2023 IEEE International Conference on Data Mining (ICDM) | 979-8-3503-0788-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDM58522.2023.00149

Hypergraph Contrastive Learning for Drug Trafficking Community Detection

Tianyi Ma[1†], Yiyue Qian[1†], Chuxu Zhang[2*], Yanfang Ye[1*]
[1]*Department of Computer Science and Engineering, University of Notre Dame, IN, USA.*
[2]*Department of Computer Science, Brandeis University, MA, USA.*
tma2@nd.edu, yqian5@nd.edu, chuxuzhang@brandeis.edu, yye7@nd.edu.

***Abstract*—In recent decades, due to the lucrative profits, the crime of drug trafficking has evolved with modern technologies. Social media, as one of the popular online platforms, have become direct-to-consumer intermediaries for illicit drug trafficking communities to promote and trade drugs. These group-wise drug trafficking activities pose significant challenges to public health and safety, requiring urgent measures to address this issue. However, existing works against the imminent problem still face limitations, such as primarily analyzing individual roles from a single perspective, ignoring the group-wise relationships, and requiring sufficient labeled samples for model training. To this end, we propose a novel *Hyper*G*raph* C*ontrastive* L*earning* framework called HyGCL-DC that employs hypergraph to model the higher-order relationships among users to detect *D*rug trafficking *C*ommunities. Firstly, we build a hypergraph called Twitter-HyDrug including online user nodes and four types of hyperedges to depict the rich group-wise relationships among these users. Then, we leverage hypergraph neural networks to model the rich relationships among nodes and hyperedges in the drug trafficking hypergraph. Furthermore, we design a hypergraph self-supervised contrast module, which integrates the augmentation from the structure view and the attribute view to enhance hypergraph representation learning over unlabeled data. Finally, we design an end-to-end framework that combines the self-supervised contrastive module and the supervised module to classify online drug trafficking communities. To comprehensively study the online drug trafficking problem and evaluate our model, we conduct extensive experiments over Twitter-HyDrug and three citation benchmark hypergraph datasets to demonstrate the effectiveness of our model. Our new data and source code are available at https://github.com/HyGCL-DC.**

***Index Terms*—hypergraph representation learning, self-supervised learning, community detection, drug trafficking**

I. Introduction

The illicit drug trafficking markets, encompassing drugs such as synthetic opioids, remain highly profitable in recent decades. As a result, the crime of drug trafficking (a.k.a. illicit drug trading) has increasingly adapted and evolved with modern technologies (e.g., social media). Recent works [1]–[3] have demonstrated that the major social media platforms, e.g., Twitter and Instagram, have become direct-to-consumer intermediaries for illicit drug trafficking, enabling drug sellers to sell drugs and drug users to purchase drugs much more easily than before. For instance, an illicit drug seller on Twitter advertises their drugs by posting drug-related content (e.g., drug street names and related hashtags), which easily attracts potential drug users to discuss and trade drugs through social media. Consequently, these activities naturally form drug trafficking communities on social media platforms, and these group-wise drug trafficking scenarios pose unprecedentedly serious challenges to social health and public safety, which needs imminent actions to address this issue.

However, existing works against drug trafficking activities still face the following limitations: (i) most works [1], [4] primarily study drug trafficking by analyzing individual roles from a single perspective (either from the drug seller side or drug user side) while ignoring the natural connections among different roles in drug trafficking communities. (ii) some existing graph models [5]–[7] merely focus on the pairwise relationships among users on social media but fail to model the higher-order group-wise relationships among these communities. For example, a drug user replies to the tweet of a drug seller inquiring about the drug price, and another drug buyer interacts with the drug user about previous purchases from the drug seller. These active and group-wise interactions among drug sellers, drug buyers, and drug users naturally form online drug communities (e.g., opioid community and depressant community). However, existing works [2], [8] do not capture more complex group-level behaviors exhibited by users within these communities. (iii) most of the existing works against drug trafficking activities require sufficient labeled samples to train models, but they underestimate the valuable information within the handy unlabeled data. For example, Roy et al. [9] spent a couple of months collecting 100,500 Instagram posts, while only 20% of posts were positive drug tweets, which is very time-consuming and effort-consuming. The aforementioned challenges inspire us to investigate the following research problem: *How do we design an effective graph representation learning framework to study drug trafficking communities comprehensively?*

To this end, we design a novel **H**yper**G**raph **C**ontrastive **L**earning framework called **HyGCL-DC** that leverages hypergraphs to model the higher-order relationships among online users to detect **D**rug trafficking **C**ommunities. To handle the first challenge, we comprehensively study online drug communities that are involved with four types of roles (i.e., drug seller, drug buyer, drug user, and drug discussant) on social media. For the second challenge, we first build a drug trafficking hypergraph called Twitter-HyDrug including online users and four types of hyperedges among these users. Then, we employ

† Equally contributed.
∗ Corresponding authors.

979-8-3503-0788-7/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDM58522.2023.00149

hypergraph neural networks (HyGNNs) to model the higher-order relationships among users and hyperedges. To solve the third challenge, inspired by self-supervised contrastive learning methods [10], [11], we design a hypergraph contrast learning method, which integrates hypergraph augmentations from the structure view and the attribute view to enhance the hypergraph representation learning over unlabeled data in hypergraphs. Furthermore, our framework is designed as an end-to-end model that combines self-supervised contrastive learning and supervised learning to detect drug communities on social media. To validate the effectiveness of HyGCL-DC, besides the newly collected data Twitter-HyDrug, we also evaluate our model over three benchmark hypergraphs. To conclude, our work makes the following contributions:

- ***Novelty:*** We devise a novel framework called HyGCL-DC, which effectively captures group-wise behaviors among online users to detect drug trafficking communities. To the best of our knowledge, this is the first work that employs hypergraph contrastive learning to detect drug trafficking communities on social media.
- ***New Data:*** To comprehensively study drug trafficking activities, we collect a new drug trafficking hypergraph from Twitter called Twitter-HyDrug, which contributes to research communities of drug trafficking and hypergraph learning.
- ***Effectiveness:*** Comprehensive experiments on three benchmark hypergraph datasets and the new data Twitter-HyDrug demonstrate the effectiveness of HyGCL-DC.

II. Related Work

Community Detection. Community detection is frequently employed in network analysis, which involves the segmentation of nodes in networks into distinct groups or clusters according to various criteria [12], [13]. Existing community detection methods can be roughly divided into three categories: optimization-based methods [14]–[16], matrix factorization methods [17], and generative models [18], [19]. The existing algorithms have demonstrated significant performance across various domains; however, these methods mainly focus on pairwise structure relationships and fail to preserve the higher-order structure relationships within the graphs. Unlike existing works, we leverage hyperedges to extract the higher-order relationships among entities to analyze online drug trafficking communities comprehensively.

Hypergraph Neural Networks. Hypergraphs are usually regarded as a generalized version of standard graphs because hypergraphs employ the hyperedge to connect multiple nodes [20], [21]. Accordingly, hypergraph neural networks models (HyGNNs) [22]–[24] have gained considerable attention in recent years with their strong ability to capture complex relationships among networks. For instance, Hypergraph Neural Network (HGNN) [20] is one of the earliest works in this field, which encodes higher-order data correlation in a hypergraph structure. Another notable work is Hypergraph Convolutional Neural Network (HyperGCN) [21], which extends Graph Convolution Network (GCN) to hypergraph through hypergraph Laplacian. Motivated by existing HyGNNs, this work proposes an effective hypergraph representation learning framework to learn the complex relationships for drug trafficking community detection on social media.

III. Preliminary

Definition III.1. Hypergraph. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X})$ denotes a hypergraph, where $\mathcal{V}$ is the set of nodes with size $N = |\mathcal{V}|$, $\mathcal{E}$ is the set of hyperedges with size $M = |\mathcal{E}|$, and $\mathcal{X}$ is the set of node attribute features where $x_i \in \mathbb{R}^d$. Unlike the pairwise edges in graphs, each hyperedge $e \in \mathcal{E}$ can connects multiple nodes and represents higher-order interactions among nodes.

Definition III.2. Community Detection. Given a network, the community detection aims to partition nodes in the network into K communities $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$, where each community C_k is a set of nodes. This work studies the overlapping community detection problem that each node can belong to multiple communities simultaneously, i.e., $\forall v_i \in \mathcal{V}$, $|\{k : v_i \in C_k, C_k \in \mathcal{C}\}| \geq 1$.

Definition III.3. Hypergraph Laplacian. In this paper, we select HyperGCN [21] as the hypergraph encoder, which leverages hypergraph Laplacian with mediators to transfer the hypergraph to weighted graphs. Given a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X})$, and a real-valued signal $\mathcal{S} \in \mathbb{R}^N$, the hypergraph Laplacian constructs a weighted graph $G_{\mathcal{S}}$ with all nodes in $\mathcal{V}$ and edges from the edge set where $\{(v_i, v_j) : (v_i, v_j) = \text{argmax}_{v_i, v_j \in e} |\mathcal{S}_i - \mathcal{S}_j|, \} \bigcup \{(v_m, v_i), (v_m, v_j) : v_m \in e \setminus \{v_i, v_j\}\}$. Formally, the symmetrically normalized hypergraph Laplacian can be formulated as follows:

$$l(\mathcal{S}) = (I - D^{-\frac{1}{2}} A_{\mathcal{S}} D^{-\frac{1}{2}})\mathcal{S}, \tag{1}$$

where A_S denotes the weighted adjacency matrix of graph $\mathcal{G}_{\mathcal{S}}$, I denotes the identity matrix, and $D = \text{diag}(d_1, \ldots, d_N)$ is the diagonal degree matrix of $G_{\mathcal{S}}$.

Problem 1. ***Hypergraph Contrastive Learning for Drug Trafficking Community Detection.*** *Given a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X})$ built on drug trafficking data, the objective is to build a hypergraph contrastive learning model $f_\phi : \mathcal{V} \rightarrow \mathbb{R}^b$ (with parameter ϕ) to project nodes into b-dimensional embeddings for drug trafficking community detection.*

IV. Methodology

In this section, we present the details of HyGCL-DC, which includes three key steps: (1) drug hypergraph construction; (2) hypergraph contrastive learning; (3) community detection.

A. Drug Hypergraph Construction

To comprehensively describe drug trafficking communities on social media, we propose to construct a drug trafficking hypergraph that integrates informative content features and complex higher-order relationships among online users. The details of content-based features and higher-order relationships are described below.

Content Feature. In this paper, we regard Twitter as an example to study online drug trafficking activities. Specifically, we consider each Twitter user as a node in our hypergraph named Twitter-HyDrug. To accurately characterize each user

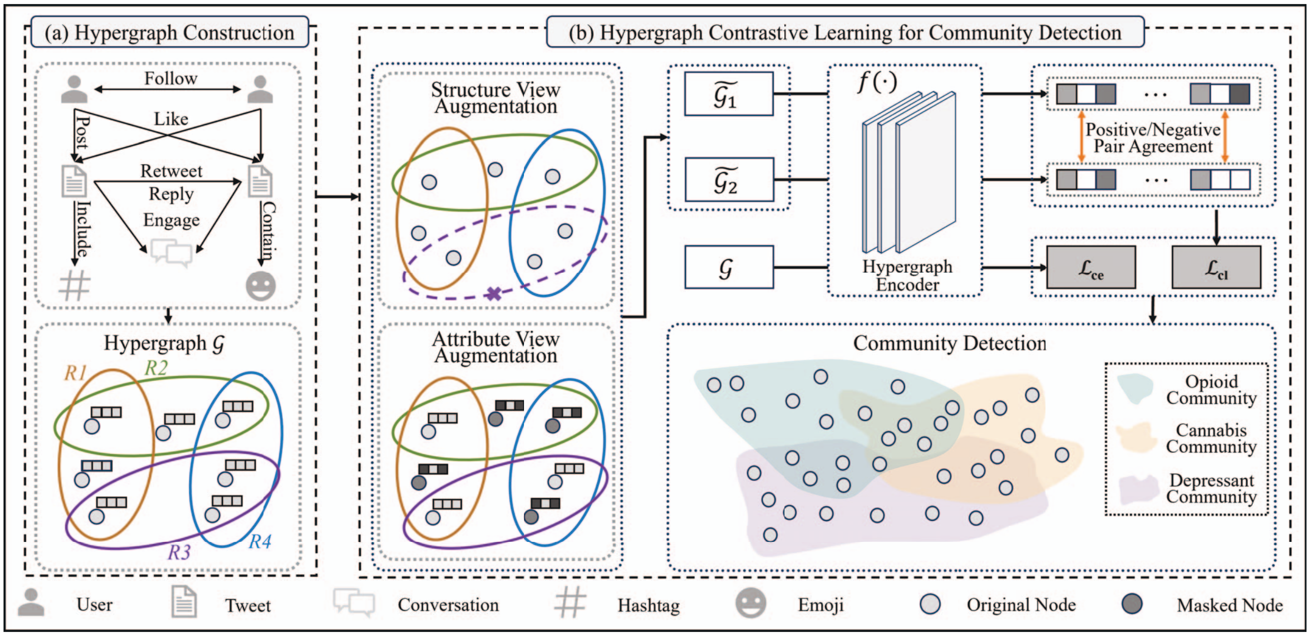

Fig. 1: The overall framework of HyGCL-DC: (a) it first constructs a hypergraph $\mathcal{G}$ based on the interactions among online drug-related users; (b) it integrates augmentations from the structure view and the attribute view to augment hypergraphs into $\widetilde{\mathcal{G}}_1$ and $\widetilde{\mathcal{G}}_2$. HyGCL-DC is designed as an end-to-end framework that integrates self-supervised contrastive learning to boost the node embeddings over unlabeled data by reaching the agreement among positive and negative embedding pairs and supervised learning with community labels for downstream drug trafficking community detection tasks.

in Twitter-HyDrug, similar to existing works [25]–[29], we concatenate the informative text content, including profile information, username, and tweets, and further leverage the pre-trained transformer-based language model, SentenceBert [30], to convert the concatenated text information to a fixed-length feature vector ($d = 384$). Each feature vector is applied to the corresponding node as the attribute feature x_i. Details about user feature generation are introduced in our GitHub page.

Hyperedge. To exhaustively depict the complex and group-wise relationships among users in Twitter-HyDrug, we define four types of hyperedges for describing the activities among users as follows: (i) *R1: users-follow-user* hyperedge relation denotes that a group of users follow a specific user in Twitter-HyDrug. The follow/following-based hyperedge aims to represent the social connections within drug trafficking communities, illustrating the friend circles involved in such illicit activities. (ii) *R2: users-engage-conversation* hyperedge relation represents that a group of users is engaged in a tweet-based conversation, encompassing activities such as posting, replying, retweeting, and liking the tweets involved within the conversation. The conversation-based hyperedge serves to portray the shared interests and topics among the group of users. (iii) *R3: users-include-hashtag* hyperedge relation indicates that a bunch of users actively discuss the specific hashtag-based topics by posting the specific hashtag in tweets or profiles. For instance, a hyperedge encompasses a group of users that post tweets on Twitter that include oxycodone, one of the opioid drugs. Note that, we follow our previous work [31] to define the drug-related hashtags. (iv) *R4: users-contain-emoji* hyperedge relation signifies that a bunch of users contains a specific drug-related emoji in their tweets or profiles. Similar to hashtags, we use emojis to describe the interested drugs in this group. Fig. 1.(a) intuitively shows how we build Twitter-HyDrug. To summarize, we build a hypergraph called Twitter-HyDrug by integrating content features and four types of hyperedge relationships among users.

B. Hypergraph Contrastive Learning

After the construction of Twitter-HyDrug, we propose to employ HyGNNs [20], [21] to model the complex relationships among hypergraphs. Besides, inspired by the self-supervised contrastive learning models [10], we devise a hypergraph self-supervised contrast learning model to enhance the hypergraph representation learning over unlabeled data.

Hypergraph Representation Learning. Our framework HyGCL-DC is applicable to any HyGNNs. In this work, we leverage a two-layer HyperGCN [21] as the encoder example to map nodes into a latent representation space. Formally, the propagation rule of a two-layer HyperGCN is defined as:

$$Z = \bar{A}^{(2)}\text{ReLU}(\bar{A}^{(1)}\mathcal{X}\bar{W}^{(1)})\bar{W}^{(2)}, \tag{2}$$

where $\bar{A}^{(1)}$ and $\bar{A}^{(2)}$ are the weighted adjacency matrices generated via Definition III.3 in the first layer and second layer, respectively. $\bar{W}^{(1)}$ is the weight matrix for the first layer, and $\bar{W}^{(2)}$ is the weight matrix for the second layer.

Hypergraph Contrastive Learning. After obtaining the node embeddings Z, we design a self-supervised contrast learning module to enhance the expressive ability of the hypergraph encoder $f(\cdot)$. The main idea of graph contrastive learning is to apply graph augmentation methods to convert graphs into different views and further achieve the agreements among node embedding pairs [11], [32]. Existing works conclude that the quality of contrastive pairs has a significant influence on graph contrastive learning, and relatively challenging contrastive pairs will enhance contrastive learning [33]–[35]. Inspired by the findings and the success of HyperGCL [10] that extends graph contrastive learning into the hypergraph field, we design a novel hypergraph contrastive model that aims to generate high-quality hypergraph contrastive pairs by perturbing the hyperedges from the structure view and corrupting node attributes from the attribute view simultaneously.

Structure View Augmentation. Based on the underlying prior that the absence of certain higher-order relations does not significantly affect the semantics of hypergraphs, We first corrupt the hypergraph structures by perturbing partial hyperedges in hypergraphs. Mention that, unlike edge perturbation in graphs that randomly removes or adds edges among nodes, we propose to merely remove hyperedges in hypergraphs as adding hyperedges for a group of nodes would be risky and would bring too much unnecessary or even harmful noise. Specifically, we generate a hyperedge masking matrix $\widetilde{\mathcal{M}}^s \in \{0,1\}^{1\times M} \sim \mathcal{B}(p_s)$. The augmented hyperedge set $\widetilde{\mathcal{E}}$ is a subset of $\mathcal{E}$, where $\widetilde{\mathcal{E}} = \{e_i : \widetilde{\mathcal{M}}^s_i = 1, e_i \in \mathcal{E}\}$.

Attribute View Augmentation. Instead of merely corrupting the hypergraph structure, we further design a hypergraph augmentation method that aims to corrupt the node attribute features for generating more challenging contrastive hypergraph pairs. Our attribute view augmentation is based on the idea that the corruption of the attribute in partial nodes would not significantly affect the semantics of nodes. To achieve this, we first generate a mask matrix $\widetilde{\mathcal{M}}^a \in \{0,1\}^{1\times N} \sim \mathcal{B}(p_a)$ to mask partial nodes. Then we corrupt the node attribute features by generating the random noise λ_i. Based on the above strategy, the augmented node attribute feature set $\widetilde{\mathcal{X}}$ is formulated as $\widetilde{\mathcal{X}} = \{x_i \cdot \widetilde{\mathcal{M}}^a_i + \lambda_i : x_i \in \mathcal{X}\}$.

Contrastive Optimization. Merely performing the hyperedge perturbation at the structure level or the attribute corruption augmentation at the node attribute level on hypergraphs is not optimal enough to generate challenging contrastive pairs for hypergraph contrastive learning. In this paper, we combine the hyperedge perturbation and the attribute corruption to generate high-quality contrastive pairs and further enhance the ability of hypergraph representation learning. Following the above augmentation methods, we first obtain the augmented hypergraph pairs $[\widetilde{\mathcal{G}}_1, \widetilde{\mathcal{G}}_2] = [(\mathcal{V}, \widetilde{\mathcal{E}}_1, \widetilde{\mathcal{X}}_1), (\mathcal{V}, \widetilde{\mathcal{E}}_2, \widetilde{\mathcal{X}}_2)]$. Then, the augmented hypergraph pairs are fed to the hypergraph encoder $f(\cdot)$ to get the node embeddings $\widetilde{Z}_1$ and $\widetilde{Z}_2$, respectively. Note that a projection head layer $h(\cdot)$ is applied to convert node embeddings from different augmented hypergraphs into the same space. Afterward, the hypergraph encoder $f(\cdot)$ is trained via optimizing the contrastive loss $\mathcal{L}_{cl}$, which attempts to maximize the consistency between the positive embedding pairs and the negative embedding pairs. The temperature-scaled contrastive loss (NT-Xent) $\mathcal{L}_{cl}$ is formulated as:

$$\mathcal{L}_{cl} = -\frac{1}{N}\log\sum_{v_i\in\mathcal{V}} \frac{\exp(\delta_{i,i}/\tau)}{\sum_{j\neq i}\exp(\delta_{i,j}/\tau)+\exp(\delta_{i,i}/\tau)}, \qquad (3)$$

where $\delta_{i,j}$ is the cosine similarity between contrastive embedding pairs $(\widetilde{z}_{1,i}, \widetilde{z}_{2,j})$. τ is the temperature hyper-parameter.

C. Community Detection

After obtaining self-supervised node embeddings generated by the hypergraph encoder, inspired by HyperGCL that designs an end-to-end framework for downstream tasks [10], we design an end-to-end framework to detect communities among online drug trafficking activities. Specifically, as illustrated in Fig. 1.(b), we first feed the original hypergraph $\mathcal{G}$ into the hypergraph encoder to generate node embeddings Z. Afterward, motivated by the existing work [36] that converts the community detection task to the node classification task, we regard the drug trafficking community detection as a node classification task that aims to classify which communities each node should belong to. As we focus on over-lapping community detection in this work, we pass node embeddings Z into a fully connected layer with the *sigmoid* function to get the probability distribution $\hat{P}$. With the ground-truth label Y, we employ the binary cross-entropy (BCE) loss as the community detection loss $\mathcal{L}_{ce}$. Then, the final objective for community detection can be formally defined as:

$$\mathcal{L} = \alpha_1\mathcal{L}_{ce} + \alpha_2\mathcal{L}_{cl}, \qquad (4)$$

where α_1 and α_2 are the trade-off hyper-parameters.

V. EXPERIMENTS

A. Datasets

New Real-world Dataset. To comprehensively study online drug trafficking activities, we build a new real-world dataset called Twitter-HyDrug to analyze drug trafficking communities on Twitter. Specifically, we first crawl the metadata through the official Twitter API [43] from Dec 2020 to Aug 2021. Afterward, following the existing work [1], we generate a keyword list that covers 21 drug types that may cause drug overdose or drug addiction problems to filter the tweets that contain drug-relevant information. Based on the keyword list, we obtain 266,975 filtered drug-relevant posts by 54,680 users. Moreover, we define six types of communities, i.e., cannabis, opioid, hallucinogen, stimulant, depressant, and others communities, based on the drug functions, and we, six researchers, spent 62 days annotating these Twitter users into six communities. The annotation rules are discussed in our provided GitHub page. To conclude, Twitter-HyDrug includes 2,936 user nodes and 33,892 hyperedges. The task of drug trafficking community detection over Twitter-HyDrug is considered the overlapping community detection (multi-label classification) problem.

Existing Benchmark Datasets. To exhaustively evaluate the effectiveness of HyGCL-DC, we also employ three citation hypergraph benchmark datasets [21]: Cora-author, Cora-citation, and Citeseer-citation. TABLE II lists the statistics of our newly collected Twitter-HyDrug and three benchmark datasets.

TABLE I: Performance comparison (Mean % ± std) of all methods for community detection. The train/validate/test ratio is 60%:20%:20%. Purple shaded numbers indicate the best result and gray shaded numbers represent the runner-up performance.

Setting		Twitter-HyDrug		Cora-author		Cora-citation		Citeseer-citation	
Group	Model	Jaccard	F1-score	Jaccard	F1-score	Jaccard	F1-score	Jaccard	F1-score
G1	K-means [37]	15.43 ± 0.33	31.53 ± 2.05	12.15 ± 4.36	22.10 ± 5.68	18.36 ± 4.91	26.64 ± 5.68	15.38 ± 4.32	20.10 ± 3.28
	BigClam [38]	23.46 ± 2.54	36.74 ± 7.24	19.73 ± 4.93	27.71 ± 5.28	21.53 ± 5.28	28.54 ± 4.29	17.21 ± 4.15	20.45 ± 4.05
	CESNA [39]	37.26 ± 4.60	40.83 ± 3.45	21.81 ± 5.04	31.02 ± 4.19	20.41 ± 3.46	34.31 ± 3.72	20.94 ± 2.97	23.15 ± 2.84
G2	GCN [5]	44.56 ± 1.03	61.64 ± 1.00	42.73 ± 3.44	70.24 ± 5.41	47.83 ± 1.46	63.91 ± 2.61	47.83 ± 0.62	51.89 ± 0.78
	GAT [6]	48.65 ± 2.02	60.35 ± 1.39	51.73 ± 8.43	67.75 ± 7.99	45.92 ± 6.91	62.44 ± 8.12	23.14 ± 3.79	37.42 ± 5.18
	GIN [40]	45.07 ± 0.82	61.74 ± 0.82	59.69 ± 4.80	70.95 ± 3.75	57.56 ± 0.83	70.60 ± 0.69	48.20 ± 2.30	65.02 ± 2.08
G3	CLARE [13]	50.17 ± 3.06	64.55 ± 3.95	54.19 ± 8.19	71.34 ± 5.26	55.26 ± 4.12	70.83 ± 3.19	48.70 ± 1.23	62.12 ± 2.67
	SEAL [12]	40.24 ± 2.37	58.92 ± 2.19	48.96 ± 6.48	60.07 ± 4.43	50.25 ± 5.10	65.26 ± 4.43	38.26 ± 1.37	56.45 ± 3.71
	Bespoke [41]	41.68 ± 3.74	59.02 ± 1.14	50.30 ± 6.25	63.19 ± 4.21	48.02 ± 3.17	64.89 ± 5.13	36.90 ± 2.93	51.64 ± 3.04
G4	HyperGCN [21]	56.83 ± 2.38	72.45 ± 1.93	66.15 ± 0.89	79.62 ± 0.64	62.86 ± 1.46	77.19 ± 1.11	55.15 ± 3.13	71.06 ± 2.58
	HGNN [20]	55.45 ± 0.44	72.16 ± 1.42	65.96 ± 0.74	79.54 ± 1.46	60.13 ± 2.14	76.39 ± 2.18	54.27 ± 1.47	68.59 ± 0.75
	HCHA [42]	52.78 ± 1.42	65.83 ± 1.42	58.84 ± 2.07	75.43 ± 1.60	56.29 ± 0.97	73.41 ± 1.81	52.89 ± 2.45	64.53 ± 1.89
Ours	**HyGCL-DC**	60.05 ± 0.54	74.85 ± 2.15	68.67 ± 0.94	81.20 ± 1.02	64.73 ± 0.14	78.59 ± 0.11	56.72 ± 2.85	72.36 ± 2.30

B. Baseline Methods

To evaluate HyGCL-DC, we compare it with twelve baseline methods which are divided into four groups: unsupervised community detection methods (G1), including K-means [37], BigClams [38], and CESNA [39]; supervised graph-based methods (G2), i.e., GCN [5], GAT [6], and GIN [40], supervised community detection methods (G3), i.e., CLARE [13], SEAL [12], and Bespoke [41], and supervised hypergraph-based methods (G4), i.e., HyperGCN [21], HGNN [20], and HCHA [42]. To fairly compare with graph methods, following existing works [10], [20], we transfer the hypergraph into a graph through clique expansion [44].

C. Experimental Settings

To evaluate the performance of our model and baseline methods, we adopt two widely-used metrics to evaluate the performance of community detection: Jaccard score [**?**], [45] and Micro-F1 score [41]. For data splitting, we use 60%/20%/20% of data as train/validate/test data respectively. Moreover, we conduct each method three times and report the average score with standard deviation (std). All experiments are conducted under the environment of the Ubuntu 16.04 OS, plus an Intel i9-9900k CPU, two GeForce GTX 2080 Ti Graphics Cards, and 64 GB of RAM.

TABLE II: The statistics of four hypergraph datasets.

	Twitter-HyDrug	Cora-author	Cora-citation	Citeseer-citation
# nodes, N	2,936	2,708	2,708	3,312
# hyperedges, M	33,892	1,072	1,579	1,079
Avg. hyperedges	2.4	4.2	3.0	3.2
# features	384	1,433	1,433	3,703
# communities, K	6	7	7	6

D. Experimental Comparison

Performance Comparison. According to TABLE I, we make the following conclusions: (i) Supervised models largely outperform all unsupervised learning models, showing that supervised learning with community labels can enhance community detection performance to a large extent. (ii) Most of the hypergraph-based models in G4 outperform graph-based models in G2, proving hypergraphs' necessity for community detection tasks. (iii) HyperGCN has better performance than other hypergraph models in G4 over four datasets, which is the motivation for selecting HyperGCN as the hypergraph encoder. (iv) Our contrastive hypergraph learning framework can enhance the ability of representation learnings in hypergraphs by comparing HyperGCN and HyGCL-DC. Besides, our model gains the best performance by comparison with all baseline models, which shows the effectiveness of our model over Twitter-HyDrug and three benchmark datasets.

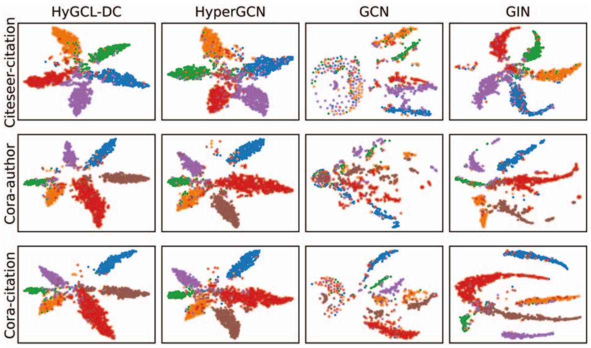

Fig. 2: Embedding Visualization over three benchmark data.

Embeddding Visualization. To further examine the effectiveness of our model intuitively, we render the embedding of three benchmark datasets generated by HyGCL-DC, HyperGCN, GCN, and GIN, respectively in Fig 2. Each unique color represents the embeddings belonging to a specific community label. We can find out that, compared HyGCL-DC with the three baseline models, HyGCL-DC shows more distinct boundaries and smaller overlapping areas. In addition, GCN and GIN appear to be unable to effectively separate different communities, especially since GCN has the largest overlapping area, which again demonstrates the effectiveness of HyGCL-DC for both overlapping (multi-label) and disjoint (multi-class) community detection tasks.

VI. Conclusion

In this work, we first collect a new drug trafficking hypergraph data called Twitter-HyDrug to study the drug trafficking community detection problem. Then we design a novel hypergraph contrastive learning framework called HyGCL-DC to detect online drug trafficking communities. To show the effectiveness of HyGCL-DC, we evaluate it over Twitter-HyDrug and other three benchmark hypergraph datasets. The empirical results on community detection tasks show the superiority of HyGCL-DC compared with baseline methods.

Acknowledgement

This work is partially supported by the NSF under grants IIS-2334193, IIS-2321504, IIS-2209814, IIS-2203262, IIS-2214376, IIS-2217239, OAC-2218762, CNS-2203261, and CMMI-2146076. Any opinions, findings, conclusions, and recommendations in this material are those of the authors and do not necessarily reflect the views of any funding agencies.

References

[1] Y. Qian, Y. Zhang, Y. Ye, and C. Zhang, "Distilling meta knowledge on heterogeneous graph for illicit drug trafficker detection on social media," in *NeurIPS*, 2021.

[2] Y. Zhang, Y. Qian, Y. Fan, Y. Ye, X. Li, Q. Xiong, and F. Shao, "dstyle-gan: Generative adversarial network based on writing and photography styles for drug identification in darknet markets," in *ACSAC*, 2020.

[3] Q. Wen, Z. Ouyang, J. Zhang, Y. Qian, Y. Ye, and C. Zhang, "Disentangled dynamic heterogeneous graph learning for opioid overdose prediction," in *KDD*, 2022.

[4] Y. Fan, Y. Zhang, Y. Ye, and X. Li, "Automatic opioid user detection from twitter: Transductive ensemble built on different meta-graph based similarities over heterogeneous information network." in *IJCAI*, 2018.

[5] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[6] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *ICLR*, 2018.

[7] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *KDD*, 2019.

[8] Y. Zhang, Y. Fan, W. Song, S. Hou, Y. Ye, X. Li, L. Zhao, C. Shi, J. Wang, and Q. Xiong, "Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network," in *WWW*, 2019.

[9] A. Roy, A. Paul, H. Pirsiavash, and S. Pan, "Automated detection of substance use-related social media posts based on image and text analysis," in *ICTAI*, 2017.

[10] T. Wei, Y. You, T. Chen, Y. Shen, J. He, and Z. Wang, "Augmentations in hypergraph contrastive learning: Fabricated and generative," in *NeurIPS*, 2022.

[11] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, "Graph contrastive learning with adaptive augmentation," in *WWW*, 2021.

[12] Y. Zhang, Y. Xiong, Y. Ye, T. Liu, W. Wang, Y. Zhu, and P. S. Yu, "Seal: Learning heuristics for community detection with generative adversarial networks," in *KDD*, 2020.

[13] X. Wu, Y. Xiong, Y. Zhang, Y. Jiao, C. Shan, Y. Sun, Y. Zhu, and P. S. Yu, "Clare: A semi-supervised community detection algorithm," in *KDD*, 2022.

[14] S. Ismail and R. Ismail, "Modularity approach for community detection in complex networks," in *IMCOM*, 2017.

[15] C. Zhou, Y. Wang, J. Zhang, J. Jiang, and D. Hu, "End-to-end modularity-based community co-partition in bipartite networks," in *CIKM*, 2022.

[16] L. Yang, X. Cao, D. He, C. Wang, X. Wang, and W. Zhang, "Modularity based community detection with deep learning." in *IJCAI*, 2016.

[17] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks," in *AAAI*, 2016.

[18] F.-Y. Sun, M. Qu, J. Hoffmann, C.-W. Huang, and J. Tang, "vgraph: A generative model for joint community detection and node representation learning," in *NeurIPS*, 2019.

[19] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *ICDM*, 2013.

[20] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *AAAI*, 2019.

[21] N. Yadati, M. Nimishakavi, P. Yadav, V. Nitin, A. Louis, and P. Talukdar, "Hypergcn: A new method for training graph convolutional networks on hypergraphs," in *NeurIPS*, 2019.

[22] R. Zhang, Y. Zou, and J. Ma, "Hyper-sagnn: a self-attention based graph neural network for hypergraphs," in *ICLR*, 2020.

[23] J. Yu, H. Yin, J. Li, Q. Wang, N. Q. V. Hung, and X. Zhang, "Self-supervised multi-channel hypergraph convolutional network for social recommendation," in *WWW*, 2021.

[24] L. Xia, C. Huang, and C. Zhang, "Self-supervised hypergraph transformer for recommender systems," in *KDD*, 2022.

[25] Y. Ye, Y. Fan, S. Hou, Y. Zhang, Y. Qian, S. Sun, Q. Peng, M. Ju, W. Song, and K. Loparo, "Community mitigation: A data-driven system for covid-19 risk assessment in a hierarchical manner," in *CIKM*, 2020.

[26] Y. Zhang, Y. Qian, Y. Ye, and C. Zhang, "Adapting distilled knowledge for few-shot relation reasoning over knowledge graphs," in *SDM*, 2022.

[27] Y. Ye, S. Hou, Y. Fan, Y. Zhang, Y. Qian, S. Sun, Q. Peng, M. Ju, W. Song, and K. Loparo, "α-satellite: An ai-driven system and benchmark datasets for dynamic covid-19 risk assessment in the united states," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[28] Y. Ye, S. Hou, Y. Fan, Y. Qian, Y. Zhang, S. Sun, Q. Peng, and K. Laparo, "α-satellite: An ai-driven system and benchmark datasets for hierarchical community-level risk assessment to help combat covid-19," *arXiv preprint arXiv:2003.12232*, 2020.

[29] Y. Qian, Y. Zhang, Y. Ye, and C. Zhang, "Adapting meta knowledge with heterogeneous information network for covid-19 themed malicious repository detection," in *IJCAI*, 2021.

[30] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *EMNLP-IJCNLP*, 2019.

[31] Y. Qian, Y. Zhang, Y. Ye, and C. Zhang, "Distilling meta knowledge on heterogeneous graph for illicit drug trafficker detection on social media-supplementary material," in *NeurIPS*, 2021.

[32] Y. Qian, C. Zhang, Y. Zhang, Q. Wen, Y. Ye, and C. Zhang, "Co-modality graph contrastive learning for imbalanced node classification," in *NeurIPS*, 2022.

[33] Y. Qian, Y. Zhang, Q. Wen, Y. Ye, and C. Zhang, "Rep2vec: Repository embedding via heterogeneous graph adversarial contrastive learning," in *KDD*, 2022.

[34] Y. Qian, Y. Zhang, N. Chawla, Y. Ye, and C. Zhang, "Malicious repositories detection with adversarial heterogeneous graph contrastive learning," in *CIKM*, 2022.

[35] Q. Wen, Z. Ouyang, C. Zhang, Y. Qian, Y. Ye, and C. Zhang, "Adversarial cross-view disentangled graph contrastive learning," *arXiv preprint arXiv:2209.07699*, 2022.

[36] Z. Chen, L. Li, and J. Bruna, "Supervised community detection with line graph neural networks," in *ICLR*, 2019.

[37] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society.*, 1979.

[38] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in *WSDM*, 2013.

[39] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *ICDM*, 2013.

[40] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *ICLR*, 2019.

[41] A. Bakshi, S. Parthasarathy, and K. Srinivasan, "Semi-supervised community detection using structure and size," in *ICDM*, 2018.

[42] S. Bai, F. Zhang, and P. H. Torr, "Hypergraph convolution and hypergraph attention," *Pattern Recognition*, 2021.

[43] "Twitter official api for developers," Twitter, 2023, https://developer.twitter.com/en/docs/twitter-api.

[44] L. Sun, S. Ji, and J. Ye, "Hypergraph spectral learning for multi-label classification," in *KDD*, 2008.

[45] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly, "Metrics for community analysis: A survey," *ACM Computing Surveys*, 2017.