Computational Approaches to Detect Illicit Drug Ads and Find Vendor Communities Within Social Media Platforms

Fengpan Zhao, Pavel Skums, Alex Zelikovsky, Eric L. Sevigny, Monica Haavisto Swahn, Sheryl M. Strasser, Yan Huang, and Yubao Wu

Abstract—The opioid abuse epidemic represents a major public health threat to global populations. The role social media may play in facilitating illicit drug trade is largely unknown due to limited research. However, it is known that social media use among adults in the US is widespread, there is vast capability for online promotion of illegal drugs with delayed or limited deterrence of such messaging, and further, general commercial sale applications provide safeguards for transactions; however, they do not discriminate between legal and illegal sale transactions. These characteristics of the social media environment present challenges to surveillance which is needed for advancing knowledge of online drug markets and the role they play in the drug abuse and overdose deaths. In this paper, we present a computational framework developed to automatically detect illicit drug ads and communities of vendors. The SVM- and CNN- based methods for detecting illicit drug ads, and a matrix factorization based method for discovering overlapping communities have been extensively validated on the large dataset collected from Google+, Flickr and Tumblr. Pilot test results demonstrate that our computational methods can effectively identify illicit drug ads and detect vendor-community with accuracy. These methods hold promise to advance scientific knowledge surrounding the role social media may play in perpetuating the drug abuse epidemic.

Index Terms—Opioid abuse epidemic, illicit drug ads, social media, text mining, deep learning, community detection

1 Introduction

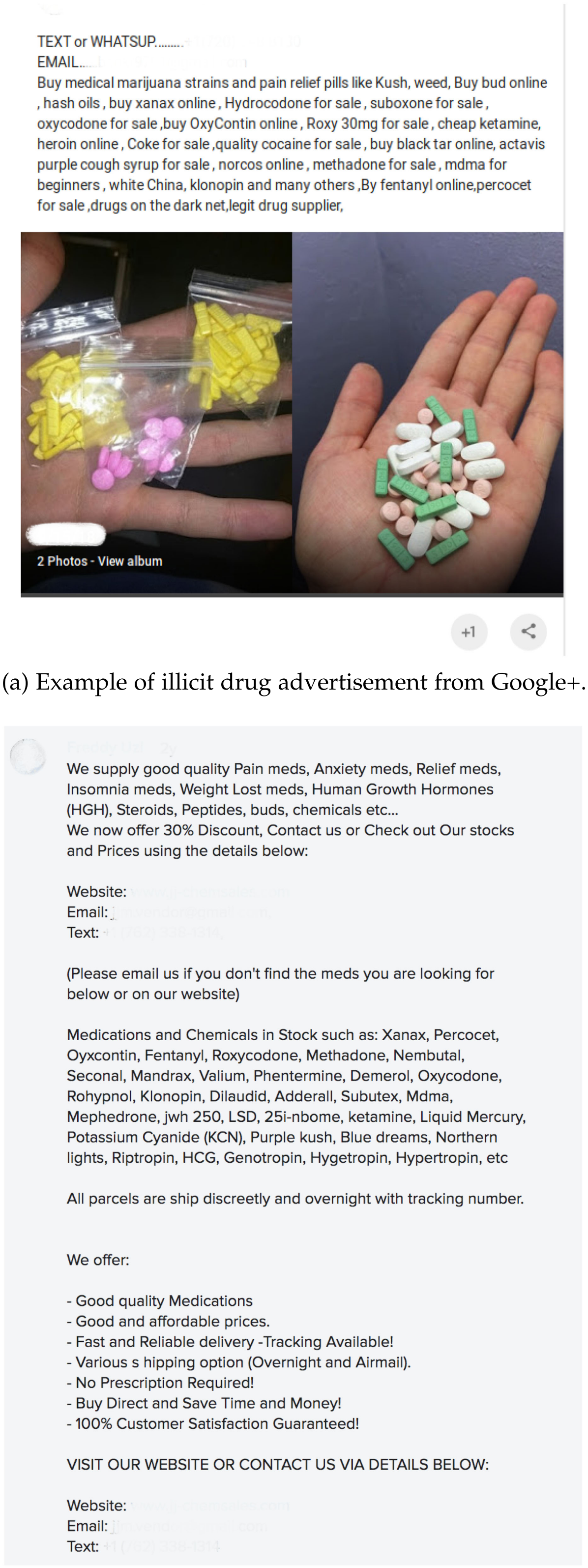
Drug overdose deaths have risen sharply over the past few years according to *Morbidity and Mortality Weekly Reports* from the Centers for Disease Control and Prevention (CDC) [47]. CDC recorded 70,237 drug overdose deaths in 2017, a 10 percent increase in drug overdose deaths from the prior year [47]. Drug overdose deaths represent an important public health priority area for prevention efforts.

One driver of the drug epidemic is enhanced drug accessibility and promotion that is facilitated via the Internet. Online drug trading platforms have emerged and are flourishing [39]. Social media platforms prove to be a popular venue whereby vendors can post illicit drug ads with relative ease, expansive reach, and with little cost. On the consumer side, drug users are able to search and find numerous vendors selling a wide range of illicit substances complete with drug information, user reviews, and encrypted web-based sale capabilities [13], [52]. Unlike traditional street drug transactions, eCommerce enables illicit drug vendors to connect and complete sales orders with drug users via social media, which bypasses direct personal contact entirely and essentially makes illicit drug sales as efficient and effective as any other online purchase. Results from a national survey on social media use from [41] indicate that an estimated 70 percent of U.S. adults ever use social media. In our research, we found that illicit drug ads are ubiquitous within most social media platforms. Figs. 1a and 1b show samples collected from Google+, Flickr. Many ads contain vendors' phone numbers, emails, Wickr IDs, and websites that enable the ability for consumers to communicate with drug vendors and initiate a drug order sale transaction. The order can be directly delivered to a specific pickup location without the disclosure of identifiable information. As such, effective and efficient surveillance tools in digital platforms are needed to monitor the activities of online drug sellers and prevent the increasing criminal use of the Internet to advertise illicit drugs.

Another surveillance avenue distinct from ads is to discover active vendors through detecting vendor communities. One confirmed active vendor would be helpful for law enforcement to identify multiple vendors linked by the same illicit substances automatically. It would also contribute to disrupting the growth of illicit drug markets [12]. However, discovering the communities of vendors is a challenging problem since there is no direct connection between vendors. Identifying meaningful communities in real-world networks has proven to be a challenging task [15], [56]. In our research, we observed that a number of vendors might sell an exclusive

- *F. Zhao, P. Skums, A. Zelikovsky, and Y. Wu are with the Department of Computer Science, Georgia State University, Atlanta, GA 30303 USA. E-mail: fzhao6@student.gsu.edu, {pskums, ywu28}@gsu.edu, alexz@cs.gsu.edu.*
- *E.L. Sevigny is with the Department of Criminal Justice and Criminology, Georgia State University, Atlanta, GA 30303 USA. E-mail: esevigny@gsu.edu.*
- *M.H. Swahn and S.M. Strasser are with the School of Public Health, Georgia State University, Atlanta, GA 30303 USA. E-mail: {mswahn, sstrasser}@gsu.edu.*
- *Y. Huang is with the Department of Software Engineering & Game Development, Kennesaw State University (KSU), Marietta, GA 30060 USA. E-mail: yhuang24@kennesaw.edu.*

Manuscript received 1 Nov. 2019; revised 16 Feb. 2020; accepted 25 Feb. 2020. Date of publication 5 Mar. 2020; date of current version 3 Feb. 2022. (Corresponding author: Fengpan Zhao.) Digital Object Identifier no. 10.1109/TCBB.2020.2978476



TEXT or WHATSUP........ [redacted]
EMAIL..... [redacted]
Buy medical marijuana strains and pain relief pills like Kush, weed, Buy bud online , hash oils , buy xanax online , Hydrocodone for sale , suboxone for sale , oxycodone for sale ,buy OxyContin online , Roxy 30mg for sale , cheap ketamine, heroin online , Coke for sale ,quality cocaine for sale , buy black tar online, actavis purple cough syrup for sale , norcos online , methadone for sale , mdma for beginners , white China, klonopin and many others ,By fentanyl online,percocet for sale ,drugs on the dark net,legit drug supplier,

(a) Example of illicit drug advertisement from Google+.

[redacted]

We supply good quality Pain meds, Anxiety meds, Relief meds, Insomnia meds, Weight Lost meds, Human Growth Hormones (HGH), Steroids, Peptides, buds, chemicals etc...
We now offer 30% Discount, Contact us or Check out Our stocks and Prices using the details below:

Website: [redacted]
Email: [redacted]
Text: [redacted]

(Please email us if you don't find the meds you are looking for below or on our website)

Medications and Chemicals in Stock such as: Xanax, Percocet, Oyxcontin, Fentanyl, Roxycodone, Methadone, Nembutal, Seconal, Mandrax, Valium, Phentermine, Demerol, Oxycodone, Rohypnol, Klonopin, Dilaudid, Adderall, Subutex, Mdma, Mephedrone, jwh 250, LSD, 25i-nbome, ketamine, Liquid Mercury, Potassium Cyanide (KCN), Purple kush, Blue dreams, Northern lights, Riptropin, HCG, Genotropin, Hygetropin, Hypertropin, etc

All parcels are ship discreetly and overnight with tracking number.

We offer:

- Good quality Medications
- Good and affordable prices.
- Fast and Reliable delivery -Tracking Available!
- Various s hipping option (Overnight and Airmail).
- No Prescription Required!
- Buy Direct and Save Time and Money!
- 100% Customer Satisfaction Guaranteed!

VISIT OUR WEBSITE OR CONTACT US VIA DETAILS BELOW:

Website: [redacted]
Email: [redacted]
Text: [redacted]

(b) Example of illicit drug advertisement from Flickr.

Fig. 1. Example of illicit drug advertisements.

substance, such as fentanyl or heroin only. Our identification of these vendors revealed their attempts to recruit distributors via social media posts. These observations led to the discovery of vendor communities. For our purposes, a community is classified as a group of nodes having more connections with the members in the same group than in the other groups or the remaining of the network [16]. There may exist active connections for vendors in the same community which may reveal patterns of their corresponding networks [35].

In this paper, we have two goals: (1) to describe computational approaches developed and deployed to identify illicit drug ads from social media platforms and (2) to describe a computational approach developed in response to the discovery of vendor communities(based on their shared drug categories). In particular, we select Google+, Flickr and Tumblr as the data source for illicit ad detection and vendor communities detection. For ad detection, we adopt two binary machine learning methods to detect the illicit drug ads from the text: the support vector machine (SVM) based method and the convolutional neural network (CNN) based method. The SVM based adopt term frequency-inverse document frequency (TF-IDF) for feature selection [50]. The CNN based method is from [28], which designed to focus on text mining. One advantage of the CNN based method is that the features can be automatically learned from the text data. For vendors' communities detection, there are three tasks. The first one is to identify the unique vendors across social platforms since vendors may have different accounts on different social platforms. In our paper, we assume the same phone number on different social platforms from the same vendor. The second task is to extract the drugs in the illicit drug ads. It is a big challenge to extract the drug names since the illicit drug posts are in an unstructured form. A long-short term memory-based method [23] is applied to extract drug names from unstructured posts to deal with this problem. The last task is to figure out the communities of vendors. We build a bipartite graph by using drug categories and vendors as two different types of nodes. The communities can be overlapped since one vendor can belong to two or more communities. Thus an overlapping community detection algorithm is needed. We adopt a matrix factorization based method [55] to detect the communities of vendors. A workflow, the whole procedure of our work for illicit drug ads detection and vendors community detection, is shown in Section 3.

2 Related Work

2.1 Online Illicit Drug Ads Detection

The illicit online drug trade has been the subject of several epidemiological and sociological studies. In particular, Mackey *et al.* [34] created a fictitious advertisement, offering consumers a way to buy drugs without a prescription. The advertisement was posted on four social media platforms: Facebook, Twitter, MySpace and Google+. Eventually, one of these accounts was blocked due to suspicious activity, but the remaining fake illicit drug advertisements were easily accessible during the duration of the experiment. A study conducted by Stroppa *et al.* [48] revealed that one-fifth of collected posts advertise counterfeit and/or illicit products online. Their research emphasized that the detection of illegal cyber-vendors and online tactics requires the development and application of sophisticated and tailored screening/detection methods.

2.2 Cyber-Vendor Community Detection

There are several studies for Community detection methods in the real bipartite graph. Isah *et al.* provided a bipartite graph model to infer hidden ties in crime data in 2015 [26]. Marin *et al.* Marin *et al.* proposed a novel model based on machine learning to find overlapped communities of malware vendors [35]. They used the categories of malware products and malware vendors to build the bipartite graph. They also provide ways to validate communities since there is no benchmark for the real-world network. For drug vendors, Duxbury *et al.* detect communities of opioid vendors on a darknet cryptomarket [12]. They characterized the network structure of vendors and buyers.

2.3 Computational Algorithms

2.3.1 Text Mining for Ads Detection

The development of tools for the detection of malicious and/or undesired advertisements in social media has been a subject of several studies. Hu *et al.* [21] provided a framework for the detection of spammers on microblogging. Zheng and colleagues [58] proposed an SVM-based machine learning model to detect spammer behavior on Sina Weibo. Agrawal *et al.* [3] introduced an unsupervised method called Reliability-based Stochastic Approach for Link-Structure Analysis, which can be used to detect topical posts on social media. Jain *et al.* [27] used convolutional and long short-term memory (LSTM) neural networks to detect spam in social media while addressing the challenges of text mining on short posts.

2.3.2 Named Entity Recognition

Supervised machine learning methods are popular for Named Entity Recognition (NER), such as Support Vector Machines, Conditional Random Fields [54]. Lafferty *et al.* first introduced CRFs in 2001 [29]. The Stanford Natural Language Processing Group proposed a 7 class model trained on the MUC 6 and MUC 7 training data sets [14] which can recognize Location, Person, Organization, Money, Percent, Date, Time in text data. Conditional Random Field (CRF) sequence models are used in Stanford NER software. Recently, Neural networks for NER become prevalent since the neural network does not need much feature selection [54]. Collobert *et al.* first provided a word-level Neural Network model [9] for NER. Moreover, deep learning models and supervised machine learning methods are combined for NER. Huang *et al.* proposed a bidirectional long-short term memory with conditional random fields which achieved a high F1-score on English CoNLL 2003 dataset [23].

2.4 Community Detection

Community detection is a central problem in machine learning and data mining since the 1980s. Depending on the properties from different networks, there are various algorithms to complete community detection tasks through identifying partitions of interacting nodes [55]. Infomap [45] and Louvain [7] are two state-of-the-art algorithms which on the basis of data mining techniques. Infomap is optimizing weighted modularity based on random walk dynamics [45]. Louvain, a greedy optimization technique of Newman-Girvan modularity [38], can unfold communities in large networks [7]. Both of OSLOM [30] and CNM [8] are based on machine learning algorithms. OSLOM searches for clusters by optimizing an important score locally [30]. It can deal with different types of graphs. CNM [8] is a fast hierarchical agglomeration algorithm for community detection in large scale network. It can cope with millions of nodes. Walktrap, a popular community detection method, is based on the combination of data mining and machine learning [42]. It is a hierarchical clustering algorithm by estimating the similarity between pairs of nodes based on random walk dynamics [42].

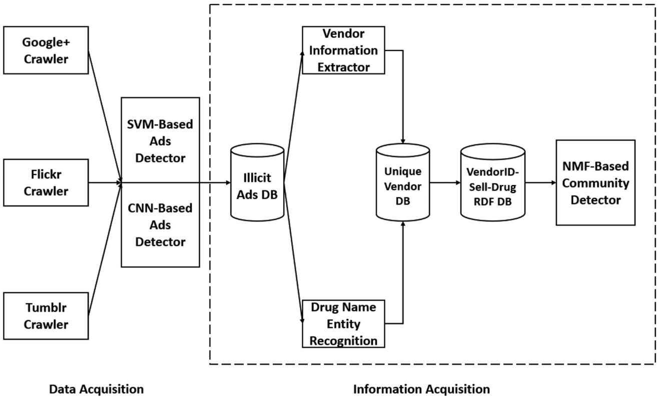


Fig. 2. Workflow for ads detection and community detection.

From the structural properties in our work, we focus on the studies for community detection in the bipartite graph. Barber proposed a model called Bipartite Recursively Induced Modules (BRIM) to detect communities in the bipartite network [5]. The goal of BRIM is finding out the optimal value of modules and then bringing out the community division. For example, using one red node starts from an arbitrary partition in red nodes and retrieves the partition of the blue nodes, which is in turn as input to find a better partition of red nodes, and iterating until modularity converges [31]. During the recursive process, the modularity Q is guaranteed to increase or at least to keep the previous modularity. Thus, the BRIM model will always find the divisions of communities at an optimal value of Q [31]. The BRIM method can also be used for determining the number of communities in the bipartite network [31]. Du *et al.* offered a novel algorithm called Bi-community Detector (BiTector) to discover the communities in the bipartite graph [10]. They use bicliques as the main ingredients to detect communities [11]. Yang and Leskovec proposed Cluster Affiliation Model for Big Networks (BigClam) for detecting communities [55]. This community detection method is based on Non-negative Matrix Factorization (NMF) to discover the affiliation weight between a node and a community.

In contrast to the previous studies, we specifically focus on the detection of illicit drug ads and discover communities of drug vendors within social media platforms, with the aim of applying epidemiological methods to investigate online enabling structures associated with opioid abuse.

3 Techniques

The workflow is shown in Fig. 2. It contains two main parts: data acquisition and information acquisition. The first part, data acquisition, is the procedure of retrieving data from digital platforms and detecting the illicit drug ads. This part includes three data crawlers: Google+ Crawler, Flickr Crawler and Tumblr Crawler, and two illicit ads detectors: SVM-based detector and CNN-based detector. The second part, the information acquisition, includes vendor information extractor, drug name recognition and vendor communities detector.

3.1 Illicit Ads Detector

Two machine learning methods are applied to classify the posts retrieved from social media: Support Vector Machine (SVM) based method and Convolutional Neural Network (CNN) based method. For both methods, the inputs are text data extracted from posts, and the outputs are the predicted labels indicating whether each post is an illicit drug ad.

3.1.1 Detecting Method 1: The SVM-Based Method

The SVM based method needs feature selection thus we have two stages: pre-processing and classification.

In pre-processing stage, text posts collected from social media are transformed into numerical feature vectors, which are further used as the inputs for the SVM classifier. It is a crucial part of traditional text mining methods because the selected features affect the performance of the classifier.

Pre-processing consists of three steps. In the first step, the stop words considered noise are removed. In the second step, the root of a word is isolated by removing tenses of verbs, which is also referred to as stemming [24]. In the third step, the term frequency-inverse document frequency (TF-IDF) features are determined [46]. The TF-IDF is the product of two statistics: term-frequency and inverse document frequency. The term frequency is calculated based on the raw count of a term (word). The inverse document frequency is a measure of how much information the word provides. For a term(word) t in a document d, the weight is calculated as [44]

$$w(t,d) = tf_{t,d} \cdot \log \frac{N}{df_i}, \tag{1}$$

where $tf_{t,d}$ is the number of occurrences of word t in document d, df_i is the total number of documents which contain the word t and N is the total number of documents. TF-IDF features computed at the pre-processing step are used to train an SVM model that can be further used to predict labels of new posts. SVM is a classical supervised learning method, which constructs a hyperplane in a multidimensional euclidean space to serve as a separator for feature vectors from two classes. We used the radial basis function (RBF) kernel SVM classifier, whose accuracy was assessed using a ten-fold cross-validation process on a labeled post text dataset manually curated by human experts.

3.1.2 Detecting Method 2: The CNN-Based Method

This method uses the TextCNN. This model can detect illicit drugs ads on social medias, identify the unique vendors across multiple social platforms and match the drugs and vendors automatically [28], which first computes a word embedding and then applies the convolutional neural networks (CNN) to perform the classification. TextCNN does not require the removal of stop words or stemming.

Word embedding which maps words or phrases to numerical vectors, was utilized to allow neural networks to process the text data. We used Word2vec, a commonly used word embedding model [36] that relies on the combination of skip-grams and continuous bag-of-words (CBOW) procedures [37]. CBOW generates a word based on the context, while skip-grams generates the context from a word. For example, if we treat {"Washington D.C.", "is", "the United States"} as a context, then CBOW will generate the word "capital". If given the word "capital", skip-grams will be able to predict the following words: 'Washington D.C.", "is", "the United States". The numerical vectors generated by word2vec are used as the input of CNN.

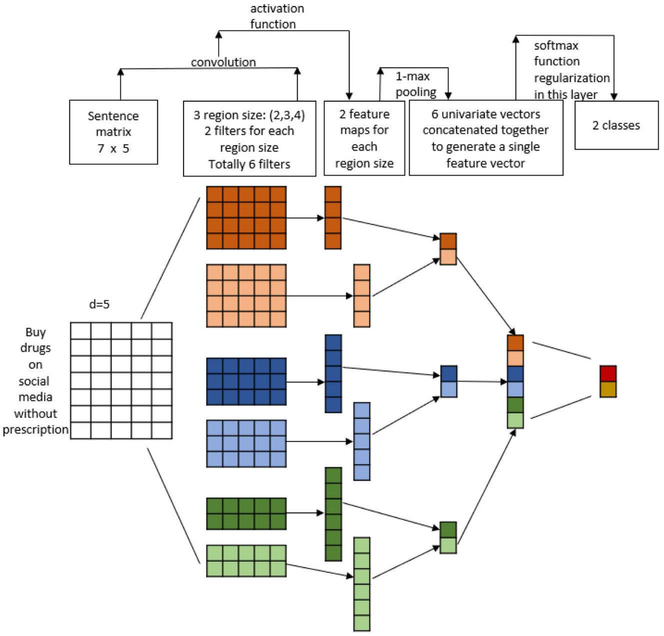


Fig. 3. Illustration of TextCNN.

TextCNN contains a single layer of neural net, which allows it to be highly scalable yet sensitive in performing text classification. Fig. 3 shows the general scheme of TextCNN [57]. Let d be the dimension of word vector. Given a sentence "Buy drugs on social media without prescription" and $d = 5$, we can generate a sentence matrix in Fig. 3. Then feature maps are generated by filters operating convolutions on the sentence matrix. Here we set the region sizes to 2, 3 and 4, and each region size has two filters. A max-pooling operation is applied to the feature map to retrieve the largest number. Therefore we can take six features from six feature maps and concatenate them together to get a feature vector which will serve as the input of the softmax layer. Finally, we complete a binary classification by using this feature vector through softmax layer.

3.2 Drug Named Entity Recognition: The LSTM-Based Method

Named entity recognition(NER) is a problem of finding named entities in a given text [54]. NER can recognize people, location, time, organization, numbers, and also support finding customized items [54]. Accurately identifying the drug names in detected illicit drug ads is a crucial part since the recognized drugs will be used to build a graph in our work and thus it will directly affect the results of community detection.

Bidirectional long-short term memory with conditional random fields(BiLSTM-CRF) [23] is one of the most widely

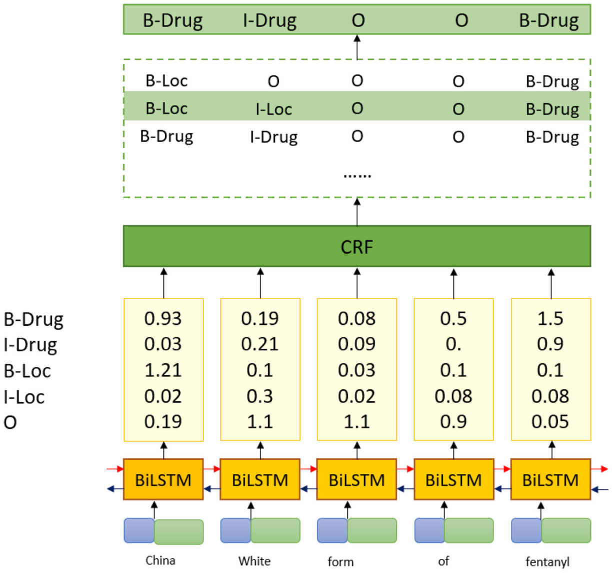


Fig. 4. Illustration of BiLSTM-CRF.

used methods for NER. In our paper, we will apply BiLSTM-CRF to complete the task of drug named entities recognition. BiLSTM-CRF contains two parts: bidirectional Long-Short Term Memory(BiLSTM) networks and Conditional Random Fields(CRF). Fig. 4 shows the structure of BiLSTM with CRF [23]. For BiLSTM, it connects two LSTM layers where the two layers have two opposite directions: forward and backward. Generally, the forward direction can process the text from the beginning to the end while the other one works from the end to the beginning. Both directions will give the results to the same output layer. In this way, the BiLSTM obtains merged hidden states from both previous and next states. For CRFs, they are on the top of BiLSTM and the inputs of CRFs are the hidden states from BiLSTM. CRFs are a type of discriminative classifier [29] and they assign the probability to a label sequence is by [51]

$$P_\lambda(Y|X) = \frac{1}{Z(X)} exp\left(\sum_{c\in C}\sum_{k}\lambda_k f_k(Y_c, X, c)\right), \tag{2}$$

where X is unsegmented characters sequence, Y is the label sequence, $Z(X)$ is a term for normalization, f_k is a feature function, and c is index for the characters which being labeled [51]. The CRFs consider the final sequence in sentence-level, which makes the prediction more reasonable than without CRFs. For example, the result in Fig. 4 before CRFs is "B-Loc O O O B-Drug". In CRFs, the whole sentence is considered, and then the result is changed to "B-DRUG I-DRUG O O B-DRUG". The "B-Drug" means the beginning of a drug name and "I-Drug" indicates a word inside a drug name. The "B-Drug" means the beginning of a location and "I-Loc" indicates a word inside a location. O stands for others.

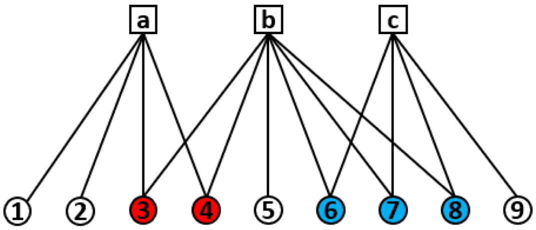


Fig. 5. Example of overlapping communities. The letters in the square stand for communities, and the numbers in circles represent of vendors. The vendor 3 and 4 in red circles could be in community a or community b or both; the number 6, 7 and 8 in blue circles could be in community b or community c or both.

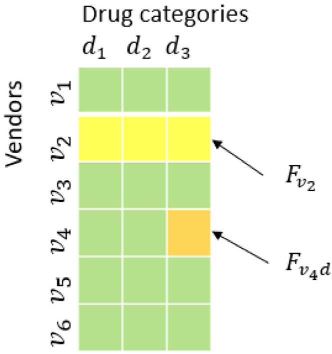


Fig. 6. Explanation of F. F_{v_2} is a vector which store the strengths of node v_2 to all communities; let d_1 is a community, $F_{v_2 d_1}$ is the strength of node v_2 to community d_1.

3.3 Community Detection

To deduce the relationship of vendors which is not possible from direct observations, we apply community detection methods to partition the vendors into disjoint communities. In our dataset, we build a bipartite graph since drugs and vendors are two different types. We use Resource Description Framework(RDF) triple stores [49] to data in the following format: "VendorId-sell-DrugName". In a bipartite network, drugs form drug communities, and vendors form vendor communities. A vendor community is composed of a group of vendors where each pair of them should have a high possibility of selling the same drugs. Meanwhile, there are overlappings in the real network, which means one vendor could belong to two or more groups. Fig. 5 gives an example of overlapping communities.

An Nonnegative Matrix factorization(NMF) [32], the Cluster Affiliation Model for Big Networks(BigClam) [55], is applied to detect the communities of vendors. BigClam is an algorithm focus on discovering overlapping communities in large networks.

BigClam use a nonnegative weight F_{uc} to represent the weight between u and c, where u is a node from node vector **V** and c is a community from community vector **C** [55]. F_{uc} indicates the strength of affiliation for node u to c. Particularly, given F_{uc} = 0, we consider u does not belong to c. The weight vector $\boldsymbol{F_u}$ represent the affiliation weights for node u to communities in community vector $\boldsymbol{C}$. Fig. 6 gives an example of matrix $\boldsymbol{F}$. BigClam provide an equation to show the probability of creating edge (u, v) which connects node u and node v

$$P(u, v) = 1 - exp(-\boldsymbol{F_u} \cdot \boldsymbol{F_v^T}), \tag{3}$$

where $P(u, v)$ will be 0 if $\boldsymbol{F_u}$ = **0** or $\boldsymbol{F_v}$ = **0** or both equal to 0. Higher $P(u, v)$ means there is a higher possibility that exists a connection between node u and v, which further indicates node u and v have higher possibility in one community.

An Nonnegative Matrix factorization(NMF) [32], the Cluster Affiliation Model for Big Networks(BigClam) [55], is applied to detect the communities of vendors. BigClam is an algorithm focus on discovering overlapping communities in large networks.

BigClam use a nonnegative weight F_{uc} to represent the weight between u and c, where u is a node from node vector $\mathbf{V}$ and c is a community from community vector $\mathbf{C}$ [55]. F_{uc} indicates the strength of affiliation for node u to c. Particularly, given $F_{uc} = 0$, we consider u does not belong to c. The weight vector $\boldsymbol{F_u}$ represent the affiliation weights for node u to communities in community vector C. Fig. 6 gives an example of matrix F. BigClam provide an equation to show the probability of creating edge (u, v) which connects node u and node v:

Yang and Leskovec spot that F is the best approximates the adjacency matrix A of graph G [55]. Therefore computing F can be solved by using matrix factorization method. Nonnegative Matrix factorization(NMF) [32] is a widely used method for matrix factorization. NMF decompose a given matrix V into two matrix W and H which let the matrix V have the approximation $V \simeq W \times H$. The W and H could be computed by minimizing the error function

$$\arg\min \|\boldsymbol{V} - \boldsymbol{W} \times \boldsymbol{H}\|, \tag{4}$$

where $\boldsymbol{W}, \boldsymbol{H} \geq \mathbf{0}$. Meanwhile, the optimization of F can be achieved by

$$\hat{\boldsymbol{F}} = \arg\max_{\boldsymbol{F} \geq \mathbf{0}} l(\boldsymbol{F}), \tag{5}$$

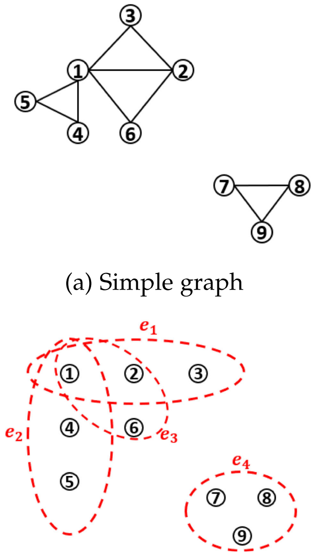
where

$$l(\boldsymbol{F}) = \sum_{(u,v)\in \boldsymbol{E}} log(1 - exp(-\boldsymbol{F_u F_v})) - \sum_{(u,v)\,\notin\, E} \boldsymbol{F_u F_v^T}. \tag{6}$$

The Eq. (5) can be viewed as nonnegative matrix factorization (NMF) [32]. If using a loss function D to replace the negative log-likelihood $-l(\boldsymbol{F})$, we have

$$\hat{\boldsymbol{F}} = \arg\min D(\boldsymbol{A}, f(\boldsymbol{F}\boldsymbol{F}^{\boldsymbol{T}})), \tag{7}$$

where D is loss function [55]. BigClam picks log-likelihood as a loss function [4] to complete NMF. F will be iteratively updated until achieving the maximum likelihood in a gradient ascent way. In 2017, Liu and Chamberlain provided a parallel way across multiple threads to speed up BigClam, which achieved 2.5 times faster than unparallelized BigClam when solving the Amazon product co-purchase network [33].



(a) Simple graph

(b) Hypergraph

Fig. 7. Simple graph vs. hypergraph. Fig. 7a shows an undirected graph where an edge connects two vendors. This graph cannot tell us how many communities are shared for two vendors. Fig. 7b shows a hypergraph clarify the complex relationships among vendors and communities.

3.4 Hypergraph as Analysis Tool

A hypergraph is a generalization of a graph, where an edge, also named hyperedge, can connect any number of nodes [6]. Comparing to a simple graph, hypergraph can clarify the complex relationships among vendors and the communities of vendors [59]. For example, suppose we obtained affiliation weight matrix F calculated by BigClam is

$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where the columns are communities $\{e_1, e_2, e_3, e_4\}$ and the rows are vendors $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Fig. 7 shows the simple graph and the hypergraph. The nodes in a simple graph are connected by an edge if there is at least one shared community. Due to one edge that can only connect two nodes in a simple graph, we miss the information on whether one vendor belongs to three or more communities. To avoid information loss, we use hypergraph to illustrate complicated relationships.

4 Experimental Results

In this section, we will show the results and evaluation of the methods. Most of the tools have been implemented in Python 3.7, and run on a DELL workstation with Intel Xeon E5-1603 2.80 GHz CPU, 32 G memory, and Ubuntu 18.04 OS. We also use MySQL and Neo4j to store data.

4.1 Data Collection

We collected data from three platforms: Google+, Flickr and Tumblr. Google+ and Flickr provide API to collect historic data. For Tumblr, we use Selenium to crawl the data [19]. The analyzed dataset has been formed by posts containing at least one of the following 30 keywords [53]:

opioid, alprazolam, amphetamine, antidepressant, benzodiazepine, buprenorphine, cocaine, diazepam, fentanyl, heroin, hydrocodone, meth, methadone, morphine, naloxone,

TABLE 1
Accuracy of the SVM Based Method and TextCNN

Methods	Pre	Recall	F-score
SVM Based Method	0.65	0.81	0.72
TextCNN	0.97	0.90	0.93

TABLE 2
Running Time of the SVM Based Method and TextCNN

Methods	Training time	Testing time
SVM Based Method	2,469s	0.023s
TextCNN	3,936 s/epoch, 10 epoch	0.034s

narcan, opana, opiate, overdose, oxycodone, oxymorphone, percocet, suboxone, subutex, pill, rehab, sober, withdrawal, shooting up, track marks

In total, 433,673 posts published from 2017/01/01 to 2017/12/31 have been collected. We collected data in 2017 since we will further analyze the relationship between drug involved death in each state and illicit drug ads in each state in future work. We labeled all the posts manually. The following examples illustrate examples of illicit drug ads from the dataset. Ads 1-3 are selling illicit drugs while ad 4 is a normal post.

1) Buy Diazepam Online etc. Our products stand the best in the market in terms of quality and pricing. We do sell at cheap prices and shipping is done expressly overnight in USA and Canada upon discrete packaging. We also have very good wholesale prices and regularly ship all over Europe and Australia. We also carry a wide range of scientific research chemicals which are labelled not for human consumption. Payments and other transactions are done securely with our company so clients are guaranteed 100 percent satisfaction. Also we provide a free catalog which contains relevant information concerning the over 100 products we sell thus we invite everyone to get a free copy of this catalog and learn more about our products.
2) Buy Tramadol Ultram online without Doctors prescription. Tramadol Pills Tramadol (Ultram) relieves pain by modifying your brains response towards what would otherwise be painful sensations. As a prescribed drug, you need a doctor's prescription to buy Tramadol Online, which is also available under different brand names such as OL-Tram and Ultram. We provide a Tramadol 50 mg, 150 mg. Buy Tramadol online, Buy Tramadol, Tramadol Uses; Side Effect, Buy Tramadol overnight without prescription, Buy Tramadol Medicine Online, Tramadol online, Order Tramadol, Generic Tramadol, Cheap Tramadol, Buy Tramadol 50 mg, Buy Tramadol 100 mg, Cheap Tramadol buy usa, Tramadol overnight US.
3) Hello, I am a vendor in high quality pharmaceutical products like Xanax, Oxycodone, Fentanyl patch, Viagra, Diazapam, Percoset, Opana, Methadone, etc. and also high quality medical marijuana strains like Og kush, Sativa, Kief,S hatter, Girls Scott, Lemon haze, Moon rock, Afghan kush, Purple haze etc., my packaging is very safe and discreet, also my delivery is 100 percent assured as we do refund or resend the same order immediately in case of any unforeseen.
4) Highlighting concerns with the pharmaceutical supply chain, the Food and Drug Administration warned McKesson, one of the nations largest wholesalers, for failing to properly handle episodes where pharmacies received tampered medicines, including three ...

   FDA scolds McKesson for naproxen in tampered oxycodone bottles -STAT-

4.2 Ads Detecting Evaluation

4.2.1 Effectiveness Evaluation

We use precision, recall and F-score as metrics to evaluate the accuracy of the classification methods [43]. Precision is defined as the ratio of predicted and ground-truth illicit ads among all predicted illicit ads, i.e., $\mathrm{Prec} = \mathrm{tp}/(\mathrm{tp} + \mathrm{fp})$. Recall is defined as the ratio of predicted and ground-truth illicit ads among all ground-truth illicit ads, i.e., $\mathrm{Recall} = \mathrm{tp}/(\mathrm{tp} + \mathrm{fn})$. The F-score is the harmonic mean of precision and recall: $\text{F-score} = 2 \cdot \mathrm{Prec} \cdot \mathrm{Recall}/(\mathrm{Prec} + \mathrm{Recall})$. We use 10-fold cross-validation procedures to evaluate the accuracy of both the SVM and CNN based methods.

In TextCNN, we set the parameters as follows: max_sequence_length 20, embedding_dim 200, validation_split 0.16, test_split 0.2 [57]. Table 1 shows the precision, recall, and F-score for SVM and TextCNN. From Table 1, we can see that TextCNN outperforms SVM in all metrics.

4.2.2 Efficiency Evaluation

Table 2 shows the running time. In Table 2, the training time represents the average running times for training ten SVM or CNN models during the ten-fold cross-validation procedure. The number of posts in the input dataset for training each model is 390,305, which is 90 percent of the total of 433,673 posts. The testing time represents the average running time of predicting the label of a single post. In each iteration of the ten-fold cross-validation, the input number of posts is 43,367 posts. We measure the average time for each post. From Table 2, we can see that the SVM based method takes less than 1 hour while the TextCNN method takes 11 hours for training. Both ways take less than 0.05 seconds for prediction.

4.3 Information Extraction From Detected Illicit Ads

4.3.1 Information Extracted Statistics

Table 3 shows the extracted data statistics. We obtained 514 unique vendors, even though we have 52,832 detected vendors. We discovered 159 drug names, although we used 30 keywords to crawl the online data.

4.3.2 Vendor Information Extraction

Finding unique users from various social platforms is a challenge problem [25]. However, there is a significant fact in our dataset: vendors have to leave their contact information to make sure customers can reach them and thus they post phone numbers or email addresses in posts. We assume the unique vendor has the same contact information and thus

TABLE 3
Data Statistics

Data type	Data size
Original Posts	433,673
Detected illicit drug ads	84,332
Detected vendors	52,832
Filtered Unique vendors	514
Recognized drug names	159

unique vendor will be detected although most of them use different account names across social media platforms, we select the contact information to match users. The contact information includes phone number, email address, WhatsApp ID, Wickr ID, etc.

We choose Regular Expression(RE) to extract contact information from detected illicit ads. Several regular expression rules are generated to cover most cases(cannot cover all cases) since the text data are unstructured. For example, the phone number in posts are $404-123-4567$, $(404)123456$, $+1404123-4567$, etc. We found 52,832 vendors and filtered 514 unique vendors based on our assumption.

4.3.3 Drug Name Entity Recognition for Drugs

To recognize the drug names, we first built customized dictionaries to train the model. The dictionaries were downloaded from the U.S. Food and Drug Administration (FDA) [2] and drug slang term list from [1]. These files are all open-source. Then we applied BiLSTM-CRF in the detected illicit drug ads. The input of BiLSTM-CRF is the vectors of words, and the output is the predicted labels for each word.

For efficiency, we selected 75,899 illicit ads(90 percent of all illicit drug ads) as our training data and the running time of training model is 865s per epoch with ten epoch; we apply the model on 8,433 drug ads which are the 10 percent of all illicit drug ads, and each ad cost 0.026s. It implies that our trained model can efficiently recognize the drug entities.

To evaluate the effectiveness, for each post on which the model is tested, we calculate the precision, recall and F-score for each entity that the model recognizes, which are 0.92, 0.89 and 0.90, respectively. It indicates that our trained model can effectively identify the drug entities.

4.4 Community Detection

We built a bipartite graph to detect the communities of vendors by using the extracted information. In this section, we first visualize a sample graph to show the relationship between vendors and drugs, and then we apply the community detection method in the bipartite graph. Finally, we discuss the detected communities of vendors.

4.4.1 Bipartite Graph

We use Resource Description Framework(RDF) triple stores [49] to store the relationship between vendor and drug in the following format: "VendorID-sell-DrugName" based on the fact that each illicit drug post contains one vendor and multiple drug names. Neo4j is used to store the data. Neo4j is a graph database and can easily display the connectivities and relationships of data [20]. Fig. 8 shows a sample graph generated by Neo4j.

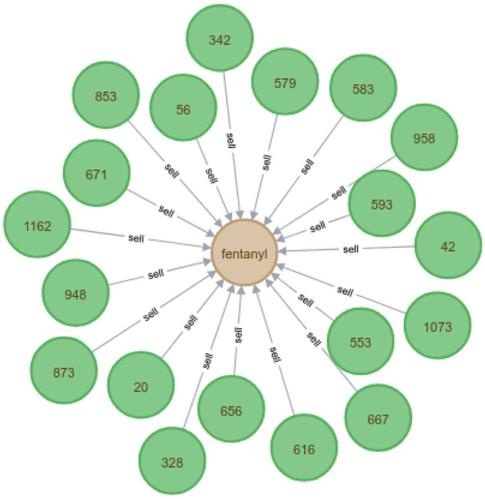


Fig. 8. A sample graph shows the vendors selling fentanyl in Google+, Flickr, and Tumblr. Green circles stand for the vendor, and the number of 740 is the vendor ID; Brown circle in the center stands for the drug (the number of records is limited to 25). This case is from a real illicit drug ad.

4.4.2 Community Detection Method Evaluation

We implemented BigClam by importing Networkx package [18]. We put the drugs into nine drug categories, according to [40]. The BigClam starts with an initial matrix $\boldsymbol{F}$ and then iteratively updates $\boldsymbol{F}$ to maximize the likelihood. We initialized matrix $\boldsymbol{F}$ by the following formula [33]:

$$F_{u'(N(u))} = \begin{cases} 1 & \text{if } u' \in N(u) \text{ and} \\ & N(u) \text{is a locally} \\ & \text{minimal neighbor}, \\ & \text{hood [17] of } u \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $N(u)$ stands for u and its neighbors. We also implemented another two state-of-the-art methods for overlapping community detection in bipartite graph: Brim [5] and BiTector [10]. For Brim, the community result can be found at an optimal value of Q [31]. For BiTector, it uses bi-cliques as the core ingredients to divide communities [11]. Brim and BiTector are also implemented in Python.

The challenge in our experiment is that there is no ground truth to evaluate our result since the network is formed from real social platforms. We first select modularity to evaluate the structure of detected communities in a computational way. Then we built a hypergraph to illustrate how it is used for law enforcement.

Modularity is a significant measure to evaluate the structure of networks. Newman proposed the definition of modularity in 2004 [38]

$$Q = \sum_{i=1}^{k} (e_{ij} - a_i^2), \tag{9}$$

where a_i is the possibility that a random edge can belong to community i, e_{ij} represent the possibility that an edge with two end nodes is in community i and j, respectively. Furthermore, e_{ii} represents the possibility that the edge with

TABLE 4
Modular Values for Community Detection

Methods	Modular Value
Unparallelized BigClam	0.6292
BRIM	0.5831
BiTector	0.4376

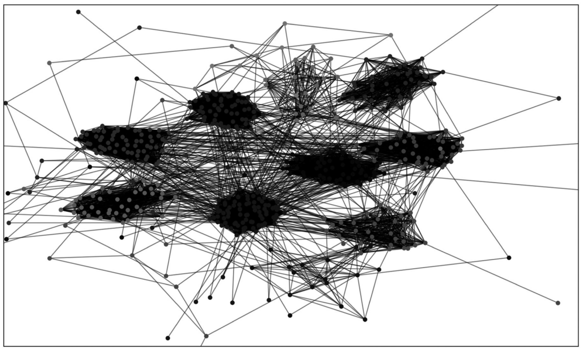


Fig. 9. Detected communities of vendors. We spotted nine main communities.

two end nodes is all in community i. Higher modularity means a better structure of the networks. Table 4 shows the results of the three state-of-the-art methods in our dataset. We can see unparallelized BigClam has the highest modular value. We also spot that the corresponding modular values from Brim and BiTector is lower than the result from BigClam.

4.4.3 Community Detection Result Discussion

In our work, we generated edges to connect vendors according to the affiliation weight matrix F calculated by BigClam. By using a vector V to store the vendors and a vector E to store the edges, we can draw a graph $G = (V, E)$ in Fig. 9. From this graph G, we spotted that there are nine main groups. However, this graph G cannot reveal the overlapping communities of vendors. We further built a hypergraph $H = (C, V)$ by using a vector C to store the communities(also known as hyperedges in hypergraph).

The problem we want to figure out from hypergraph H is how the communities are overlapped. We draw a degree distribution in the hypergraph to answer this question. The degree of a vendor in hypergraph can be considered as the number of shared communities. The degree of node v in hypergraph is defined as [22] $d(v) = \sum_{v \in V, c \in \boldsymbol{C}} w(c)$, where $w(c)$ is the weight of hyperedge c. In our hypergraph, for any hyperedge $c \in \boldsymbol{C}$, $w(c)$ is 1.0. Fig. 10 shows degree frequency histogram of hypergraph $\boldsymbol{H}$. We can see there are more than 200 vendors shared 2 communities, more than 80 vendors shared 3 communities. It indicates that more than half of the vendors are shared two or more communities. The maximum number of shared communities is 8, and the frequency is 2, which is also the smallest frequency. It indicates that the two vendors are active in eight groups. This result shows that our work would contribute to discovering the high influence vendors and thus disrupting the growth of illicit drug markets for law enforcement.

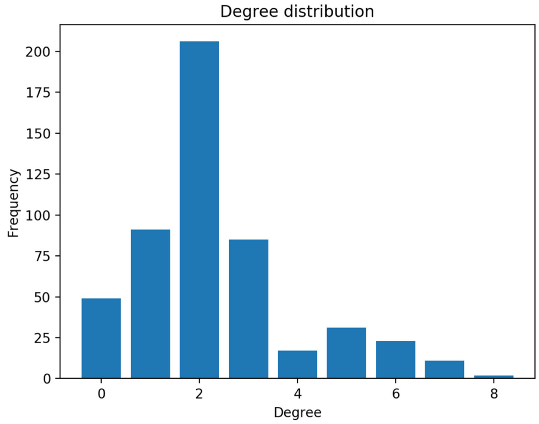
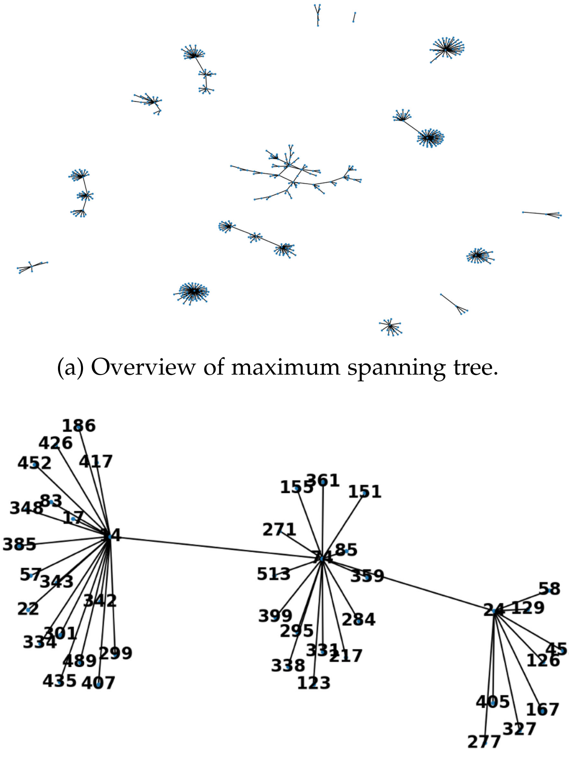


Fig. 10. Degree distribution from hypergraph.



(a) Overview of maximum spanning tree.

(b) A case from attributed maximum spanning tree.

Fig. 11. Maximum spanning tree from hypergraph.

Another problem we want to know is how the communities associate with drug types. To figure out this, we built a maximum spanning tree by generating weight between vendors. The weight is the number of shared communities between two vendors. For example, the weight between node 1 and node 2 in Fig. 7a is 2 since they share two communities: c_1 and c_2. According to these weights, we draw a maximum spanning tree in Fig. 11a. We observed separated groups in this maximum spanning tree since some communities are not sharing the same drug. We checked this MST by analyzing the drug types sold in each community. We found that for any isolated group, more than 86 percent vendors are selling the same kind of drugs. In Fig. 11b, we selected 43 vendors by zooming in Fig. 11a and found that there are 93 percent vendors sell drugs in the category of fentanyl.

5 Conclusion

Social media platforms offer a vast frontier for illicit drug e-Commerce to thrive and may represent an important enabler of the current opioid epidemic. Thus tools for monitoring and analysis of online drug markets are needed to advance epidemiological studies of illicit drug trading, which play a critical role in informing intervention and prevention applications. In this paper, we demonstrate that machine-learning-based methods which can efficiently capture illicit drug advertisements and detect communities of vendors. These tools hold vast potential for health care practitioners, law enforcement officials, and researchers to extract and analyze valuable data related to the opioid abuse epidemic, which can be examined to better understand the dynamics of online drug markets, trade, and vendor behaviors. These insights are essential in the development of tailored recommendations and public health intervention strategies that align with the social media environment.

References

[1] Drug Enforcement Administration, "Slang terms and code words: A reference for law enforcement personnel," 2018. Accessed: Jul. 21, 2019. [Online]. Available: https://publicintelligence.net/dea-drug-slang-code-words-2018/

[2] U.S. Food & Drug Administration, "Drugs@fda data files," 2019. Accessed: Jul. 17, 2019. [Online]. Available: https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files

[3] M. Agrawal and R. L. Velusamy, "R-SALSA: A spam filtering technique for social networking sites," in *Proc. IEEE Students' Conf. Elect. Electron. Comput. Sci.*, 2016, pp. 1–7.

[4] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. Berlin, Germany: Springer, 1998, pp. 199–213.

[5] M. J. Barber, "Modularity and community detection in bipartite networks," *Phys. Rev. E*, vol. 76, no. 6, 2007, Art. no. 066102.

[6] C. Berge, *Hypergraphs: Combinatorics of Finite Sets*, vol. 45. Amsterdam, The Netherlands: Elsevier, 1984.

[7] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Statist. Mech.: Theory Experiment*, vol. 2008, no. 10, 2008, Art. no. P10008.

[8] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, no. 6, 2004, Art. no. 066111.

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, no. Aug., pp. 2493–2537, 2011.

[10] N. Du, B. Wang, B. Wu, and Y. Wang, "Overlapping community detection in bipartite networks," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, 2008, pp. 176–179.

[11] D. Duan, Y. Li, R. Li, and Z. Lu, "Incremental K-clique clustering in dynamic social networks," *Artif. Intell. Rev.*, vol. 38, no. 2, pp. 129–147, 2012.

[12] S. W. Duxbury and D. L. Haynie, "The network structure of Opioid distribution on a darknet cryptomarket," *J. Quantitative Criminol.*, vol. 34, no. 4, pp. 921–941, 2018.

[13] Flash Eurobarometer, "401, 2014. Young people and drugs Report. Conducted by TNS Political & Social at the request of the European Commission, Directorate-General for Justice," 2014. Accessed: Jul. 20, 2019.

[14] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics*, 2005, pp. 363–370.

[15] S. Fortunato and M. Barthelemy, "Resolution limit in community detection," *Proc. Nat. Academy Sci. USA*, vol. 104, no. 1, pp. 36–41, 2007.

[16] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Academy Sci. USA*, vol. 99, no. 12, pp. 7821–7826, 2002.

[17] D. F. Gleich and C. Seshadhri, "Vertex neighborhoods, low conductance cuts, and good seeds for local community methods," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2012, pp. 597–605.

[18] A. Hagberg, P. Swart, and D. S. Chult, "Exploring network structure, dynamics, and function using networkx," in *Proc. 7th Python Sci. Conf.*, 2008, pp. 11–15.

[19] A. Holmes and M. Kellogg, "Automating functional tests using selenium," in *Proc. AGILE*, 2006, pp. 6-275.

[20] F. Holzschuher and R. Peinl, "Performance of graph query languages: Comparison of cypher, gremlin and native access in Neo4j," in *Proc. Joint EDBT/ICDT Workshops*, 2013, pp. 195–204.

[21] X. Hu, J. Tang, Y. Zhang, and H. Liu, "Social spammer detection in microblogging," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2633–2639.

[22] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang, "Learning hypergraph-regularized attribute predictors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 409–417.

[23] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*, [Online]. Available: http://arxiv.org/abs/1508.01991

[24] D. A. Hull, "Stemming algorithms: A case study for detailed evaluation," *J. Amer. Soc. Inf. Sci.*, vol. 47, no. 1, pp. 70–84, 1996.

[25] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 522–525.

[26] H. Isah, D. Neagu, and P. Trundle, "Bipartite network model for inferring hidden ties in crime data," in *Proc. IEEE/ACM Int. Conf. Advances Social Netw. Anal. Mining*, 2015, pp. 994–1001.

[27] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," *Ann. Math. Artif. Intell.*, vol. 85, no. 1, pp. 21–44, 2019.

[28] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: https://doi.org/10.3115/v1/d14-1181

[29] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[30] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," *PloS One*, vol. 6, no. 4, 2011, Art. no. e18961.

[31] K. Li and Y. Pang, "A unified community detection algorithm in complex network," *Neurocomputing*, vol. 130, pp. 36–43, 2014.

[32] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, 2007.

[33] C. H. Liu and B. P. Chamberlain, "Speeding up BigClam implementation on snap," 2017, *arXiv: 1712.01209*. [Online]. Available: https://doi.org/10.4230/OASIcs.ICCSW.2018.1

[34] T. K. Mackey and B. A. Liang, "Global reach of direct-to-consumer advertising using social media for illicit online drug sales," *J. Med. Internet Res.*, vol. 15, no. 5, 2013, Art. no. e105.

[35] E. Marin, M. Almukaynizi, E. Nunes, and P. Shakarian, "Community finding of malware and exploit vendors on darkweb marketplaces," in *Proc. 1st Int. Conf. Data Intell. Secur.*, 2018, pp. 81–84.

[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[38] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, 2004, Art. no. 026113.

[39] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *J. Amer. Med. Informat. Assoc.*, vol. 22, no. 3, pp. 671–681, 2015.

[40] Department of Vermont Health Access, "Department of vermont health access pharmacy benefit management program," 2019. Accessed: Sep. 17, 2019. [Online]. Available: https://dvha.vermont.gov/for-providers/vermont-pdl-effective-02-16-17-hepc-changes.pdf

[41] A. Perrin and M. Anderson, "nd share of US adults using social media, including facebook, is mostly unchanged since 2018," *Pew Res. Cent.* Accessed: Apr. 18, 2019, 2019. [Online]. Available: https://www. pewresearch. org/fact-tank/2019/04/10/share-of-us-adults-using-socialmedia-including-facebook-is-mostly-unchanged-since-2018

[42] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *J. Graph Algorithms Appl.*, vol. 10, no. 2, pp. 191–218, 2006.

[43] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[44] J. Ramos *et al.*, "Using TF-IDF to determine word relevance in document queries," in *Proc. 1st Instructional Conf. Mach. Learn.*, 2003, pp. 133–142.

[45] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Nat. Academy Sci. USA*, vol. 105, no. 4, pp. 1118–1123, 2008.

[46] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1986.

[47] L. Scholl, P. Seth, M. Kariisa, N. Wilson, and G. Baldwin, "Drug and opioid-involved overdose deaths—United States, 2013–2017," *Morbidity Mortality Weekly Rep.*, vol. 67, no. 5152, 2019, Art. no. 1419.

[48] A. Stroppa, D. di Stefano, and B. Parrella, "Social media and luxury goods counterfeit: A growing concern for government, industry and consumers worldwide," *The Washington Post*, pp. 1–50, 2016.

[49] J. Sun and Q. Jin, "Scalable RDF store based on HBase and MapReduce," in *Proc. 3rd Int. Conf. Adv. Comput. Theory Eng.*, 2010, pp. V1-633–V1-636.

[50] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.

[51] H. Tseng, P.-C. Chang, G. Andrew, D. Jurafsky, and C. D. Manning, "A conditional random field word segmenter for Sighan Bakeoff 2005," in *Proc. 4th SIGHAN Workshop Chin. Lang. Process.*, 2005, pp. 168–171.

[52] C. Walsh, "Drugs, the internet and change," *J. Psychoactive Drugs*, vol. 43, no. 1, pp. 55–63, 2011.

[53] Y. Wu, P. Skums, A. Zelikovsky, D. C. Rendon, and X. Liao, "Predicting opioid epidemic by using twitter data," in *Proc. Int. Symp. Bioinf. Res. Appl.*, 2018, pp. 314–318.

[54] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 2145–2158.

[55] J. Yang and J. Leskovec, "Overlapping community detection at scale: A nonnegative matrix factorization approach," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 587–596.

[56] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowl. Inf. Syst.*, vol. 42, no. 1, pp. 181–213, 2015.

[57] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," 2015, *arXiv:1510.03820*. [Online]. Available: http://arxiv.org/abs/1510.03820

[58] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, pp. 27–34, 2015.

[59] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2007, pp. 1601–1608.

Fengpan Zhao received the BS degree in electronic science and technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2012, and the MS degree in computer science from the University College Dublin, Dublin, Ireland, in 2014. He is currently working toward the PhD degree in the Department of Computer Science, Georgia State University, Atlanta, Georgia. His research focus on data mining, machine learning, and big data.

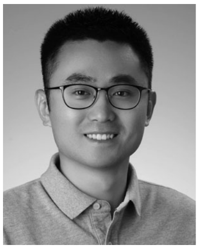
Pavel Skums received the BS and MS degrees in mathematics and the PhD degree in computer science from Belarusian State University, Minsk, Belarus, in 2004 and 2007, respecively. He is an assistant professor with the Department of Computer Science, Georgia State University. His research areas are computational genomics and computational biology, where his research focuses on studying epidemiology and the evolution of highly mutable RNA viruses, such as the human immunodeficiency virus (HIV) and the hepatitis C virus (HCV). Before joining Georgia State University in 2016, he was a regular fellow and associate service fellow in the National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention of the US Centers for Disease Control and Prevention (CDC), where his research was recognized by several CDC awards. He is the author of more than 80 refereed publications. His research has been supported by NIH.

Alex Zelikovsky received the BS degree in mathematics and the MS degree in computer science and mathematics from Moldova State University, Chişinău, Moldova, in 1980 and 1982, respectively, and the PhD degree in computer science from the Byelorussian Academy of Sciences, Minsk, Belarus, in 1989. Currently, he is a distinguished University professor with the Department of Computer Science, Georgia State University, Atlanta. His research interests include bioinformatics, discrete algorithms, VLSI CAD, combinatorial optimization, computational geometry, computational biology, and graph theory.

Eric L. Sevigny received the PhD degree in public and international affairs from the University of Pittsburgh, Pittsburgh, Pennsylvania, in 2006. He is associate professor of criminal justice and criminology with Georgia State University. His research interests lie at the intersection of drugs, crime and public policy, particularly around issues of sentencing and incarceration, drug courts, the measurement of drug-related problems including cryptomarkets, and the public health, safety, and economic impacts of marijuana and other drug policy reforms. His work appears in an array of interdisciplinary journals, including the *Criminology and Public Policy*, the *Journal of Quantitative Criminology*, the *International Journal of Drug Policy*, the *Drug and Alcohol Dependence*, the *Journal of Policy Analysis and Management*, and the *Social Indicators Research*.

Monica Haavisto Swahn is a distinguished professor in population health sciences with the School of Public Health, Georgia State University. Her research focuses on alcohol and drug use and associated harm, particularly in vulnerable populations. She also examines the structural drivers for alcohol and drug use and is developing strategies for harnessing technology for research and policy impact. Her research has been funded by the NIH and CDC and she is also a former fulbright scholar.

Sheryl M. Strasser is an associate professor with the School of Public Health, Georgia State University and part of its Partnership for Urban Health Research. She also is an affiliate faculty member of the Gerontology Institute. She previously held faculty positions within the New York University Colleges of Dentistry and Nursing and the University of Alabama at Birmingham School of Medicine. She has collaborated on externally-funded projects related to prescription drug abuse, youth alcohol use and prevention, tobacco control, and marijuana research totally more than $7 million in awards ranging from the National Institute of Aging, National Institute of Minority Health and Health Disparities, National Institute of Dental and Craniofacial Research, the Agency for Healthcare Research and Quality, as well as the RAND Foundation, among others. She has also completed a fellowship with the Institute of Healthcare Improvement in Cambridge, Massachusetts. She specializes in interdisciplinary health promotion planning and evaluation research focusing on substance abuse and prevention and healthy aging. She utilizes innovative teaching modalities to connect with graduate students-and she currently teaches graduate courses in intervention and evaluation research, public health ethics, as well as social and behavioral aspects of public health.



Yan Huang received the PhD degree from the Department of Computing Science, Georgia State University, Atlanta, Georgia. He is currently an assistant professor with the Department of Software Engineering & Game Development, Kennesaw State University (KSU). His research focus on cyber-security & privacy, IoT, and federated learning.

Yubao Wu received the PhD degree from Case Western Reserve University, Cleveland, Ohio, in 2016. After that, he joined the Department of Computer Science, Georgia State University as an assistant professor. He worked on big data, data mining, and their applications in social domains. He has published several papers in top-tier data mining conferences and journals.