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

Anonymous Author(s)

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 together 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, a meta-path based approach is exploited to incorporate higher-level semantics to establish relatedness over traffickers. To efficiently measure the relatedness over nodes (i.e., traffickers) in 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 *uIdentifier* to classify if a given pair of traffickers are the same individual. Comprehensive experiments on data collections from four 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.

CCS CONCEPTS

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

KEYWORDS

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

WWW'19, May 13-17, 2019, San Francisco, CA, USA

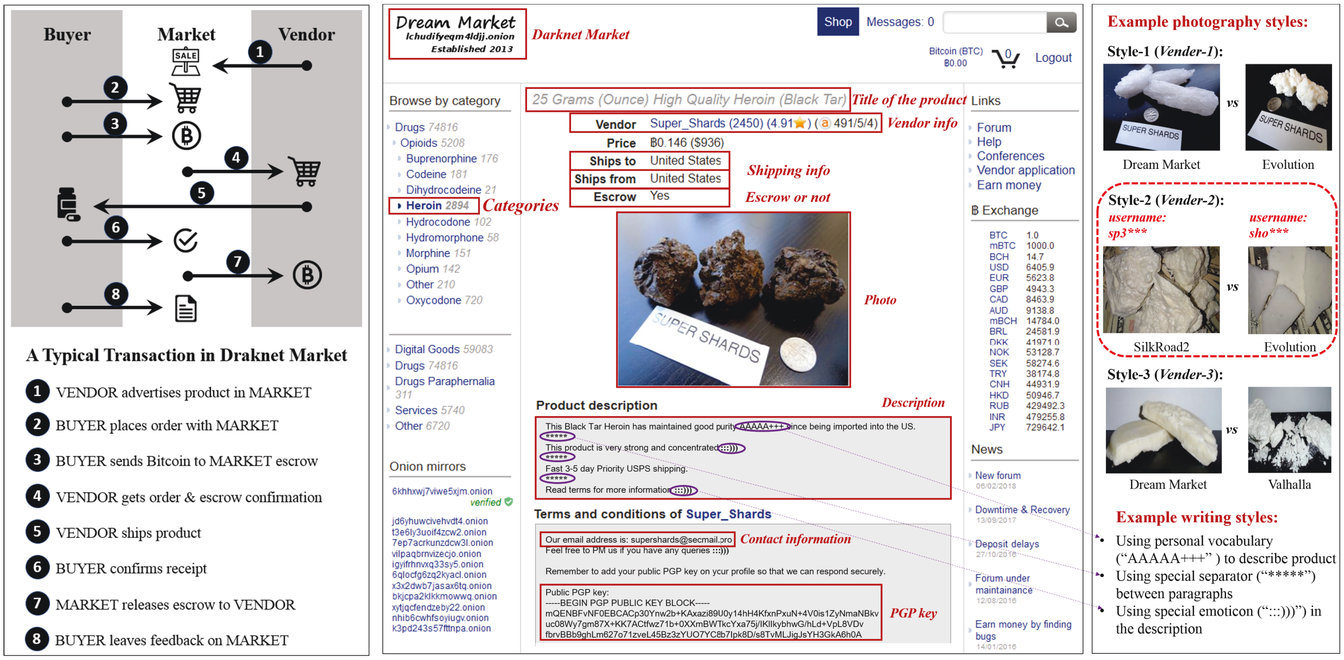
ACM ISBN 978-1-4503-5639-8/18/04…$15.00
https://doi.org/10.1145/3178876.3186094

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 [6, 15, 17, 28, 33, 42, 43]. 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 [34], Dream Market [27], Valhalla [40], 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 [10].

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 [8]. However, one of the key challenges is that drug traffickers may maintain multiple accounts across different markets for the reasons [3, 6, 35] such as (1) ripper: an old account has lost the trust of other members (i.e., scamming), (2) branding: a trafficker (i.e., vendor) creates an alias to positively review his/her own products or services, and (3) anonymity: a user wants to conceal the identities. Such case can also happen within the same market: if an old account has been banned by a moderator (e.g. due to abuse), the user may create a new one without a sweat. Linking different accounts to the same individuals is essential to track their status and better understand the online drug trafficking ecosystem [45]. Given the growing scale of darknet markets and the large number of user accounts, it is simply impossible to manually link suspicious accounts and track their latest status. Therefore, it is highly desirable to develop novel methodologies that can automatically link multiple accounts of the same individuals in darknet markets.

To automate the process, some of existing approaches [2, 7, 20] relied on stylometry analysis which aims at linking different accounts to the same user based on his/her writing styles (e.g., as illustrated in Figure 1.(b) and (c), the vendor likes using specific emoticon in the product description). Since drug traffickers in darkent markets have to prove the possession of illegal drugs by posting their own product photos, their distinct photography styles might



(a) A typical transaction in darknet market. (b) Illicit drug advertised by a trafficker (i.e., vender) in darknet market. (c) Illustration of writing & photography styles.

Figure 1: Illustration of drug trafficking in Darknet market.

be revealed by the posted photos. A recent research [45] proposed to link multiple accounts of the same vendors in different darknet markets based on their distinct photography styles (e.g., the way to display products and camera model as shown in Figure 1.(c)). Though each kind of analysis has shown success in fingerprinting underground market participants, using them respectively may suffer different challenges (e.g., stylometry analysis only is sensitive to the language of content [45], while photography style analysis only may face the challenge of intrinsic ambiguity arising from resale or photo plagiarizing). *Can we leverage both writing and photography styles to develop an integrated framework for drug trafficker identification in darknet markets?* To the best of our knowledge, this has yet studied in the open literature.

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) [26] for modeling and use meta-path based approach [36] to build up relatedness over vendors. 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. The major contributions of this work are summarized as follows:

- ***A novel feature representation elegantly integrating writing and photography styles to characterize drug traffickers in darknet markets.*** To comprehensively characterize vendors in darknet markets, we extract various features including (1) vendors' posted texts and writing styles, (2) vendors' posted photos and photography styles, (3) attributed features of vendors and drugs, and (4) different kinds of relations among them. To depict vendors, drugs, texts, photos and their associated attributes as well as the rich relationships among them, it is important to model them properly in order to facilitate the task of drug trafficker identification. Though heterogeneous information network (HIN) [36] 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 present attributed HIN (AHIN) for representation and then exploit meta-path based approach to characterize the relatedness over vendors. Despite being a natural way to handle different types of relations and entities associated with additional attributes, we propose such elegant model leveraging both writing and photography styles for drug trafficker identification in darknet markets at the first attempt.

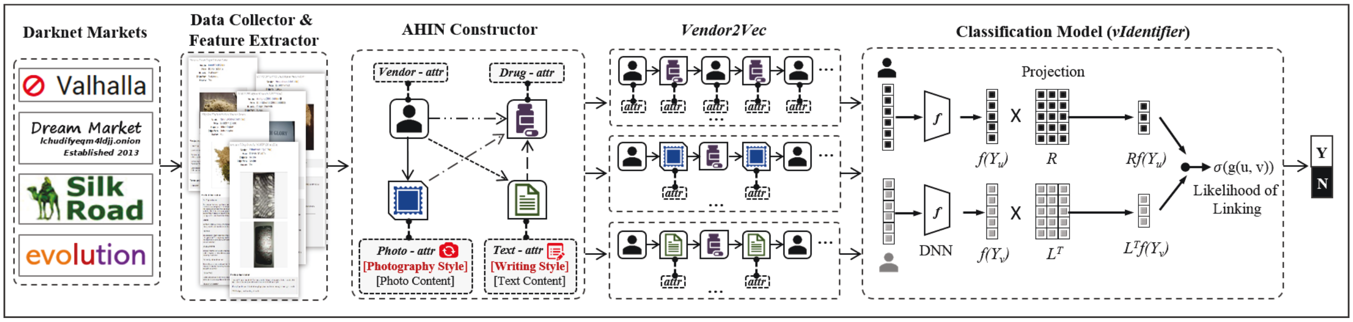


Figure 2: System architecture of *uStyle-uID*.

- ***A new AHIN embedding model***. HIN embedding models such as [12, 13] have been proposed to learn latent representations of nodes in HIN. However, the existing HIN embedding models are unable to deal with the attribute information associated with each entity; without the attribute information such as writing and photography styles, the learned embeddings are low quality and less useful to our application. To address this issue, our proposed model named *Vendor2Vec* leverages complementary attribute information of each node to guide the meta-path based random walk for path-instance sampling.
- ***An intelligent system to automate drug trafficker identification in darknet markets.*** Based on the data collected from four darknet markets (i.e., Valhalla, Dream Market, SilkRoad2 and Evolution), we design and develop an intelligent system named *uStyle-uID* integrating our proposed method to automate drug trafficker identification. Comprehensive experimental studies are provided to validate the performance of our developed system in comparisons with alternative approaches. *Only through collective and persistent effort of the whole society, can we build up a drug-free world - one community a time.*

2 SYSTEM ARCHITECTURE

The overview of our developed system *uStyle-uID* for drug trafficker identification in darknet markets is shown in Figure 2, which consists of the following components:

- **Data Collector.** A set of web crawling tools are first developed to collect data from darknet markets (e.g., Valhalla and Dream Market widely used for illicit drug trading). The collected data includes vendors' profiles and their posted contents (e.g., texts and photos).
- **Feature Extractor.** Based on the data collected from the previous module, it extracts the following features for further analysis: (1) vendors' posted texts and writing styles; (2) vendors' posted photos and photography styles; (3) attributed features of vendors (i.e., usernames, contact information and PGP keys) and drugs (i.e., drug category, shipping and escrow information); (4) various kinds of relation-based features, including *vendor-drug*, *vendor-text*, *vendor-photo*, *text-drug* and *photo-drug* relations. (See Section 3.1 for details.)
- **AHIN Constructor.** Based on the above extracted features, an attributed HIN (AHIN) is first constructed to model different relationships among different types of entities associated with additional attributes; and then different meta-paths are built from the AHIN to capture the relatedness over vendors in darknet markets from different views. (See Section 3.2 for details.)
- ***Vendor2Vec.*** In this module, a new AHIN embedding model *Vendor2Vec* is proposed to learn the low-dimensional representations for the nodes (i.e., vendors). The proposed model leverages complementary attribute information of each node to guide meta-path based random walk for path-instance sampling; then a skip-gram model is utilized to learn effective representations of vendors based on AHIN. (See Section 3.3 for details.)
- **Classification Model.** After the latent representation learning for the nodes in AHIN using *Vendor2Vec*, the pairwise vendors represented by the learned low-dimensional vectors will be fed to the devised *vIdentifier* to train the classification model, based on which any new given pair of vendors can be predicted as either they are the same individual or not. (See Section 3.4 for details.)

3 PROPOSED METHOD

In this section, we introduce the detailed approaches of how we represent vendors in darknet markets leveraging their writing and photography styles as well as other attributed features and relations, and how we solve the problem of drug trafficker identification based on the representations.

3.1 Feature Extraction

We propose to characterize vendors in darknet markets in a comprehensive view by extracting various features, including (1) vendors' posted texts and writing styles, (2) vendors' posted photos and photography styles, (3) attributed features of vendors and drugs, and (4) different kinds of relations among vendors, drugs, posted texts and photos. The feature extraction is described as follows.

(1) Posted text and writing style feature 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 propose to exploit *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: *lexical*, *syntactic* and *structural*. 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 (all, upper, or lower case), *ii*) number of digits/white spaces/special characters (e.g., '%', '\$', '@'), *iii*) number of words, *iv*) average word length, and *v*) vocabulary richness. Vocabulary richness is useful to gain an understanding about an author's use of vocabulary and the complexity of the author's language, which can be calculated by Yule's K measure, Yule's I measure, or Simpson's D measure [39]. 2) ***Syntactic features*** capture the writing style from the sentence structure. In this category, we adopt four types of features: *i*) frequency of punctuation (e.g., '.', '!', '?'), *ii*) frequency of function word (e.g., for, to, the), *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) extracted by using Stanford log-linear parts-of-speech tagger [38]. 3) ***Structural features*** represent the way an author organizes the layout of his/her posted text. We consider the following four types of structural features: *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. Table 1 summarizes the features we extract to describe the writing style of each vendor. For each posted text by a vendor, 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 for the further analysis.

Table 1: Extracted features for writing style representation.

Category	Feature
Lexical	Number of characters (all, upper, or lower case) Number of digits/white spaces/special characters Number of words Average word length Vocabulary richness
Syntactic	Frequency of punctuation Frequency of function word Number of sentences beginning with a capital letter Frequency of parts-of-speech n-grams
Structural	Total number of paragraphs Indentation of paragraph Whether there's separator between paragraphs Number of words/sentences/characters per paragraph

(2) Posted photo and photography style extraction. To represent the content of a posted photo, we propose to utilize *image2vec* [14] to convert it into a fixed length feature vector (empirically, we set the dimension to 100). Since drug traffickers in darkent markets have to prove the possession of illegal drugs by posting their own product 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. This information can serve as an important source of knowledge about how vendors capture the photo and what tools are involved in the process of production which often includes: *i*) camera make and model, *ii*) camera angle, *iii*) exposure time, *iv*) focal length, and *v*) image size. Those EXIF-related features have been widely employed in the field of digital image forensics [16]. 2) ***High-level features*** are extracted from the photo's original content. Since color tones and saturation play important roles in photography and color psychology, HSV (hue, saturation, value) color space often offers a convenient representation for content analysis of photos [9]. Accordingly, we first convert the photo into its HSV 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 [9]; 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 for the further analysis.

(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* which is used to denote a vendor, *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* to denote this drug, *escrow information* and *shipping information* (e.g., from where and to where). Then, we apply one-hot encoding [47] to convert the extracted features to a binary feature vector to be an attribute associated with each vendor/drug.

(4) Relation-based Features. In order to characterize the rich relations among vendors, drugs, posted texts and photos, we propose to extract the following relation-based features. 1) ***R1***: To depict the relation that a vendor sells a drug, we build the ***vendor-sell-drug*** matrix **S** where each element $s_{i,j} \in \{0, 1\}$ indicates whether vendor i sells drug j. 2) ***R2***: To describe the relation between a vendor and a text, we generate the ***vendor-write-text*** matrix **W** where each element $w_{i,j} \in \{0, 1\}$ denotes if vendor i writes text j. 3) ***R3***: To denote the relation that a vendor posts a photo, we build the ***vendor-post-photo*** matrix **P** where each element $p_{i,j} \in \{0, 1\}$ indicates whether vendor i posts photo j. 4) ***R4***: To represent whether a text describes a drug, we generate the ***text-describe-drug*** matrix **D** where each element $d_{i,j} \in \{0, 1\}$ indicates whether text i describes drug j. 5) ***R5***: To depict whether a photo characterizes a drug, we generate the ***photo-characterize-drug*** matrix **C** where each element $c_{i,j} \in \{0, 1\}$ denotes if photo i characterizes drug j.

3.2 AHIN Construction

After the feature extraction in Section 3.1, each vendor, drug, text, or photo is associated with its related attributes represented by a feature vector. For example, a *vendor* is associated with attributes of *username*, *PGP key* and *contact information* represented by its corresponding feature vector, while a *photo* is associated with attributes of its *photo content* and *photography style* represented by its concatenated feature vector. In order to depict vendors, drugs, texts, photos and their associated attributes as well as the rich relationships among them (i.e., ***R1***-***R5***), it is important to model them properly in order to facilitate the task of drug trafficker identification. Though heterogeneous information network (HIN) [36] 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 HIN (AHIN) for representation.

Definition 3.1. ***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 $\mathcal{X}_i$ 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 ϕ: $\mathcal{V} \to \mathcal{T}$ and a *relation* type mapping ψ: $\mathcal{E} \to \mathcal{R}$, 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*** [26] 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 $\mathcal{R}$.

AHIN provides not only the network structure of data associations but also a high-level abstraction of categorical associations. In our application, we have four entity types (i.e., vendor, drug, text and photo) and five types of relations among them (i.e., ***R1-R5***); meantime, each entity is also attached with an extracted feature vector representing its associated attributes as described in Section 3.1. Based on the definitions above, the network schema for AHIN in our application is shown in Figure 3 (to facilitate the illustration, attribute information is shown in its original form), which enables the entities to be represented in a comprehensive manner utilizing both their content-based and relation-based information.

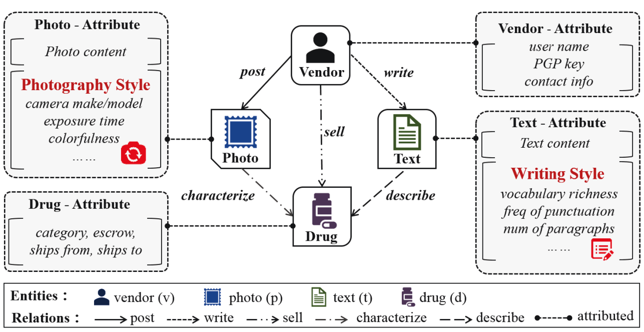


Figure 3: Network schema for AHIN.

Different types of entities and relations motivate us to use a machine-readable representation to enrich the semantics of relatedness among vendors in darknet markets. Toward this objective, we propose to adopt the concept of meta-path [36] to formulate higher-order relationships among entities in AHIN. Here, we follow this concept and extend it to our application.

Definition 3.2. ***Meta-path***[36]. A meta-path $\mathcal{P}$ is a path defined on the network schema $\mathcal{T}_{\mathcal{G}} = (\mathcal{T}, \mathcal{R})$, and is denoted in the form of $T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} ... \xrightarrow{R_L} T_{L+1}$, which defines a composite relation $R = R_1 \cdot R_2 \cdot \ldots \cdot R_L$ between types T_1 and T_{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 numerous meta-paths. In our application, we focus on three most meaningful meta-paths for characterizing the relatedness over vendors in darknet markets - i.e., ***PID1-PID3*** as shown in Figure 4. These three meta-paths jointly characterize the relatedness between two vendors from different views: (1) ***PID1***: $vendor \xrightarrow{sell} drug \xrightarrow{sell^{-1}} vendor$ means that two vendors can be connected through the path that they both sell the same kind of drug (e.g., heroin); (2) ***PID2***: $vendor \xrightarrow{post} photo \xrightarrow{characterize} drug \xrightarrow{characterize^{-1}} photo \xrightarrow{post^{-1}} vendor$ 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); and (3) ***PID3***: $vendor \xrightarrow{write} text \xrightarrow{describe} drug \xrightarrow{describe^{-1}} text \xrightarrow{write^{-1}} vendor$ denotes that two vendors can be connected if their written texts describe the same kind of drug.

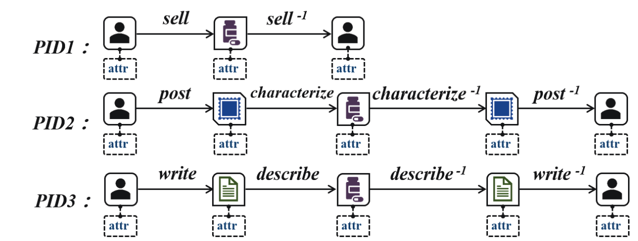


Figure 4: Meta-paths built for drug trafficker identification.

3.3 *Vendor2Vec*

To quantify the relatedness over AHIN entities, traditional approaches [21, 36, 46] often rely on matrix factorization for similarity measure. Unfortunately, the computational cost of matrix factorization is expensive and not scalable; meantime, it suffers from various statistical drawbacks due to ignorance of uncertainty [18]. To tackle this problem, we propose to develop a scalable representation learning method for AHIN. We first formalize the problem of AHIN representation learning as following.

Definition 3.3. ***AHIN Representation Learning*** [12, 13]. Given an AHIN $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, the representation learning task is to learn a function $f : \mathcal{V} \to \mathbb{R}^D$ 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.

Due to the heterogeneous property of AHIN (i.e., network consisting of multi-typed entities and relations), it is difficult to directly apply conventional homogeneous network embedding techniques (e.g., DeepWalk [32], LINE [37], node2vec [18]) to learn the latent representations for AHIN. To resolve this difficulty, recently developed HIN embedding models such as *metapath2vec* [12] employed meta-path based random walks and heterogeneous skip-grams to learn the latent representations for HIN. However, *metapath2vec* is unable to deal with the attribute information associated with each entity; without the attribute information such as writing and photography styles, the generated path instances are of low quality and less useful to our application.

We propose to address this issue based on the following observation. At each step of random walk, the walker should wander to a neighbor of higher proximity with a larger probability because such a neighbor might contain more reliable semantics by completing a closer link. For example, as illustrated in Figure 5, *Text-1*, *Text-2* and *Text-3* posted by *Vendor-1*, *Vendor-2* and *Vendor-3* both describe the same drug (i.e., *Drug-1*: GBL). If the walker is located at *Drug-1*, *metapath2vec* counting on a uniform sampling strategy would pick the two neighbor nodes (*Text-2* vs. *Text-3*) with equal probability; however, if we consider the attribute feature vector attached to each text (including text content and writing style information extracted in Section 3.1), *Text-2* would arguably receive more attention (and priority) because it possesses more similar text content and writing styles to *Text-1*. Along this line of reasoning, we propose to leverage the attribute information of each node to guide meta-path based random walk and develop a novel *attribute-aware* AHIN embedding model named *Vendor2Vec*. More specifically, to learn node representations for vendors in AHIN, our *Vendor2Vec* model consists of two steps: (1) attribute-aware meta-path guided random walk is used to sample high-quality path instances first; (2) a skip-gram model is then utilized to learn the node representation. We will elaborate them next.

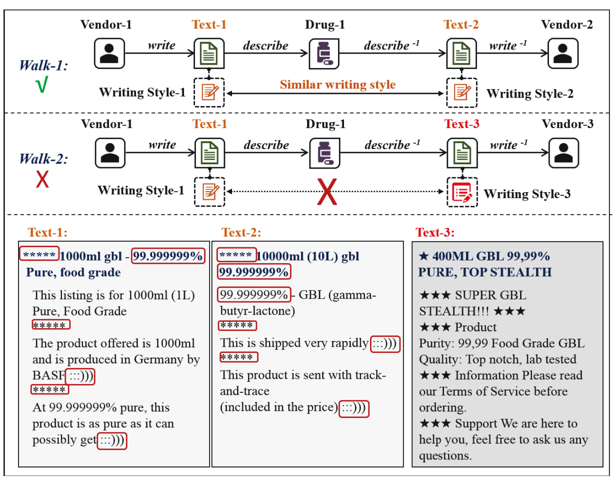


Figure 5: Example of attribute-aware setting.

Attribute-aware meta-path guided random walk. Given a source node v_j in homogeneous network, a random walk is defined by a stochastic process with random variables $v_j^1, v_j^2, ..., v_j^k$ such that v_j^{k+1} is a node chosen at random from the neighbors of node v_k. The transition probability $p(v_j^{i+1}|v_j^i)$ at step i is the normalized probability distributed over the neighbors of v_j^i by ignoring their node types. Apparently, this formulation is not applicable to heterogeneous networks in which semantic and structural correlations among different types of nodes matter. A remedy to this formulation is to use *attribute-aware* meta-path to guide a random walker in an AHIN while generating the paths for multiple types of nodes. Given an AHIN $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ with schema $\mathcal{T}_\mathcal{G} = (\mathcal{T}, \mathcal{R})$, and a meta-path scheme $\mathcal{P}$ in the basic form: $T_1 \to \cdots T_t \to T_{t+1} \cdots \to T_L \to T_{L+1}$, we put a random walker to traverse the AHIN and then we propose to calculate transition probability at step i as following:

$$p(v^{i+1}|v_t^i, \mathcal{P}) = \begin{cases} \dfrac{sim(\mathbf{f}_{v'}, \mathbf{f}_{v^{i+1}})}{\sum_{v^c \in N_{T_{t+1}}(v_t^i)} sim(\mathbf{f}_{v'}, \mathbf{f}_{v^c})} \\ \qquad (v_t^i, v^{i+1}) \in \mathcal{E}, \phi(v^{i+1}) = \phi(v') = T_{t+1} \\ \dfrac{1}{|N_{T_{t+1}}(v_t^i)|} \\ \qquad (v_t^i, v^{i+1}) \in \mathcal{E}, \phi(v^{i+1}) = T_{t+1}, v' = \emptyset \\ 0 \qquad \text{otherwise,} \end{cases} \tag{1}$$

where v' denotes the latest entity the walker visited is with the same type of v^{i+1}, $sim(\mathbf{f}_{v^i}, \mathbf{f}_{v^j})$ is the similarity between two entities' attribute vectors (e.g, it can be calculated by using cosine similarity measure), ϕ is the node type mapping function, $N_{T_{t+1}}(v_t^i)$ denote T_{t+1} type of neighborhood of node v_t^i, v^c denotes a node in $N_{T_{t+1}}(v_t^i)$. The paths generated by our attribute-aware meta-path guided random walk are able to preserve both the semantic and structural relations between different types of nodes, which greatly facilitates the transformation of AHIN structures to skip-gram. Since we have three different meta-paths in our application, we simply combine the path instances sampled via each meta-path, which will be then fed to the skip-gram model.

Skip-gram model. After mapping the word-context concept in a text corpus to an AHIN via attribute-aware meta-path guided random walk (i.e., a sentence in the corpus corresponds to a sampled path and a word corresponds to a node), we apply the skip-gram model [29] to the paths and propose to maximize the probability of observing a node's neighborhood (within a window w) conditioned on its current representation. The objective function of our skip-gram model is given by:

$$\arg\max_{Z} \sum_{-w \le k \le w, j \ne k} \log p(v_{j+k}|Y(v_j)), \tag{2}$$

where $Y(v_j)$ is the current representation vector of v_j, $p(v_{j+k}|Y(v_j))$ is defined by the softmax function:

$$p(v_{j+k}|Y(v_j)) = \frac{\exp(Y(v_{j+k}) \cdot Y(v_j))}{\sum_{q=1}^{|V|} \exp(Y(v_q) \cdot Y(v_j))}. \tag{3}$$

To improve the efficiency of our implementation, we first apply hierarchical softmax technique [31] to solve Eq.(3); then use stochastic gradient descent (SGD) [4] to train the skip-gram model.

3.4 Classification Model

Given an AHIN $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ and the designed meta-paths, for every node $v \in \mathcal{V}$ (i.e., vendor), we can learn its latent representation (i.e., embedding vector $Y_v \in \mathbb{R}^D$) encoding both writing and photograph styles using the above proposed *Vendor2vec*. The problem of determining if a given pair of vendors are the same individual can be considered as a link prediction in an AHIN. For homogeneous information networks, the method proposed in [1] has achieved great success for the task of link prediction; however, low-rank asymmetric projection cannot be directly applied to handle heterogeneous properties of AHIN in our application. In this paper, utilizing the proposed method in [1], we propose to devise a classification model for AHIN (named *vIdentifier* as shown in Figure 6) to predict the likelihood of the link between two nodes.

As deep neural network (DNN) has been found useful to reduce overfitting and improve generalization because its hidden layers

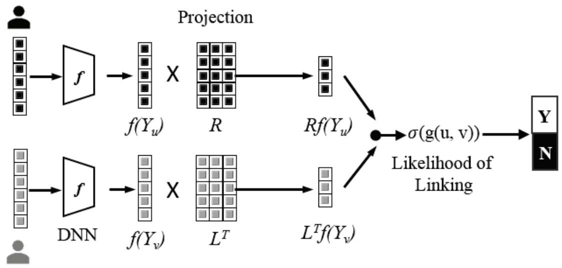


Figure 6: *vIdentifier* for drug trafficker identification.

can reveal hidden correlations among the input data [1], we consider a novel application of DNN into node embedding here - i.e., to learn $f_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^d$ that maps a vendor's learned latent representation onto a low-dimensional manifold. The mapping function f_θ is defined as [1]:

$$f_\theta : Y_u \rightarrow FC_{\{\mathbf{W}_1,\mathbf{b}_1\}} \rightarrow BatchNorm \rightarrow relu \\ \rightarrow FC_{\{\mathbf{W}_2,\mathbf{b}_2\}} \rightarrow BatchNorm \rightarrow f_\theta(Y_u), \tag{4}$$

where $FC_{\{\mathbf{W},\mathbf{b}\}}$ is a fully-connected layer with weight matrix $\mathbf{W}$ and bias vector $\mathbf{b}$, BatchNorm is described in [22], $relu(x) = max(0, x)$ is an element-wise activation function, and $\theta = \{\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2, ...\}$.

Given a pair of vendors, to define a general link function $g(u, v) \in \mathbb{R}$ to determine whether they are the same individual, we consider a low-rank affine projection in the manifold space [1]:

$$g(u, v) = f(Y_u)^T \times M \times f(Y_v), \tag{5}$$

where the low-rank projection matrix $M = L \times R$ with $L \in \mathbb{R}^{d \times b}$ and $R \in \mathbb{R}^{b \times d}$, $b < d < D$. We can factor $g(u, v)$ into an inner product $\langle L^T f(Y_u), Rf(Y_v)\rangle$.

Then, we utilize the method proposed in [1] to conduct the optimization, whose function is further devised as following in our application:

$$Pr(\mathcal{G}) \propto \prod_{u,v \in \mathcal{V}} \sigma(g(u, v))^{D_{uv} f(u,v)} (1 - \sigma(g(u, v)))^{1-f(u,v)}, \tag{6}$$

where $\sigma(x) = 1/(1 + exp(-x))$ is the standard logistic,

$$f(u, v) = \begin{cases} 1 & \text{if } u \text{ and } v \text{ are the same individual} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

D_{uv} is the frequency that vendors u and v co-occur within a specific window in the path instances sampled by our above proposed attribute-aware meta-path random walk.

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we conduct four sets of experiments using data collections from darknet markets to fully evaluate the performance of our developed system *uStyle-uID* for drug trafficker identification: (1) In the first set of experiments, we evaluate the effectiveness of different kinds of features; (2) In the second set of experiments, we examine the proposed method *Vendor2Vec* by comparisons with state-of-the-art network embedding models; (3) In the third set of experiments, we compare our developed system *uStyle-uID* with alternative methods; (4) In the last set of experiments, we evaluate the parameter sensitivity, stability and scalability of *uStyle-uID*.

4.1 Experimental Setup

To fully evaluate our proposed method, we have collected the data from four different darknet markets (i.e., *Valhalla* and *Dream Market* that are still active and widely used for illicit drug trading, *SilkRoad2* and *Evolution* that have been shut down but were the most popular ones for drug trafficking). For the former two darknet markets, we develop a set of crawling tools to scrape weekly snapshots in these two markets from June 2017 to August 2017. The scraped snapshots contain vendor profiles and the product pages. For the latter two markets, we collect their public data dumps including vendors and product pages. Though SilkRoad2 and Evolution have been shut down, the data from these markets still provides us valuable information and knowledge to investigate the online drug trafficking ecosystem. Note that these darknet markets may include the products beyond drugs, but we only focus on the drug-related products. After collecting the data from these darknet markets, we merely retain the vendors who at least posted two different drugs, each of which should be at least with one transaction (i.e., this information can be indicated in the comments). Due to the anonymity of darknet markets, it is difficult to access the actual ground-truth. Similar to the approach used in [45], we propose an alternative solution: for a given vendor, we randomly split his/her posted texts and photos into two even parts to form a positive example; then we randomly match this given vendor to the other in the same darknet market to generate a negative example. In this pseudo setting, the attribute vector attached to each vendor node is set as null. We summarize the data we collect and extract from each market as below.

- **Valhalla** (denoted as **Val**) is a Finnish market starting in October 2013. It is still active and has become one of the largest global market for illicit drug trading [11]. By the date, we have collected the data from 522 vendors related to 191 kinds of sold drugs posted in 13,150 threads in Valhalla. After feature extraction and based on the designed network schema, the constructed AHIN has 27,535 nodes (i.e., 1,044 nodes with type of vendor, 191 nodes with type of drug, 13,150 nodes with type of text, 13,150 nodes with type of photo) and 54,698 edges including relations of *R1-R5*.
- **Dream Market** (denoted as **DM**), founded in November 2013, is one of the most popular, most well-stocked and most reputed darknet markets that exists today [41]. We have collected the data from 2,547 vendors related to 67,270 posted threads in Dream Market. The constructed AHIN for this market has 139,493 nodes and 281,080 edges.
- **SilkRoad2** (denoted as **SR2**), the successor of the well-known darkent market of SilkRoad, was established in November 2013. It was growing quickly but was taken down by authorities in November 2014 [45]. We have extracted the information from 15,231 thread posted by 681 vendors in the data dump of SilkRoad2. The constructed AHIN has 31,354 nodes and 66,648 edges.
- **Evolution** (denoted as **Evol**), founded in January 2014, was the largest darknet marketplace after the SilkRoad2. In March 2015, its the administrators unexpectedly shut down the market and took away all the bitcoins that users deposited to the market, the value of which was estimated to be $11.7 million dollars [45]. We have extracted the information from 36,798 threads posted by 1,650 vendors in the data dump of Evolution. The constructed AHIN for this market has 79,451 nodes and 164,792 edges.

In the following experiments, we conduct ten-fold cross validations and use accuracy (denoted as *ACC*) and *F1* as performance measures to evaluate different methods in drug trafficker identification. The parameters for *Vendor2Vec* are empirically set as follows: node dimension $D = 100$, walks per node $r = 10$, walk length $l = 80$ and window size $w = 10$, while for the parameter of d in *vIdentifier* is set as 30 and the remaining ones are consistent with [1].

4.2 Comparisons of Different Features

In this set of experiments, based on the datasets described in Section 4.1, in our developed system *uStyle-uID*, we first evaluate the effectiveness of different features for drug trafficker identification.

- **Text-based**. This type of features only consider the vendors' posted text information, including (1) *text content only* (denoted as ***f-1***): each posted text is represented by the converted fixed dimensional feature vector, (2) *writing style only* (denoted as ***f-2***): each posted text is represented by its extracted stylometry feature vector, and (3) *text content and writing style* (denoted as ***f-3***): the concatenation of *f-1* and *f-2*.
- **Photo-based**. This type of features only consider the vendors' posted photo information, including (1) *photo content only* (denoted as ***f-4***): each posted photo is represented by the converted fixed dimensional feature vector, (2) *photography style only* (denoted as ***f-5***): each posted photo is represented by its extracted photography feature vector, and (3) *photo content and photography style* (denoted as ***f-6***): the concatenation of *f-4* and *f-5*.

Table 2: Comparisons of different features.

Metric	Method	Feature	Val	DM	SR2	Evol
ACC	Text-based	***f-1***	0.780	0.798	0.782	0.792
		f-2	0.792	0.807	0.795	0.805
		f-3	0.803	0.819	0.806	0.815
	Photo-based	***f-4***	0.795	0.808	0.796	0.805
		f-5	0.807	0.817	0.808	0.816
		f-6	0.818	0.828	0.819	0.827
	uStyle-uID	/	**0.876**	**0.903**	**0.881**	**0.889**
F1	Text-based	***f-1***	0.782	0.784	0.772	0.784
		f-2	0.796	0.788	0.784	0.790
		f-3	0.804	0.809	0.792	0.797
	Photo-based	***f-4***	0.785	0.794	0.792	0.793
		f-5	0.798	0.806	0.796	0.802
		f-6	0.810	0.817	0.806	0.809
	uStyle-uID	/	**0.865**	**0.894**	**0.868**	**0.879**

The experimental results are illustrated in Table 2, from which we can see that different features show different performances in drug trafficker identification. To put this into perspective, (1) the relatedness over vendors depicted by style-based correlations (***f-2***: writing styles and ***f-5***: photography styles) perform better than content-based correlations (***f-1***: text contents and ***f-4***: photo contents), the reason behind which is that a vendor in the darknet market could post different drugs such that the posted texts and photos might vary but their writing and photography styles could be similar; (2) feature engineering (***f-3***: augmented features of text content and writing style and ***f-6***: augmented features of photo content and photography style) helps the performance of drug trafficker identification; (3) the photo-based features (***f-4***, ***f-5***, and ***f-6***) perform better than the text-based ones (***f-1***, ***f-2***, and ***f-3***), which demonstrates that vendors' uploaded photos are more effective than the posted texts in distinguishing their identities (i.e., this result is consistent with the conclusion drawn in [45]); (4) our proposed method *uStyle-uID* which integrates different levels of semantics, leveraging both writing and photograph styles, obtains a significantly better performance than the others (i.e., ***f-1*** - ***f-6***) in drug trafficker identification.

4.3 Network Embedding Model Comparisons

In this set of experiments, we evaluate our proposed method *Vendor2Vec* by comparisons with several state-of-the-art network embedding models, which include followings:

- **DeepWalk** [32] 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 property and attribute information, and directly feed the AHIN for embedding.
- **node2vec** [18] is generalized from DeepWalk. It learns node vectors by capturing node pairs within w-hop neighborhood via *parameterized* random walks. Same as DeepWalk, we directly feed the AHIN for embedding.
- **metapath2vec** [12] is a HIN embedding method which embeds the semantic information of a single meta-path. Since we have three different meta-paths in our application, we simply combine the path instances sampled via each meta-path, then feed to skip-gram model.

Table 3: Comparisons of network embedding models.

Metric	Method	Val	DM	SR2	Evol
ACC	DeepWalk	0.703	0.714	0.699	0.707
	node2vec	0.726	0.731	0.722	0.729
	metapath2vec	0.741	0.754	0.744	0.748
	Vendor2Vec	**0.876**	**0.903**	**0.881**	**0.889**
F1	DeepWalk	0.682	0.688	0.669	0.679
	node2vec	0.703	0.710	0.692	0.709
	metapath2vec	0.718	0.735	0.727	0.731
	Vendor2Vec	**0.865**	**0.894**	**0.868**	**0.879**

For these embedding methods, we use the same parameters as *Vendor2Vec*. Table 3 illustrates the results of different network representation learning models for drug trafficker identification. From Table 3, we can see that *Vendor2Vec* consistently and significantly outperforms all baselines for drug trafficker identification in all four darket markets in terms of *ACC* and *F1*. That is to say, *Vendor2Vec* learns significantly better vendor representations than current state-of-the-art embedding models. The success of *Vendor2Vec* lies in: (1) the proper consideration and accommodation of the heterogeneous property of AHIN (i.e., the multiple types of nodes and relations); (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 photography styles, the generated path instances are of low quality and less useful to our application).

4.4 Comparisons with Alternative Approaches

In this set of experiments, based on the dataset described in Section 4.1, we compare our developed system *uStyle-uID* with alternative approaches: (1) based on all extracted features described in Section 3.1, we feed all the features (i.e., ***f-3***, ***f-6***, and feature vectors of vendors and drugs) into a generic DNN [19] with multiple fully connected layers and a softmax layer to make the identification (denote as *Hybrid-DNN*); (2) we replace the classification model *uIdentifier* in *uStyle-uID* by a generic DNN (denote as *AHIN-DL*); (3) we replace the classification model *uIdentifier* in *uStyle-uID* by the support vector machine (SVM) (denote as *AHIN-SVM*). For the generic DNN, we implement the model in Keras [45] and retain the default parameters. 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 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 demonstrates that, for drug trafficker identification in darknet markets, *uStyle-uID* using meta-path based approach over AHIN is able to build the higher-level semantic and structural connection between vendors with a more expressive and comprehensive view and thus achieves better performance. We also note that the proposed system *uStyle-uID* significantly outperforms other baselines (i.e., *AHIN-DNN* and *AHIN-SVM*), which demonstrates that our proposed classification model *uIdentifier* indeed helps the performance in drug trafficker identification compared with the generic DNN and state-of-the-art shallow learning classification model.

Table 4: Comparisons of other alternative approaches.

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

4.5 Evaluation of Parameter Sensitivity, Stability and Scalability

In this set of experiments, we first conduct the **sensitivity** analysis of how different choices of parameters (dimension D in *Vendor2Vec* and d in *uIdentifier*) will affect the performance of *uStyle-uID* in drug trafficker identification. From the results shown in Figure 7.(a) and (b), we can observe that the performance tends to be stable in all four different darket markets when we varied the number of node dimensions in the model from $D = 50$ to $D = 250$ or $d = 20$ to $d = 100$. Therefore, *uStyle-uID* is not strictly sensitive to the parameters and is able to reach high performance under a cost-effective parameter choice. We then further evaluate the **stability** and **scalability** of *uStyle-uID*. Figure 7.(c) shows the overall receiver operating characteristic (ROC) curves of *uStyle-uID* based on the ten-fold cross validations; it achieves an impressive 0.903 average TP rate at the 0.194 average FP rate on Dream Market for drug trafficker identification. Figure 7.(d) shows the running time of *uStyle-uID* with different sizes of the datasets, which illustrates that the running time is quadratic to the number of samples. When dealing with more data, approximation or parallel algorithms can be developed. From the results and analysis above, *uStyle-uID* is efficient and scalable for large-scale AHIN mining with large numbers of nodes.

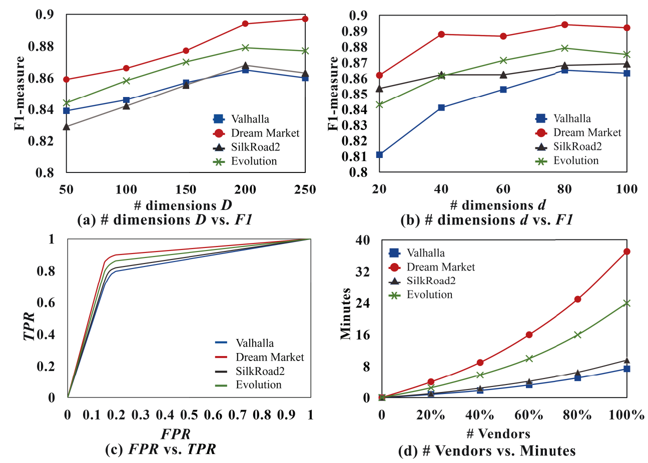
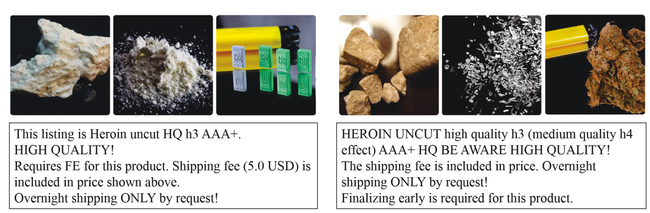


Figure 7: Parameter sensitivity, stability and scalability

5 CROSS-MARKET DRUG TRAFFICKER IDENTIFICATION AND CASE STUDIES

To better understand and gain deeper insights into the ecosystem of drug trafficking in darknet markets, in this section, 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 including (1) manually read the product descriptions (e.g., sometimes the pair themselves admit in their posts about their other accounts), (2) examine the type of drugs they sell, and (3) check vendor reviews. Among these 798 detected cross-market pairs, 726 pairs (90.09%) are with high confidence that they are the same individuals and 22 pairs are uncertain (2.76%).



(a) Vendor "The***shop" on Evolution (b) Vendor "***Store" on Valhalla

Figure 8: An example of detected cross-market vendor pair.

As shown in Figure 8, for one of our detected cross-market vendor pairs, though "The***shop" on Evolution and "***Store" on Valhalla have different usernames, PGP keys and contact information, they sell similar drugs (i.e., heroin and MDMA) 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.

After detecting the cross-market vendor pairs using *uStyle-uID*, we conduct further analysis. Figure 9 shows the detected vendor pairs cross every two markets. From Figure 9, we can see that there's largest number of detected cross-market vendor pairs in Dream Market and Valhalla (i.e., 246 pairs); the reason behind is that Dream Market and Valhalla are still active and are among the most popular ones for illicit drug trading. This also demonstrates that it is common that a portion of vendors usually operate multiple accounts across different markets to increase their sales. We also find that many vendors would transfer their drug trafficking activities to other markets (by creating new accounts) after one marketplace was taken down. For example, after SilkRoad2 was shut down in November 2014, among 198 SilkRoad2-Evolution detected vendor paris, about 70% of the vendors created new accounts to conduct the illicit activities in Evolution (i.e., they registered in Evolution after November 2014).

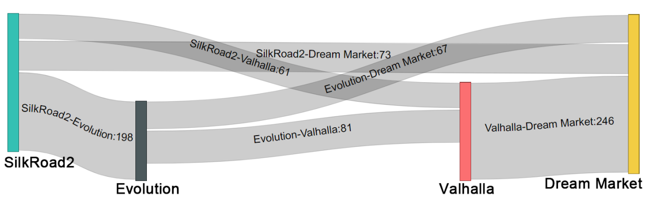


Figure 9: Cross-market drug trafficker pairs.

The above studies reveal that using our proposed methodology to perform the surveillance of darknet markets can be a valuable and supplementary way to facilitate the understanding of the behavioral processes of drug trafficking. The studies based on the detected cross-market vendor pairs using our developed system *uStyle-uID* also demonstrate that knowledge gained from darkweb data mining could facilitate the insight into online drug trafficking ecosystem, which can help devise effective interventions.

6 RELATED WORK

To combat drug trafficking in darknet makrets, there have been many research efforts on darknet market data analysis [5, 11, 20, 45]. Among these studies, there have been some methods proposed to tackle the challenges of authorship identification in darknet markets. These approaches mainly relied on either stylometry analysis or photography style analysis. For example, Ho et al. [20] built a framework using writing style analysis to link multiple accounts in different Darkweb forums; Wang et al. [45] presented an approach that constructed a series deep neural networks to model distinct photography styles for linking the multiple accounts of the same darknet vendors. Different from existing works, in this paper, besides vendor profiles and drug information, we propose to leverage both writing and photography styles together with their contents for drug trafficker identification.

In order to depict vendors, drugs, texts, photos and their associated attributes as well as the rich relationships among them, it is important to model them properly in order to facilitate the task of drug trafficker identification. Though HIN has shown the success of modeling different types of entities and relations (e.g., scientific publication network analysis [36], document analysis based on knowledge graph [44]), it it has limited capability of modeling additional attributes attached to entities. To address this challenge, we propose to use attributed HIN (AHIN) for representation. 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 [32], node2vec [18], and LINE [37]. Unfortunately, due to the heterogeneous properties of HIN, it's difficult to directly apply them for HIN representation learning. To tackle this challenge, metapath2vec [12], HIN2vec [13] have been proposed for HIN representation learning, which are based on meta-path scheme. However, these HIN embedding models are unable to deal with the attribute information associated with each entity; without the attribute information such as writing and photography styles, the generated path instances are of low quality and less useful to our application. To address this issue, we propose *Vendor2Vec* to learn the desirable node representations in AHIN which is capable to not only preserve both the semantics and structural correlations among different types of nodes but also leverages the attribute information attached on each node in AHIN.

7 CONCLUSION

To combat drug trafficking in darknet markets, 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 for drug trafficker identification at the first attempt. To depict vendors, drugs, texts, photos and their associated attributes (e.g., a *text* is associated with attributes of its *text content* and *writing style*, a *photo* is associated with attributes of its *photo content* and *photography style*) as well as the rich relationships among them, we present 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 nodes (i.e., vendors) in 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 transform the drug trafficker identification task into a link prediction problem and further present a learning model named *uIdentifier* to solve the problem. The promising experimental results on the collected datasets from four darknet markets (i.e., Valhalla, Dream Market, SilkRoad2 and Evolution) demonstrate that *uStyle-uID* which integrates our proposed method outperforms alternative approaches. This work also demonstrates that knowledge gained from darkweb data mining could facilitate the insight into drug trafficking ecosystem, which can help devise effective interventions.

REFERENCES

[1] Sami Abu-El-Haija, Bryan Perozzi, and Rami Al-Rfou. 2017. Learning edge representations via low-rank asymmetric projections. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1787–1796.

[2] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. Doppelgänger finder: Taking stylometry to the underground. In *Security and Privacy (SP), 2014 IEEE Symposium on*. IEEE, 212–226.

[3] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. DoppelgÃďnger finder: Taking stylometry to the underground. In *IEEE Symposium on Security and Privacy (SP)*.

[4] Léon Bottou. 1991. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes* 91, EC2 (1991).

[5] Julian Broséus, Damien Rhumorbarbe, Caroline Mireault, Vincent Ouellette, Frank Crispino, and David Décary-Hétu. 2016. Studying illicit drug trafficking on Darknet markets: structure and organisation from a Canadian perspective. *Forensic science international* 264 (2016), 7–14.

[6] Julia Buxton and Tim Bingham. 2015. The rise and challenge of dark net drug markets. *Policy Brief* 7 (2015).

[7] Prima Chairunnanda, Nam Pham, and Urs Hengartner. 2011. Privacy: Gone with the Typing! Identifying Web Users by Their Typing Patterns.. In *SocialCom/PASSAT*. 974–980.

[8] Richard Colbaugh and Kristin Glass. 2011. Proactive defense for evolving cyber threats. In *Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on*. IEEE, 125–130.

[9] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *ECCV*. Springer, 288–301.

[10] Hakan Demirbuken et al. 2011. The Global Afghan Opium Trade: A Threat Assessment. *UNODC Report* (2011), 3–4.

[11] Martin Dittus, Joss Wright, and Mark Graham. 2018. Platform Criminalism: The'last-mile'geography of the darknet market supply chain. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 277–286.

[12] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*. ACM, 135–144.

[13] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning. In *CIKM*. ACM, 1797–1806.

[14] Dario Garcia-Gasulla, Eduard Ayguadé, Jesús Labarta, Javier Béjar, Ulises Cortés, Toyotaro Suzumura, and R Chen. 2017. A visual embedding for the unsupervised extraction of abstract semantics. *Cognitive Systems Research* 42 (2017), 73–81.

[15] Michael Gilbert and Nabarun Dasgupta. 2017. Silicon to syringe: Cryptomarkets and disruptive innovation in opioid supply chains. *International Journal of Drug Policy* 46 (2017), 160–167.

[16] Thomas Gloe and Rainer Böhme. 2010. The'Dresden Image Database'for benchmarking digital image forensics. In *Proceedings of the 2010 ACM Symposium on Applied Computing*. 1584–1590.

[17] P. Griffiths and J. Mounteney. 2017. Disruptive potential of the internet to transform illicit drug markets and impact on future patterns of drug consumption. *Clinical Pharmacology & Therapeutics* 101, 2 (2017).

[18] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *KDD*. ACM, 855–864.

[19] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* 29, 6 (2012), 82–97.

[20] Thanh Nghia Ho and Wee Keong Ng. 2016. Application of stylometry to darkweb forum user identification. In *International Conference on Information and Communications Security*. Springer, 173–183.

[21] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. 2002. Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* 97, 460 (2002), 1090–1098.

[22] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[23] Dev Kar and Joseph Spanjers. 2017. Transnational crime and the developing world. (2017).

[24] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In *CVPR*, Vol. 1. IEEE, 419–426.

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

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

[27] Dream Market. 2018. *Dream Market*. http://6khhxwj7viwe5xjm.onion.

[28] James Martin. 2014. Drugs on the dark net: How cryptomarkets are transforming the global trade in illicit drugs. (2014).

[29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*. 3111–3119.

[32] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *KDD*. ACM, 701–710.

[33] Damien Rhumorbarbe, Ludovic Staehli, Julian Broseus, Quentin Rossy, and Pierre Esseiva. 2016. Buying drugs on a Darknet market: A better deal? Studying the online illicit drug market through the analysis of digital, physical and chemical data. *Forensic science international* 267 (2016), 173–182.

[34] SilkRoad. 2018. *SilkRoad*. http://silkroad7rn2puhj.onion.

[35] Martijn Spitters, Femke Klaver, Gijs Koot, and Mark van Staalduinen. 2015. Authorship analysis on dark marketplace forums. In *Intelligence and Security Informatics Conference (EISIC)*.

[36] 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. *VLDB Endowment* 4, 11 (2011), 992–1003.

[37] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *WWW*. International World Wide Web Conferences Steering Committee, 1067–1077.

[38] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 173–180.

[39] Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32, 5 (1998), 323–352.

[40] Valhalla. 2018. *Valhalla*. http://valhallaxmn3fydu.onion.

[41] Joe Van Buskirk, S Naicker, RB Bruno, C Breen, and A Roxburgh. 2016. Drugs and the Internet. (2016).

[42] Joe Van Buskirk, Amanda Roxburgh, Michael Farrell, and Lucy Burns. 2014. The closure of the Silk Road: What has this meant for online drug trading? *Addiction* 109, 4 (2014).

[43] Joe Van Buskirk, Amanda Roxburgh, Sundresan Naicker, and Lucinda Burns. 2015. A response to Dolliver's "Evaluating drug trafficking on the Tor network". *International Journal of Drug Policy* 26, 11 (2015).

[44] Chenguang Wang, Yangqiu Song, Haoran Li, and Jiawei Zhang. 2016. Text Classification with Heterogeneous Information Network Kernels. In *AAAI*. 2130–2136.

[45] Xiangwen Wang, Peng Peng, Chun Wang, and Gang Wang. 2018. You are your photographs: Detecting multiple identities of vendors in the darknet marketplaces. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. ACM, 431–442.

[46] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *TPAMI* 29, 1 (2007), 40–51.

[47] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*. 649–657.