Your Style Your Identity: Leveraging Writing and Photography Styles for Drug Trafficker Identification in Darknet Markets over Attributed Heterogeneous Information Network

Yiming Zhang, Yujie Fan
Department of CSEE
West Virginia University, WV, USA
{ymzhang,yf0004}@mix.wvu.edu

Wei Song, Shifu Hou
Department of CSEE
West Virginia University, WV, USA
{ws0016,shhou}@mix.wvu.edu

Yanfang Ye*, Xin Li
Department of CSEE
West Virginia University, WV, USA
{yanfang.ye,xin.li}@mail.wvu.edu

Liang Zhao
Department of IST
George Mason University, VA, USA
lzhao9@gmu.edu

Chuan Shi
School of Computer Science
B UPC, Beijing, China
shichuan@bupt.edu.cn

Jiabin Wang, Qi Xiong
Tencent Security Lab
Tencent, Guangdong, China
luciferwang@tencent.com

ABSTRACT
Due to its anonymity, there has been a dramatic growth of underground drug markets hosted in the darknet (e.g., Dream Market and Valhalla). To combat drug trafficking (a.k.a. illicit drug trading) in the cyberspace, there is an urgent need for automatic analysis of participants in darknet markets. However, one of the key challenges is that drug traffickers (i.e., vendors) may maintain multiple accounts across different markets or within the same market. To address this issue, in this paper, we propose and develop an intelligent system named uStyle-uID leveraging both writing and photography styles for drug trafficker identification at the first attempt. At the core of uStyle-uID is an attributed heterogeneous information network (AHIN) which elegantly integrates both writing and photography styles along with the text and photo contents, as well as other supporting attributes (i.e., trafficker and drug information) and various kinds of relations. Built on the constructed AHIN, we propose a new network embedding model Vendor2Vec to learn the low-dimensional representations for the nodes in AHIN, which leverages complementary attribute information attached in the nodes to guide the meta-path based random walk for path instances sampling. After that, we devise a learning model named vIdentifier to classify if a given pair of traffickers are the same individual. Comprehensive experiments on the data collections from four different darknet markets are conducted to validate the effectiveness of uStyle-uID which integrates our proposed method in drug trafficker identification by comparisons with alternative approaches.

KEYWORDS
Darknet Market; Drug Trafficker Identification; Attributed Heterogeneous Information Network (AHIN); Network Embedding.

ACM Reference Format:

1 INTRODUCTION
The market of illicit drugs (e.g., cannabis, cocaine, heroin) is considerably lucrative - i.e., the estimated yearly revenue for the global market reached about $426-$652 billion in 2017 [23]. Driven by such remarkable profits, the crime of drug trafficking (a.k.a. illicit drug trading) has never stopped but co-evolved along with the advance of modern technologies [4, 16, 17, 28, 32, 38, 39]. Darknet, as a hidden part of the Internet, employs advanced encryption techniques to protect the anonymity of its users. The markets hosted in the darknet are built on The Onion Router (TOR) service to hide the IP address, the escrow system, the encrypted communication tools like Pretty Good Privacy (PGP), and the virtually untraceable cryptocurrency (e.g., bitcoin) to facilitate anonymous transactions among participants. Figure 1.(a) illustrates a typical transaction in darknet markets. Due to its anonymity, there has been a dramatic growth of underground drug markets hosted in the darknet (e.g., Silk Road 3 [33], Dream Market [27], Valhalla [37], known as “eBay of drugs” or “Amazon of drugs”). Illegal trading of drugs in these markets has turned into a serious global concern because of its severe consequences on society (e.g., violent crimes) and public health at regional, national and international levels [8].

To combat drug trafficking in the cyberspace, there is an urgent need for analysis of participants in darknet markets, as it could provide valuable insight to the investigation of drug trafficking ecosystem and prediction of future incidents while building proactive defenses [6]. However, one of the key challenges is that drug traffickers may maintain multiple accounts across different markets or in the same market for the reasons [2, 4, 34] such as ripper (i.e.,
Figure 1: Illustration of drug trafficking in Darknet market.

In this paper, we propose to leverage both writing and photography styles to develop an intelligent system (named uStyle-uID) to automatically link multiple accounts of the same individuals for drug trafficker identification in darknet markets. In uStyle-uID, given a pair of vendors (denoted by their usernames in the related markets), to determine whether they are the same individual, we not only analyze their posted contents (i.e., including their posted texts and photos), but also consider their writing styles and photography styles as well as other supporting attributes (i.e., vendor and drug information) and various kinds of relations. To depict vendors, drugs, texts, photos and their associated attributes as well as the rich relations among them, we present an attributed heterogeneous information network (AHIN) for modeling. To tackle the challenge of high computation cost and memory constraint of measuring the relatedness over vendors in the constructed AHIN, we propose a new network embedding model named Vendor2Vec to learn low-dimensional attribute-aware embeddings for the nodes in AHIN. The proposed Vendor2Vec model leverages complementary attribute information of each node to guide the meta-path based random walk for path-instance sampling; then a skip-gram model [30] is utilized to learn effective node representations for AHIN. Finally, based on the learned latent representations of the nodes (i.e., vendors) in AHIN, we devise a learning model named vIdentifier to classify whether a given pair of vendors are the same individual.

2 PROPOSED METHOD

The overview of our developed system uStyle-uID for drug trafficker identification in darknet markets is shown in Figure 2. In this section, we will introduce the detailed approaches which are integrated in uStyle-uID for drug trafficker identification.
2.1 Feature Extraction

We propose to characterize vendors in darknet markets in a comprehensive view by extracting various features. 

(1) Posted text and writing style extraction. To fingerprint a vendor based on his/her posted texts, we consider both his/her posted text content and writing style. For text content, we apply doc2vec [25] to convert each text of variant size into a fixed length feature vector (empirically, we set the dimension to 100). For writing style [2, 20], we propose to extract multi-scale stylometry features at three different levels as follows: 1) Lexical features can be further divided into character-based and word-based groups to capture stylistic traits. At this level, we extract i) number of characters, ii) number of digits/white spaces/special characters, iii) number of words, iv) average word length, and v) vocabulary richness [36]. 2) Syntactic features capture the writing style from the sentence structure. In this category, we adopt i) frequency of punctuation, ii) frequency of function word, iii) number of sentences beginning with a capital letter, and iv) frequency of parts-of-speech n-grams (we set \(n = 3\) in our case). 3) Structural features represent the way an author organizes the layout of his/her posted text. We consider i) total number of paragraphs, ii) indentation of paragraph, iii) whether there’s separator between paragraphs, and iv) number of words/sentences/characters per paragraph. For each posted text, we then concatenate its converted feature vector representing the posted text content and the feature vector describing its writing style as an attribute associated with this posted text. 

(2) Posted photo and photography style extraction. To represent the content of a posted photo, we propose to utilize image2vec [15] to convert it into a fixed length feature vector (empirically, we set the dimension to 100). Since drug traffickers in darknet markets have to prove the possession of illegal drugs by posting their own photos, their distinct photography style might be revealed by the posted photos. We propose to capture the photography style by extracting its low-level and high-level features. 1) Low-level features refer to the information that can be directly obtained from a photo’s exchangeable image file format (EXIF) data, which include i) camera make and model, ii) camera angle, iii) exposure time, iv) focal length, and v) image size. 2) High-level features are extracted from the photo’s original content. We first convert the photo into its HSV (hue, saturation, value) representation and then extract the following five types of high-level features: i) colorfulness, ii) exposure of light, iii) saturation, iv) hue count, and v) contrast. In our current implementation, colorfulness, exposure of light and saturation are calculated using the method in [7]; while hue count and contrast are measured by [24]. For each posted photo by a vendor, we then concatenate its converted feature vector representing the posted photo content and the feature vector describing its photography style as an attribute associated with this posted photo.

(3) Attributed features of vendors and drugs. Besides the above extracted features, vendors’ basic information and drugs they sell also play an important role in resolving their identities. Therefore, we further extract three kinds of features to depict each vendor: username, PGP key and contact information. Note that, for username, we first apply standard string matching techniques to measure the similarity of two usernames, if their similarity is greater than a user-specific threshold, we regard these two usernames as the same (e.g., “MF***Jones” and “MF***J0nes”). For each drug, we further extract its category, escrow information and shipping information (e.g., from where and to where). Then, we apply one-hot encoding [41] to convert the extracted features to a binary feature vector to be an attribute associated with each vendor/drug.

2.2 AHIN Construction

Though heterogeneous information network (HIN) [35] has shown the success of modeling different types of entities and relations, it has limited capability of modeling additional attributes attached to entities. Thus, to depict vendors, drugs, texts, photos and their associated attributes as well as the rich relationships among them, we propose to use attributed HIN (AHIN) for representation.

Definition 2.1. Attributed heterogeneous information network (AHIN) [26]. Let \(\mathcal{T} = \{T_1, ..., T_m\}\) be a set of \(m\) entity types. For each entity type \(T_i\), let \(X_i\) 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 \(f_j = (f_{j1}, f_{j2}, ..., f_{j|A_i|})\). An AHIN is defined by a graph \(\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})\) with an entity type...
We formalize the AHIN representation learning problem as below.

To reduce the computation and space cost in network mining, scalable representation learning method \[ \text{Vendor2Vec} \] which consists of attribute-aware meta-path random walk and skip-gram model. Given an AHIN \( G = (V, E, A) \) with schema \( TG = (T, R) \), and a meta-path scheme \( \mathcal{P} \) in the basic form: \( T_1 \rightarrow \cdots \rightarrow T_i \rightarrow \cdots \rightarrow T_l \rightarrow T_{l+1} \), we use attribute-aware meta-path to guide a random walker in AHIN, the transition probability at step \( i \) is calculated as:

\[
p(u^{i+1} | v^i, \mathcal{P}) = \begin{cases} 
\frac{\text{sim}(f(v^i), f(u^{i+1}))}{\sum_{u'' \in N_{T_{i+1}}(v^i)} \text{sim}(f(v^i), f(u''))} & \text{if } u^{i+1} \in E, \phi(u^{i+1}) = \phi(u') = T_{l+1} \\
\frac{|N_{T_{i+1}}(v^i)|}{(v^i, u^{i+1}) \in E, \phi(u^{i+1}) = T_{l+1}, v' = 0} & \text{otherwise,}
\end{cases}
\]

where \( v' \) denotes the latest entity the walker visited is with the same type of \( u^{i+1} \), \( \text{sim}(f(v^i), f(u^{i+1})) \) is the similarity between two entities’ attribute vectors (e.g., it can be calculated by using cosine similarity measure), \( \phi \) is the node type mapping function, \( N_{T_{i+1}}(v^i) \) denote \( T_{l+1} \) type of neighborhood of node \( v^i \), \( u^{i+1} \) denotes a node in \( N_{T_{i+1}}(v^i) \). Since we have three different meta-paths, we simply combine the path instances sampled via each meta-path, and feed them into the skip-gram model \[29\] to learn the node embeddings.

Figure 3: Network schema and meta-paths.

In our application, we have four entity types and five types of relations among them; meantime, each entity is also attached with an extracted feature vector representing its associated attributes. Based on the definitions above, the network schema for AHIN in our case is shown in Figure 3.(a) (to facilitate the illustration, attribute information is shown in its original form). Then, we adopt the concept of meta-path \[35\] to formulate higher-order relationships among entities in AHIN. In our application, we focus on three most meaningful meta-paths (i.e., \( \text{PID1-3} \) as shown in Figure 3.(b)) to jointly characterize the relatedness between two vendors from different views: (1) \( \text{PID1} \) means that two vendors can be connected through the path that they both sell the same kind of drug (e.g., heroin); (2) \( \text{PID2} \) denotes that two vendors can be linked if their posted photos describe the same kind of drug (e.g., as shown in Figure 1(c)), the vendor with username of "sp3***" in SilkRoad2 and the vendor with username of "sho***“ in Evolution can be linked via this meta-path; (3) \( \text{PID3} \) indicates that two vendors can be connected if their written texts describe the same kind of drug.

2.3 Vendor2Vec

To reduce the computation and space cost in network mining, scalable representation learning method \[10, 14\] for AHIN is in need. We formalize the AHIN representation learning problem as below.

**Definition 2.2. AHIN Representation Learning** \[10, 14\]. Given an AHIN \( G = (V, E, A) \), the representation learning task is to learn a function \( f : V \rightarrow \mathbb{R}^D \) that maps each node \( v \in V \) to a vector in a \( D \)-dimensional space \( \mathbb{R}^D \), \( D \ll |V| \) that are capable of preserving both structural and semantic relations among them.

To solve this problem, we propose a novel attribute-aware AHIN embedding model named Vendor2Vec which consists of attribute-aware meta-path random walk and skip-gram model. Given an AHIN \( G = (V, E, A) \) with schema \( TG = (T, R) \), and a meta-path scheme \( \mathcal{P} \) in the basic form: \( T_1 \rightarrow \cdots \rightarrow T_i \rightarrow \cdots \rightarrow T_l \rightarrow T_{l+1} \), we use attribute-aware meta-path to guide a random walker in AHIN, the transition probability at step \( i \) is calculated as:

\[
p(u^{i+1} | v^i, \mathcal{P}) = \begin{cases} 
\frac{\text{sim}(f(v^i), f(u^{i+1}))}{\sum_{u'' \in N_{T_{i+1}}(v^i)} \text{sim}(f(v^i), f(u''))} & \text{if } u^{i+1} \in E, \phi(u^{i+1}) = \phi(u') = T_{l+1} \\
\frac{|N_{T_{i+1}}(v^i)|}{(v^i, u^{i+1}) \in E, \phi(u^{i+1}) = T_{l+1}, v' = 0} & \text{otherwise,}
\end{cases}
\]
we develop a set of crawling tools to scrape weekly snapshots from
darknet markets to fully evaluate the performance of our proposed
method. For these experiments, we use the Valhalla, Dream Mar-
ket, SilkRoad2 and Evolution. To ensure the quality and
feasibility of our crawled dataset, we conduct a detailed
evaluation of the crawled dataset. The results from Table 3 show that
 Vendor2Vec consistently and significantly outperforms all state-
of-the-art embedding models. The success of Vendor2Vec lies in:
(1) the better consideration and accommodation of the hetero-
genous property of AHIN; (2) the advantage of the attribute setting
and the proposed attribute-aware meta-path guided random walk for
sampling the high-quality path instances (i.e., without the attribute
information such as writing and photograph styles, the generated
path instances are of low quality and less useful to our application).

3.4 Comparisons with Alternative Approaches

In this set of experiments, based on our collected datasets, we com-
pare our developed system uStyle-uid with alternative approaches:
(1) feeding all the features (i.e., $f_3$, $f_6$, and feature vectors of vendors and drugs) into a generic DNN [19] to make the identification (denoted as Hybrid-DNN); (2) replacing the $vIdentifier$ in $uStyle-uID$ by a generic DNN (denoted as AHIN-DNN); (3) replacing the $vIdentifier$ in $uStyle-uID$ by SVM (denote as AHIN-SVM). For the generic DNN, we implement the model in Keras [40] and retain the default parameters. The experimental results are illustrated in Table 4. From the results we can observe that AHIN-DNN added the knowledge represented as AHIN performs better than Hybrid-DNN, which shows that using meta-path based approach over AHIN is able to build the higher-level semantic connection between vendors with a more expressive view. We also note that $uStyle-uID$ significantly outperforms other baselines, which demonstrates that $uIdentifier$ indeed helps the performance compared with the generic DNN and state-of-the-art shallow learning classification model.

Table 4: Comparisons of other alternative approaches.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Method</th>
<th>Val</th>
<th>DM</th>
<th>SR2</th>
<th>Evol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>Hybrid-DNN</td>
<td>0.831</td>
<td>0.839</td>
<td>0.833</td>
<td>0.838</td>
</tr>
<tr>
<td></td>
<td>AHIN-DNN</td>
<td>0.854</td>
<td>0.876</td>
<td>0.856</td>
<td>0.863</td>
</tr>
<tr>
<td></td>
<td>AHIN-SVM</td>
<td>0.843</td>
<td>0.851</td>
<td>0.847</td>
<td>0.848</td>
</tr>
<tr>
<td></td>
<td>$uStyle-uID$</td>
<td>0.876</td>
<td>0.903</td>
<td>0.881</td>
<td>0.889</td>
</tr>
<tr>
<td>$F1$</td>
<td>Hybrid-DNN</td>
<td>0.809</td>
<td>0.818</td>
<td>0.812</td>
<td>0.814</td>
</tr>
<tr>
<td></td>
<td>AHIN-DNN</td>
<td>0.841</td>
<td>0.864</td>
<td>0.845</td>
<td>0.855</td>
</tr>
<tr>
<td></td>
<td>AHIN-SVM</td>
<td>0.832</td>
<td>0.837</td>
<td>0.832</td>
<td>0.840</td>
</tr>
<tr>
<td></td>
<td>$uStyle-uID$</td>
<td>0.865</td>
<td>0.894</td>
<td>0.868</td>
<td>0.879</td>
</tr>
</tbody>
</table>

4 CROSS-MARKET DRUG TRAFFICKER IDENTIFICATION AND CASE STUDIES

To better understand and gain deeper insights into the ecosystem of drug trafficking in darknet markets, we further apply our developed system $uStyle-uID$ for cross-market drug trafficker identification. For the detected cross-market vendor pairs, we further sample 798 pairs and validate them using conclusive evidences. Among these 798 detected cross-market pairs, 726 pairs (90.09%) are with high confidence that they are the same entities and 22 pairs are uncertain (2.76%). As shown in Figure 4, for one of our detected cross-market vendor pairs, though “The***shop” on Evolution and “****Store” on Valhalla have different usernames, PGP keys and with similar writing and photography styles. After further investigation, we find that “The***shop” and “****Store” are both the members of a group named “DrugShop” in Finland according to the description in the terms and conditions of vendors. This indicates that they might work together as an organization and we can link them to the same group of drug traffickers. Such kind of information can provide investigative insight for law enforcement to trace their activities and thus to build proactive defenses.

5 RELATED WORK

To combat drug trafficking in darknet markets, there have been many research efforts on darknet market data analysis [3, 9, 20, 40]. Among these studies, there have been some methods proposed to tackle the challenges of authorship identification such as [20, 40]. However, these approaches mainly relied on either stylometry analysis or photography style analysis. Different from existing works, we propose to leverage both writing and photography styles together with their contents for drug trafficker identification.

In order to depict different entities, associated attributes and the rich relationships among them, it is important to model them properly. Though HIN has shown the success of modeling different types of entities and relations [11–13, 21, 35, 42, 43], it has limited capability of modeling additional attributes attached to entities. To address this challenge, we propose to use AHIN for representation. To better address representation learning for HIN, many efficient network embedding methods have been proposed such as meta-graph2vec [11], metapath2vec [10], HIN2vec [14]. However, these models are unable to deal with the attribute information associated with each entity. To address this issue, we propose Vendor2Vec to learn the desirable node representations in AHIN.

6 CONCLUSION

To combat drug trafficking, in this paper, we design and develop an intelligent system named $uStyle-uID$ to automate drug trafficker identification in darknet markets. In $uStyle-uID$, we propose to leverage both writing and photography styles at the first attempt. To depict vendors, drugs, texts, photos and their associated attributes as well as the rich relationships among them, we present a structural AHIN to model them which gives the vendors higher-level semantic representations. Then, a meta-path based approach is used to characterize the semantic relatedness over vendors. To efficiently measure the relatedness over vendors in AHIN, we propose a new network embedding model Vendor2Vec which leverages complementary attribute information attached in the nodes to guide the meta-path based random walk for path instances sampling. Then, we transform the identification task into a link prediction problem and further present a learning model named $vIdentifier$ to solve the problem. The promising experimental results on the collected datasets from four darknet markets demonstrate that $uStyle-uID$ outperforms alternative approaches.

7 ACKNOWLEDGMENTS

Y. Zhang, Y. Fan, W. Song, S. Hou, Y. Ye and X. Li’s work is partially supported by the DoJ/NIJ under grant NIJ 2018-75-CX-0032; the NSF under grants CNS-1618629, CNS-1814825, CNS-1845138 and OAC-1839909; the WV Higher Education Policy Commission Grant (HEPC.dsr.18.5); and the WVU Research and Scholarship Advancement Grant (R-844).