with different sizes of the dataset. Figure 7.(b) shows that the running time is quadratic to the number of samples. When dealing with more data, approximation or parallel algorithms can be developed. Figure 7.(c) shows the overall receiver operating characteristic (ROC) curves of *iDetective* based on the ten-fold cross validations; it achieves an impressive performance both in Hack Forums and Nulled. From the results and analysis above, *iDetective* is efficient and scalable for large-scale AHIN mining with large numbers of nodes.

4.6 Case Studies

To better understand and gain deeper insights into the ecosystem of underground forums, in this section, using our developed system *iDetective*, we further identify 105 key players in Hack Forums. We

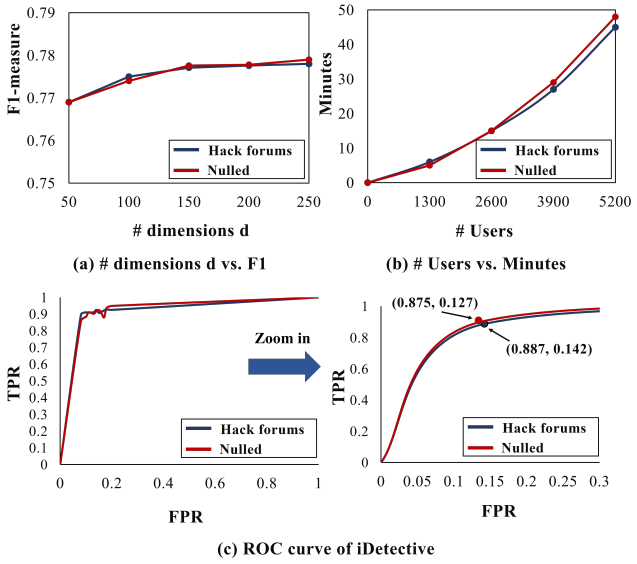


Figure 7: Parameter sensitivity, scalability and stability

then retrieve all the posts generated by these users and further analyze their online activities. We have several interesting findings: (1) most of the identified key players barely focused on a particular kind of product or service; (2) as shown in Table 4, the most prevalent products or services were social media account hacking products or services; (3) Figure 8 shows the trend of illicit activities in Hack Forums in recent years (i.e., the supply of products or services of *social media account hacking* and *botnet & web attack* has had exponential growth; while the supply of *software cracking* related products or services has slowly decreased). These findings also reveal that, as the popularity of social media platforms grows, there is an imminent need to improve the security of online social media accounts.

Table 4: Different kinds of products or services advertised by identified key players in Hack Forums.

Type	#Products	# Key players
Social media account hacking	172	34
Botnet and web attack	146	25
Software cracking	99	19
Web vulnerability and exploit	96	19
Others	79	24

To facilitate the understanding of different activities performed by key players and non-key players in Hack Forums, we compare the 105 identified key players with the 108 identified non-key players (i.e., vendors). Figure 9.(a) shows the price distributions of the products or services advertised by key players and non-key players. From Figure 9.(a), we can observe that (1) the price of the advertised products or services is in the range of [$1, $599], majority of which are lower than $50; (2) compared with non-key players, key players are more experienced in pricing their products or services

(i.e., 10.5% of products or services offered by non-key players were priced lower than $5 or greater than $100; while key players priced their products or services in a more reasonable range), the reason behind which may be that key players could better analyze the market demands and supplies. Figure 9.(b) shows the distributions of view numbers of products or services advertised by key players and non-key players. From Figure 9.(b), we can see that (1) the products or services offered by key players had more views than those offered by non-key players; (2) 87.3% of the products or services offered by key players were viewed more than 100 times, while only 38.7% offered by non-key players were viewed more than 100 times. The analysis also indicates that the products or services advertised by key players are more popular than the ones offered by non-key players, as they had better knowledge and expertise to offer quality products or services.

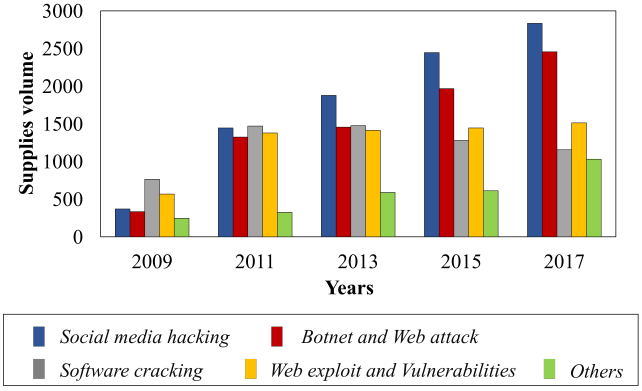


Figure 8: The trend of different kinds of products or services advertised by identified key players in Hack Forums.

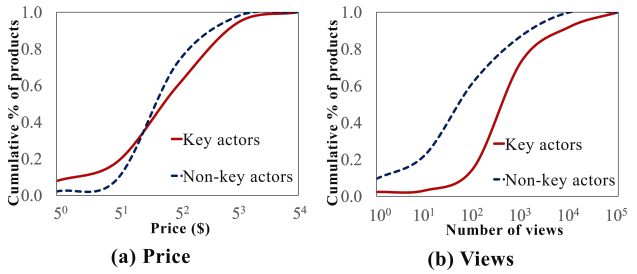


Figure 9: Key players vs. non-key players on price and views.

The above studies reveal that using the automatic methods and 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. The studies based on the identified key players in Hack Forums using our developed system *iDetective* also demonstrate that knowledge gained from underground forum data mining could facilitate the insight into cybercrime ecosystem, which can help devise effective interventions.

5 RELATED WORK

To combat the cybercrimes which have become increasingly dependent on the online underground forums, there have been many research efforts on underground forum analysis [19, 29, 39], which can be categorized into three areas: (1) the first type focuses on identifying the threats found in the content and other content-related features; (2) the second type mainly works on understanding the cybercriminal community structure and social relationships; (3) the third type focuses on identifying the most influential cybercriminal community members. Our work is one of the third type. There have been some existing works on key member identification, for examples: Anwar et al. [2] embedded users radicalness and association in a customized PageRank algorithm to rank the influential users in the web forums; Abbasi et al. [1] characterized users with content features and structural features and perform the X-means clustering algorithm to identify expert hackers in the hacker forums; Yang et al. [37] incorporated the message similarity and response immediacy features with link analysis to determine the impact and the neighborhood of the influential users; Tang et al. [35] used Bipartite Graph analysis and developed a user interest and topic detection model to predict user participation in the Dark Web; other advanced techniques including deep learning [24] are developed to profile sellers from their advertisements. Different from the existing works, in this paper, we consider various kinds of relationships and propose to utilize a structured AHIN to represent underground forum users for key player identification.

HIN is proposed to model different types of entities and relations and has been applied to various applications, such as scientific publication network analysis [33], document analysis based on knowledge graph [9, 10, 36] and malware detection [8, 16, 38]. Several measures (e.g., meta-path [33] and meta-structure [18]) have already been proposed for relevance computation over HIN entities. In order to reduce the high computation and space cost in network mining, many efficient network embedding methods have been proposed to address representation learning for homogeneous network, such as DeepWalk [30], node2vec [13], and LINE [34]. Unfortunately, due to the heterogeneous properties of HIN, it's difficult to directly apply them for HIN representation learning. To tackle this challenge, several HIN embedding approaches such as metapath2vec [7] have been proposed. However, existing HIN embedding models fail to incorporate the attribute information attached to the nodes in AHIN. To solve this problem, in this paper, we propose a novel model *Player2Vec* to learning node representations in AHIN for key player identification.

6 CONCLUSION

To combat the illicit activities in underground forums, in this paper, we design and develop an intelligent system named *iDetective* to automate the identification of key players in underground forums. In *iDetective*, we first introduce an AHIN for user representation and use a meta-path based approach to incorporate higher-level semantics to build up relatedness over users in underground forums; then we propose *Player2Vec* to efficiently learn node (i.e., user) representations in AHIN for key player identification. In *Player2Vec*, we first map the constructed AHIN to a multi-view network which consists of K single-view attributed graphs encoding the relatedness over users depicted by K designed meta-paths; then we employ GCN to learn embeddings of each single-view attributed graph; later, an attention mechanism is designed to fuse

different embeddings learned based on different single-view attributed graphs for final representations. Based on the large-scale data collected from different underground forums (i.e., Hack Forums, Nulled) and the obtained ground-truth, the promising experimental results demonstrate that *iDetective* which integrates our proposed method outperforms other baselines in key player identification in online underground forums.

ACKNOWLEDGMENTS

Y. Zhang, Y. Fan, Y. Ye's work is partially supported by the NSF under grants CNS-1618629, CNS-1814825, CNS-1845138, OAC-1839909, and III-1908215, the NIJ 2018-75-CX-0032, the WV HEPC Grant (HEPC.dsr.18.5), and the WVU RSA grant (R-844); L. Zhao's work is partially supported by the NSF under grants 1755850, 1841520 and 1907805, Jeffress Trust Award, and NVEDIA GPU Grant.

REFERENCES

[1] Ahmed Abbasi, Weifeng Li, Victor Benjamin, Shiyu Hu, and Hsinchun Chen. 2014. Descriptive analytics: Examining expert hackers in web forums. In *IEEE Joint Intelligence and Security Informatics Conference*. IEEE, 56–63.

[2] Tarique Anwar and Muhammad Abulaish. 2015. Ranking radically influential web forum users. *IEEE Transactions on Information Forensics and Security* 10, 6 (2015), 1289–1298.

[3] James Atwood and Don Towsley. 2016. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1993–2001.

[4] BlackHatworld. 2018. . https://www.blackhatworld.com/.

[5] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203* (2013).

[6] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan S Kankanhalli. 2018. Aˆ 3NCF: An Adaptive Aspect Attention Model for Rating Prediction.. In *International Joint Conferences on Artificial Intelligence*. 3748–3754.

[7] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 135–144.

[8] Yujie Fan, Shifu Hou, Yiming Zhang, Yanfang Ye, and Melih Abdulhayoglu. 2018. Gotcha-sly malware! Scorpion: a metagraph2vec based malware detection system. In *KDD*.

[9] Yujie Fan, Yiming Zhang, Shifu Hou, Lingwei Chen, Yanfang Ye, Chuan Shi, Liang Zhao, and Shouhuai Xu. 2019. iDev: Enhancing Social Coding Security by Cross-platform User Identification Between GitHub and Stack Overflow. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. IJCAI, 2272–2278. https://doi.org/10.24963/ijcai.2019/315

[10] Yujie Fan, Yiming Zhang, Yanfang Ye, and Xin Li. 2018. Automatic Opioid User Detection from Twitter: Transductive Ensemble Built on Different Meta-graph Based Similarities over Heterogeneous Information Network.. In *International Joint Conference on Artificial Intelligence*. 3357–3363.

[11] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1797–1806.

[12] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth Tnternational Conference on Artificial Intelligence and Statistics*. 249–256.

[13] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 855–864.

[14] HackForums. 2018. . https://hackforums.net/.

[15] Mikael Henaff, Joan Bruna, and Yann LeCun. 2015. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163* (2015).

[16] Shifu Hou, Yanfang Ye, Yangqiu Song, and Melih Abdulhayoglu. 2017. Hindroid: An intelligent android malware detection system based on structured heterogeneous information network. In *KDD*.

[17] Binbin Hu, Chuan Shi, Wayne Xin Zhao, and Philip S Yu. 2018. Leveraging Meta-path based Context for Top-N Recommendation with A Neural Co-Attention Model. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1531–1540.

[18] Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, and Xiang Li. 2016. Meta structure: Computing relevance in large heterogeneous information networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1595–1604.

[19] Aleksandar Hudic, Katharina Krombholz, Thomas Otterbein, Christian Platzer, and Edgar Weippl. 2014. Automated Analysis of Underground Marketplaces. In *IFIP International Conference on Digital Forensics*. Springer, 31–42.

[20] Mohammad Karami and Damon McCoy. 2013. Understanding the emerging threat of ddos-as-a-service. In *The 6th USENIX Workshop on Large-Scale Exploits and Emergent Threats*.

[21] Thomas N Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907* (2016).

[22] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. 1188–1196.

[23] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper Insights Into Graph Convolutional Networks for Semi-Supervised Learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[24] Weifeng Li and Hsinchun Chen. 2014. Identifying top sellers in underground economy using deep learning-based sentiment analysis. In *IEEE Joint Intelligence and Security Informatics Conference*. IEEE, 64–67.

[25] Xiang Li, Yao Wu, Martin Ester, Ben Kao, Xin Wang, and Yudian Zheng. 2017. Semi-supervised clustering in attributed heterogeneous information networks. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1621–1629.

[26] Giancarlo De Maio, Alexandros Kapravelos, Yan Shoshitaishvili, Christopher Kruegel, and Giovanni Vigna. 2014. Pexy: The other side of exploit kits. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. 132–151.

[27] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 55–60.

[28] Nulled. 2018. . https://www.nulled.to.

[29] Sergio Pastrana, Daniel R Thomas, Alice Hutchings, and Richard Clayton. 2018. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale. In *Proceedings of the 27th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1845–1854.

[30] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 701–710.

[31] Statista. 2018. *Global retail e-commerce sales 2014-2021*. https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/.

[32] Yizhou Sun and Jiawei Han. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* 3, 2 (2012), 1–159.

[33] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment* 4, 11 (2011), 992–1003.

[34] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[35] Xuning Tang, Christopher C Yang, and Mi Zhang. 2012. Who will be participating next? predicting the participation of Dark Web community. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*. ACM, 1–7.

[36] Chenguang Wang, Yangqiu Song, Haoran Li, and Jiawei Zhang. 2016. Text Classification with Heterogeneous Information Network Kernels. In *Thirtieth AAAI Conference on Artificial Intelligence*. 2130–2136.

[37] Christopher C Yang, Xuning Tang, and Bhavani M Thuraisingham. 2010. An analysis of user influence ranking algorithms on dark web forums. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*. ACM, 10.

[38] Yanfang Ye, Shifu Hou, Lingwei Chen, Jingwei Lei, Wenqiang Wan, Jiabin Wang, Qi Xiong, and Fudong Shao. 2019. Out-of-sample Node Representation Learning for Heterogeneous Graph in Real-time Android Malware Detection. In *IJCAI*.

[39] Yiming Zhang, Yujie Fan, Wei Song, Shifu Hou, Yanfang Ye, Xin Li, Liang Zhao, Chuan Shi, Jiabin Wang, and Qi Xiong. 2019. Your Style Your Identity: Leveraging Writing and Photography Styles for Drug Trafficker Identification in Darknet Markets over Attributed Heterogeneous Information Network. In *The World Wide Web Conference*. ACM, 3448–3454.

[40] Yiming Zhang, Yujie Fan, Yanfang Ye, Liang Zhao, Jiabin Wang, Qi Xiong, and Fudong Shao. 2018. KADetector: Automatic Identification of Key Actors in Online Hack Forums Based on Structured Heterogeneous Information Network. In *IEEE International Conference on Big Knowledge*. IEEE, 154–161.