

GNN-based Graph Anomaly Detection with Graph Anomaly Loss

Tong Zhao¹, Chuchen Deng¹, Kaifeng Yu¹, Tianwen Jiang^{1,2}, Daheng Wang¹, Meng Jiang¹

¹Department of Computer Science and Engineering, University of Notre Dame, USA

²Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
{tzhao2, cdeng, kyu2, dwang8, mjiang2}@nd.edu, twjiang@ir.hit.edu.cn

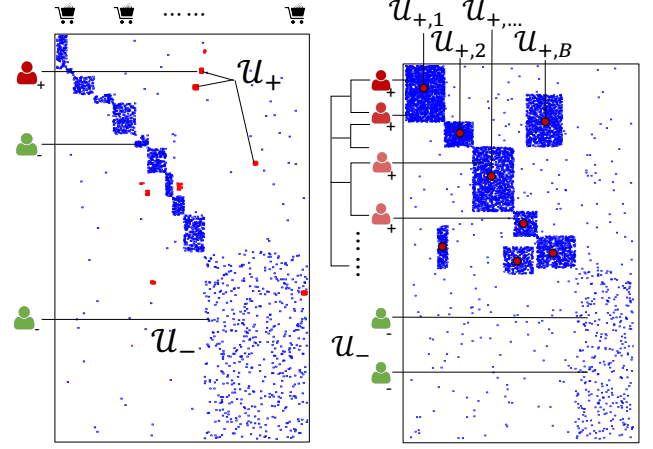
ABSTRACT

Graph neural networks (GNNs) have been widely used to learn node representations from graph data in an unsupervised way for downstream tasks. However, when applied to detect anomalies (e.g., outliers, unexpected density), they deliver unsatisfactory performance as existing loss functions fail. For example, any loss based on random walk (RW) algorithms would no longer work because the assumption that anomalous nodes were close with each other could not hold. Moreover, the nature of class imbalance in anomaly detection tasks brings great challenges to reduce the prediction error. In this work, we propose a novel loss function to train GNNs for anomaly-detectable node representations. It evaluates node similarity using global grouping patterns discovered from graph mining algorithms. It can automatically adjust margins for minority classes based on data distribution. Theoretically, we prove that the prediction error is bounded given the proposed loss function. We empirically investigate the GNN effectiveness of different loss variants based on different algorithms. Experiments on two real-world datasets show that they perform significantly better than RW-based loss for graph anomaly detection.

1 INTRODUCTION

Detecting anomalies from large-scale graphs is an important task on many real-world applications. There are two main types of graph anomalies: *graph outliers* (e.g., fake reviewers) and *unexpected dense blocks* in graph’s adjacency matrix (e.g., spammers, botnets). To learn node representations for such downstream tasks, Graph Neural Networks (GNNs) have been highly recognized for their abilities of aggregating attributed information from local neighborhood [33, 51, 53]. Usually, the models have two to four layers (i.e., #hops in the local neighborhood), sufficient for aggregation; to train the model parameters in an unsupervised manner (when labels are hardly available), random walk (RW) algorithms can discover more “global properties” (i.e., node pair-wise similarity in longer distance) that form *RW-based loss* on the last layer [19, 55].

However, we found that existing GNN models performed poorly on benchmarks in the task of graph anomaly detection. The reason is that graph anomalies do not have the aforementioned RW-based global properties. In other words, nodes of the same class might not be closer in the graph than those of different classes. For example, the graph outliers (e.g., fake reviewer group \mathcal{U}_+ in Figure 1a) do not have to be connected nor have common neighbors. They share global properties of being outliers (away from the majority) on the graph. Another example is that when a graph’s adjacency matrix has multiple dense blocks (e.g., social botnet groups $\{\mathcal{U}_{+,i}\}_{i=1}^B$ in Figure 1b), the nodes in different blocks do not have to be connected while having similar global properties of creating unexpected density. How to effectively train GNNs for graph anomaly detection by capturing proper global properties is important and non-trivial.



(a) Anomalous node groups forming graph outliers (b) Anomalous node groups forming dense blocks

Figure 1: New global properties and new loss functions must be defined to train effective GNNs for graph anomaly detection. Nodes in different anomalous groups might not be reachable within random-walk distance.

Designing a proper loss is non-trivial because the GNN models, after trained with the loss, are expected to accurately predict minority groups (e.g., outliers) from class-imbalanced data [7]. Compared with the node population of entire graph, graph outliers (e.g., fake reviewers) and/or nodes in dense blocks (e.g., botnet accounts) are the minority. Such severe imbalance is detrimental to model performance: When representations were not properly trained, the models would meet over-fitting on the minority classes and would perform poorly on test/unseen data [7]. Imbalanced machine learning has been studied from many perspectives [10, 23]; however, to the best of our knowledge, no existing work has been done to address the problem of reducing predictive error on imbalanced data for unsupervised graph representation learning.

Present work. In this paper, we evaluate node pair-wise similarity using the “global properties” discovered by *graph mining algorithms*, and we present a novel *error-bounded Graph Anomaly Loss (GAL)* that is designed based on the similarities to learn effective node representations for the task of graph anomaly detection.

Here the graph mining algorithms refer to the algorithms that use heuristics (e.g., greedy search, spectral methods) to identify abnormal node groups from graph data. For example, Akoglu *et al.* [1], Rayana *et al.* [45], and Kumar *et al.* [37] proposed efficient algorithms to detect graph outliers on web platforms by measuring the distance of their behavioral patterns from the pattern of

the majority. Jiang *et al.* [29, 30] and Hooi *et al.* [25] identified botnets by measuring the unexpectedness of dense bipartite cores and greedily looking for them in large social networks. However, these algorithms ignore node individual identity and attributed information, assuming all the nodes in a group must have the same characteristics and the same label. Though the assumption is too strong to work on itself, these algorithms capture global structural information which can be used to train GNN parameters, specifically, to supervise the process of feature aggregation from local neighborhood. The GNN models will take advantage of both the local and global contexts for effective graph anomaly detection.

Our proposed GAL can be generalized for both categories of graph mining algorithms, i.e., *graph outlier detection* [1, 37] and *dense block detection* [25, 29, 30] algorithms, to train GNNs for different purposes. It can be applied to all types of existing GNN algorithms such as GCN [33], GAT [51], and GRAPH-SAGE [19]. Moreover, GAL automatically encourages larger margins for minority classes. It learns proper margins from the global imbalanced data distribution discovered by graph mining algorithms. So it creates better generalization on predicting minority classes. Theoretically, we obtain and prove the bound on prediction error.

Here we summarize the important features of our proposed GAL:

- **Generalizability:** GAL variants include loss functions for different tasks of graph anomaly detection such as outlier detection and unexpected density detection. It can be applied to train an arbitrary graph neural algorithm.
- **Effectiveness:** GAL captures global properties when random walk-based loss fails. It trains GNNs effectively for graph anomaly detection. Experiments on two real-world datasets, aiming at detecting two different kinds of anomalies, demonstrate that any GNN algorithm trained by the proposed GAL can perform significantly better.
- **Theoretical guaranteed performance:** GAL creates better generalization on patterns of minority groups. It maintains a bounded test prediction error on imbalanced data.

2 RELATED WORK

In this section, we survey research work of five related topics.

Imbalanced learning. Learning with imbalanced data has always been a challenging problem for machine learning. Most existing work focused on sampling and generating techniques. These algorithms either under-sample/over-sample the data objects [9, 40] or generate new data objects for the minority classes [10, 18, 22]. [32, 35] proved that generalization error for both linear and non-linear models with hinge losses is bounded. Recently, Cao *et al.* [7] showed that the error bound could be found on imbalanced datasets.

Graph representation learning. The goal is to learn node representations in a low-dimensional space using random-walk paths or factorized features [16, 17, 20, 43, 50]. These algorithms are transductive as they directly train node embeddings for individual nodes and require retraining or additional training to generate embeddings for new nodes. DEEPWALK [43] and NODE2VEC [17] learned node embeddings by performing word embedding models WORD2VEC on “corpus” of nodes generated by random walk. BiNE [16] extended DEEPWALK and optimized for bipartite graphs. DEEPFD [52] learned the node embeddings with an encoder-decoder

structure specifically for the task of fraud detection. AdONE [6] was an unsupervised auto-encoder based model that learns outlier resistant embeddings from attributed networks.

Graph neural networks. GNNs are deep learning architectures for graph structured data. The core idea is to learn node representations through local neighborhoods. Kipf *et al.* [33, 34] proposed graph convolutional network (GCN) for semi-supervised graph representation learning. GCN is a transductive model that requires the calculation of whole graph Laplacian during training. Many inductive GNNs [12, 41, 51, 53, 54] that follow a neighborhood aggregation scheme are proposed in recent years. In these models, the representation of a node is computed by recursively aggregating representations of its neighbors. Bi-HGNN [39] learned hierarchical node representations from bipartite graphs that contains community information for recommendation tasks.

Graph outlier detection. The goal is to find outlier nodes in large graphs [3]. Traditional density-based clustering methods [8, 13, 21] regarded nodes in sparse regions as outliers. Similar approaches have been developed for bipartite graphs [49]. SPOTLIGHT [14] detected outlier anomalies in streaming graphs. ODDBALL [2] detected outlier anomalies in weighted graphs. BIRDNEST [24] was a Bayesian inference model for rating networks that modeled user rating behaviors and detected outliers. REV2 [37] was an iterative algorithm that calculated reviewer fairness scores and considered users with low scores as outliers. TWOFACE [31] uncovered distant user nodes that serve structurally similar roles of being review manipulators. DOMINANT [11] was a graph auto-encoder based deep model that detects anomalous nodes from attributed networks. REPEN [42] unifies representation learning and outlier detection to learn representations that are tailored for the random distance-based outlier detection methods.

Unexpected density detection. The goal is to find suspicious nodes by looking for dense blocks in the graph’s adjacency matrix. SPOKEN [44] found the “spokes” pattern on pairs of eigenvectors of graphs. LOCKINFER [30] identified pattern of communities based on singular vectors of graphs. FBBOX [47] located mini-scale attacks missed by spectral techniques. CATCHSYNC [29] and CROSSPOT [28] found the lockstep behaviors made by fraudulent users. Several methods [25, 46, 48] utilized dense subgraph detection algorithms on graph or tensor data to locate the suspicious dense blocks. Hooi *et al.* [26] showed that the edge density-based suspiciousness of subgraph can be maximized with approximation guarantee. Shin *et al.* [48] adopted and extended the density approximation guarantee to dense subtensor detection. Zhao *et al.* [56] proposed an actionable algorithm to block the dense subgraphs on bipartite graphs.

3 PROPOSED METHOD

3.1 Problem Definition

The goal of our approach is to learn low-dimensional representations of user nodes on a bipartite graph for detecting anomalous users. Suppose that \mathcal{U} is the set of users, \mathcal{V} is the set of items (e.g., products, hashtags), and $\mathcal{R} = \{r_{u,v} | u \in \mathcal{U}, v \in \mathcal{V}\}$, where $r_{u,v}$ denotes the weight of the edge between node u and node v . The problem is defined as follows:

Given a bipartite graph $G(\mathcal{U}, \mathcal{V}, \mathcal{R})$ and a set of user’s node feature vectors $\{\mathbf{x}_u \in \mathbb{R}^{d_x}, \forall u \in \mathcal{U}\}$ (where d_x is the dimension of raw

features), **find** a mapping function of the representations of user nodes $f : u \in \mathcal{U} \rightarrow \mathbf{z}_u \in \mathbb{R}^d$ where d is the number of latent dimensions in user embeddings. We expect the user representations are optimized for the task of anomaly detection by preserving both user's node attribute information and proper global properties.

Currently most GNNs are designed for homogeneous graphs by default. GNNs generate embeddings of a node by aggregating the embeddings from nodes in its local neighborhood. The assumption is that neighboring nodes have related information and/or similar characteristics. However, on a bipartite graph, this assumption does not hold as neighbors are different types of nodes. Hence in order to have GNNs better perform on bipartite graphs, rather than aggregating embeddings from immediate neighbors, we aggregate from 2-hop neighbors which are the same type as the target node.

3.2 Distribution-aware Anomaly Margin Loss

Traditionally, in order to train the network model in an unsupervised manner, RW-based loss functions are often applied to learn the output representations, $\mathbf{z}_u, \forall u \in \mathcal{U}$, and to tune the weight matrices $\mathbf{W}^k, \forall k \in \{1, \dots, K\}$ via stochastic gradient descent. The RW-based loss encourages nearby nodes to have similar representations as well as enforcing the representations of disparate nodes to be distinct, which can be formatted as [54]:

$$\mathcal{L}_{\text{RW}}(u) = \mathbb{E}_{u_+ \sim \mathcal{U}_{u_+}, u_- \sim \mathcal{U}_{u_-}} \max\{0, \mathbf{z}_u^T \mathbf{z}_{u_+} - \mathbf{z}_u^T \mathbf{z}_{u_-} + \Delta\}, \quad (1)$$

where Δ denotes a fixed margin hyper-parameter, \mathcal{U}_{u_+} denotes the set of user nodes that are reachable with a fixed length random-walk starting from u , and \mathcal{U}_{u_-} denotes $\mathcal{U} \setminus \mathcal{U}_{u_+}$. However, as we mentioned before, it is neither proper nor effective when the task is to detect anomalies on graphs. Task-specific loss functions are desired. Fortunately, researchers have proposed algorithms about learning with imbalanced data [7, 32, 35] and mining anomalies from graph data [3, 26, 29, 37]. Hence we propose a task-specific class-distribution-aware graph anomaly loss that is able to utilize the results of these unsupervised algorithms. Here we first introduce the class-distribution-aware margin loss function.

Let y_u denote the the label of user node u (note that user nodes in different dense blocks should have different labels). We assume that the class-conditional distribution $\mathcal{P}(u|y_u)$ is the same at training and testing. Then let \mathcal{P}_j denote the class-conditional distribution, that is, $\mathcal{P}_j = (u|y_u = j)$. For our graph neural network model $f : \mathcal{U} \rightarrow \mathbb{R}^d$, we use function $g : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$ to denote the similarity of the representations of any two user nodes u and u' :

$$g(u, u') = f(u)^T \cdot f(u'), \quad (2)$$

and $L_{\text{bal}}[g]$ to denote the standard 0-1 test error on the balanced data distribution:

$$L_{\text{bal}}[g] = \Pr_{(u,j) \sim \mathcal{P}_{\text{bal}}} [\min_{y_{u_+}=j} g(u, u_+) < \max_{y_{u_-} \neq j} g(u, u_-)]. \quad (3)$$

The error L_j for class j is then defined similarly as:

$$L_j[g] = \Pr_{u \sim \mathcal{P}_j} [\min_{y_{u_+}=j} g(u, u_+) < \max_{y_{u_-} \neq j} g(u, u_-)]. \quad (4)$$

Let n_j be the number of user nodes in class j and $S_j = u : y_u = j$ denote the set of user nodes with label j . Define the training margin

for class j as:

$$\gamma_j = \min_{u \in S_j} \left(\min_{y_{u_+}=j} g(u, u_+) - \max_{y_{u_-} \neq j} g(u, u_-) \right), \quad (5)$$

where $\gamma_{\min} = \min\{\gamma_1, \dots, \gamma_j\}$ is the widely used training margin in previous studies [35]. Then we let $L_{\gamma,j}$ denote the margin loss for class j when training:

$$L_{\gamma,j}[g] = \Pr_{u \sim \mathcal{P}_j} [\min_{y_{u_+}=j} g(u, u_+) < \max_{y_{u_-} \neq j} g(u, u_-) + \gamma], \quad (6)$$

and let $\hat{L}_{\gamma,j}$ denote its empirical variant. For a hypothesis class \mathcal{G} , we use $\mathfrak{R}(\mathcal{G})$ to denote the empirical Rademacher complexity of margin for class j :

$$\hat{\mathfrak{R}}_j(\mathcal{G}) = \frac{1}{n_j} \mathbb{E}_{\sigma} \left[\sup_{g \in \mathcal{G}} \sum_{u \in S_j} \sigma_u [\min_{y_{u_+}=j} g(u, u_+) - \max_{y_{u_-} \neq j} g(u, u_-)] \right], \quad (7)$$

where σ is a vector of i.i.d. uniform $\{-1, +1\}$ bits. Here we consider the bound below for balanced test distribution by considering the margin of each class, which allows us to design distribution-aware margin loss function that is suitable for the imbalanced data.

THEOREM 3.1. [7] *With probability $1 - \delta$ over the randomness of the training data, for all choices of class-dependent margins $\gamma_1, \gamma_2, \dots, \gamma_k > 0$, all hypotheses $g \in \mathcal{G}$ will have balanced-class generalization bounded by:*

$$L_{\text{bal}}[g] \leq \frac{1}{k} \left(\sum_{j=1}^k \hat{L}_{\gamma_j,j}[g] + \frac{4}{\gamma_j} \hat{\mathfrak{R}}_j(\mathcal{G}) + \varepsilon_j(\gamma_j) \right), \quad (8)$$

where $\varepsilon_j(\gamma) \triangleq \sqrt{\frac{\log \log_2 \left(\frac{2 \max_{u, v \in \mathcal{U}, g \in \mathcal{G}} |g(u, v)|}{\gamma} \right) + \log \frac{2c}{\delta}}{n_j}}$ is typically a low-order term in n_j . Concretely, the Rademacher complexity $\hat{\mathfrak{R}}_j(\mathcal{G})$ will typically scale as $\sqrt{\frac{C(\mathcal{G})}{n_j}}$ for some complexity measure $C(\mathcal{G})$, in which case:

$$L_{\text{bal}}[g] \leq \frac{1}{k} \left(\sum_{j=1}^k \hat{L}_{\gamma_j,j}[g] + \frac{4}{\gamma_j} \sqrt{\frac{C(\mathcal{G})}{n_j}} + \varepsilon_j(\gamma_j) \right). \quad (9)$$

Note that although the losses and empirical Rademacher complexity of margins are defined different from those in Theorem 2 in [7], it can be proved that the above inequality still holds.

The balanced generalization error bound (Eq. 9) suggests that in order to improve the generalization of minority classes, we should enforce larger margins for them. However, manually assigning larger margins for minority classes may lead to sub-optimal margin for the frequent class and hence hurt the model's performance. Thus here we take the binary classification problem as an example of showing how to obtain the optimal trade-off.

When $k = 2$, we aim to optimize the balanced generalization error bound in Eq. 9, which can be simplified to (after removing constant factors, common factor $C(\mathcal{G})$ and low order term $\varepsilon_j(\gamma_j)$) [7]

$$\frac{1}{\gamma_1 \sqrt{n_1}} + \frac{1}{\gamma_2 \sqrt{n_2}}. \quad (10)$$

Although the it is hard to get the optimal margins with the above equation as they are complicate functions of the parameters in $g(\cdot)$, we can figure out the relative scales between the two margins.

Suppose we have $\gamma_1^*, \gamma_2^* > 0$ that minimize the equation above, we observe that any $\gamma_1' = \gamma_1^* - \delta$ and $\gamma_2' = \gamma_2^* + \delta$ (where $-2\gamma_2^* < \delta < \gamma_1^*$) can be realized by the same parameters with a shifted bias term. Therefore, for γ_1^*, γ_2^* to be optimal, the following inequality must be satisfied [7],

$$\frac{1}{\gamma_1^* \sqrt{n_1}} + \frac{1}{\gamma_2^* \sqrt{n_2}} \leq \frac{1}{(\gamma_1^* - \delta) \sqrt{n_1}} + \frac{1}{(\gamma_2^* + \delta) \sqrt{n_2}}, \quad (11)$$

which implies that

$$\gamma_1^* \propto n_1^{-1/4}, \text{ and } \gamma_2^* \propto n_2^{-1/4}. \quad (12)$$

Given the trade-off above, we can define our margin loss as

$$\mathcal{L}(u) = \max\{0, \max_{y_{v'} \neq y_u} g(u, v') - \min_{y_v = y_u} g(u, v) + \Delta_{y_u}\}, \quad (13)$$

where $\Delta_{y_u} = \frac{C}{n_{y_u}^{1/4}}.$

Here C is a constant hyper-parameter. When applying the above loss function on real-world graphs, it is impossible to enumerate all node pairs when calculating the minimum and maximum distances. Hence we use positive and negative sampling to approximate the distances. That is, we propose the following margin loss function in our GAL,

$$\mathcal{L}(u) = \mathbb{E}_{u_+ \sim \mathcal{U}_{u_+}, u_- \sim \mathcal{U}_{u_-}} \max\{0, g(u, u_-) - g(u, u_+) + \Delta_{y_u}\}, \quad (14)$$

where $\Delta_{y_u} = \frac{C}{n_{y_u}^{1/4}}.$

Here \mathcal{U}_{u_+} denotes the set of user nodes that has the same label as u , \mathcal{U}_{u_-} denotes $\mathcal{U} \setminus \mathcal{U}_{u_+}$, and $n_{y_u} = |\mathcal{U}_{u_+}|$.

3.3 Estimating the Parameters of GAL

Recall that our task is unsupervised. When applying the distribution-aware anomaly margin loss (Eq. 14), we are not aware of the label for each user node $u \in \mathcal{U}$ during the training process. Therefore, we utilize the results of existing unsupervised graph-based *outlier detection* [3, 37] and *dense blocks detection* [26, 29] algorithms to estimate the user node sets \mathcal{U}_{u_+} and \mathcal{U}_{u_-} for each user node u . We call GAL utilizing the results of these algorithms respectively as GAL with *graph outlier loss* and GAL with *dense block loss*.

3.3.1 Graph outlier detection algorithms.

Given a graph, they assign binary labels to nodes in an unsupervised way. We use \mathcal{U}_o and \mathcal{U}_n to denote the set of outlier nodes and the set of normal nodes, respectively. Here are the three categories of the algorithms:

- *Feature-based graph outlier detection.* These methods define the outlierness score of node u based on a pair of its particular features a_u and b_u [2, 29]:

$$\text{outlierness}(u) = |b_u - \hat{b}_u| \text{ or } \frac{\max(b_u, \hat{b}_u)}{\min(b_u, \hat{b}_u)} \cdot \log(|b_u - \hat{b}_u| + 1), \quad (15)$$

where \hat{b}_u is the predicted feature value based on the observed a_u . Intuitively, the measure is the “distance to fitting line (a, b).” Akoglu *et al.* [2] adopted four basic features such as number of neighbors, number of edges, total weight, and principal eigenvalue of the weighted adjacency matrix. Power laws were observed between the

features with a large population of nodes (i.e., $b_u \propto a_u^\gamma$, γ is a constant). Big distance to the power-law fitting line indicates the role of graph outlier. Jiang *et al.* [29] proposed two high-order features: one is called synchronicity which describes how similar a node’s neighbors are with each other in the space of basic features (e.g., degree, PageRank); the other is called normality that describes how similar the neighbor nodes are with every node in the space. They found the synchronicity had a parabolic lower limit of the normality (i.e., $\text{sync}_u \propto \alpha \cdot \text{norm}_u^2 + \beta$, α and β are constants) and designed an outlierness scoring function to catch the suspicious nodes. Big synchronicity and small normality indicate suspiciousness.

- *Structure-based graph outlier detection.* These methods define the outlierness score using the graph structure. They assume that the majority of users have low outlierness score. Then users with high outlierness scores can be reported as outliers [37]:

$$\text{outlierness}(u) = 1 - \frac{\sum_{v \in \mathcal{N}_V(u)} \hat{R}(u, v) + \alpha_1 \cdot \mu_f + \alpha_2 \cdot \Pi_U(u)}{|\mathcal{N}_V(u)| + \alpha_1 + \alpha_2}, \quad (16)$$

where $\hat{R}(u, v)$ is the normality score of the rating from u to item v , μ_f is the prior belief of u ’s normality score given by BIRDNEST [24]. $\Pi_U(u)$ is u ’s behavior normality score. α_1 and α_2 are constants.

- *Model-based graph outlier detection.* The idea behind these methods is that the majority of the graph, or say, the structural dependency can be learned by a specific graph model (e.g., compression model, generative model) and the outliers deviate significantly from the model [8, 15].

GAL with graph outlier loss. The graph outlier loss function encourages the pairs of positive nodes (outliers) to have similar representations and encourages the pairs of negative nodes (i.e., normal nodes) to have similar representations. It enforces that the representations of pairs of positive and negative nodes are highly distinct. Therefore, when sampling for each user node $u \in \mathcal{U}$ in the loss function (Eq 14), we let

$$\mathcal{U}_{u_+} = \begin{cases} \mathcal{U}_n & , \text{ if } u \in \mathcal{U}_n \\ \mathcal{U}_o & , \text{ if } u \in \mathcal{U}_o \end{cases}, \mathcal{U}_{u_-} = \begin{cases} \mathcal{U}_o & , \text{ if } u \in \mathcal{U}_n \\ \mathcal{U}_n & , \text{ if } u \in \mathcal{U}_o \end{cases}. \quad (17)$$

3.3.2 Density-based graph anomaly detection algorithms.

These algorithms defined a measurement on how suspicious a subgraph is with respect to the size and high density in a large graph, then employed an efficient algorithm scheme (e.g., greedy search) to detect the subgraphs of high suspiciousness. Suppose the detection algorithm finds B dense subgraphs (i.e., “blocks”) $\{\mathcal{B}_i = (\mathcal{U}_{b,i} \subseteq \mathcal{U}, \mathcal{V}_i \subseteq \mathcal{V})\}_{i=1}^B$, where $\mathcal{U}_{b,i}$ and \mathcal{V}_i denote the set of user nodes and item nodes in block \mathcal{B}_i , respectively. $\mathcal{U}_n = \mathcal{U} \setminus \cup_{i=1}^B \mathcal{U}_{b,i}$ denotes the set of normal nodes that have never participated in dense blocks. For block \mathcal{B}_i , we denote the size by $n_i = |\mathcal{U}_{b,i}|$ and $m_i = |\mathcal{V}_i|$; we denote the number of ratings in the block by $c_i = |\{r_{u,v} > 0 | u \in \mathcal{U}_{b,i}, v \in \mathcal{V}_i\}|$. The suspiciousness score is defined in different ways in different approaches:

- *Average Degree (AD)* [4, 5]: $\text{susp}^{AD}(\mathcal{B}_i) = \frac{c_i}{n_i}$;
- *Singular Value (SV)* [44, 47]: $\text{susp}^{SV}(\mathcal{B}_i) = \frac{c_i}{\sqrt{n_i \cdot m_i}}$;
- *Kullback–Leibler divergence of Density (KL)* [28]:

$$\text{susp}^{KL}(\mathcal{B}_i) = n_i \cdot m_i \cdot D_{KL}(\rho_i \| p); \quad (18)$$

where $\rho_i = \frac{c_i}{n_i \cdot m_i}$ is the block density, $p = \frac{|\{r_{u,v} > 0 | u \in \mathcal{U}, v \in \mathcal{V}\}|}{|\mathcal{U}| \cdot |\mathcal{V}|}$ is the data density, and $D_{KL}(\rho_i \parallel p) = p - \rho + \rho \log \frac{\rho}{p}$ is the KL divergence between ρ_i and p .

GAL with dense block loss. Suspicious user nodes may perform in multiple blocks. So the graph density loss should be designed based on the following assumptions: (i) user node pairs in the same dense block have similar representations; (ii) pairs of normal nodes have similar representations; (iii) for each suspicious user u , the representations of normal nodes and the nodes in blocks that do not include u are dissimilar with u 's representation.

Therefore, considering the possibility of one user occurring in multiple dense blocks, when sampling for each user node $u \in \mathcal{U}$ in the loss function (Eq 14), we have

$$\mathcal{U}_{u+} = \begin{cases} \mathcal{U}_n & , \text{ if } u \in \mathcal{U}_n \\ \bigcup_{i=1, u \in \mathcal{U}_{b,i}}^B \mathcal{U}_{b,i} & , \text{ if } u \notin \mathcal{U}_n \end{cases}, \quad (19)$$

$$\mathcal{U}_{u-} = \begin{cases} \bigcup_{i=1}^B \mathcal{U}_{b,i} & , \text{ if } u \in \mathcal{U}_n \\ \mathcal{U} \setminus \bigcup_{i=1, u \in \mathcal{U}_{b,i}}^B \mathcal{U}_{b,i} & , \text{ if } u \notin \mathcal{U}_n \end{cases}. \quad (20)$$

4 EXPERIMENTS

4.1 Experimental Settings

4.1.1 Datasets.

We evaluate our proposed GAL with two real-world datasets: **Bitcoin-Alpha** is a trust network of Bitcoin trading users on the Alpha platform [38], where each edge indicates a rating from one user to another with a rating score. **Tencent-Weibo** is a user-posts-hashtag graph from a micro-blogging platform [27].

4.1.2 Baseline methods and GAL variants.

Unsupervised dense block detection methods.

- **FRAUDAR** [26]: it catches suspicious dense subgraphs with theoretical bounded densities for camouflage;
- **CATCHSYNC** [29]: it captures the synchronized behavior patterns in rating networks and social networks;
- **LOCKINFER** [30]: it uses singular vectors of adjacency matrix to find anomalous groups of users in spectral subspaces.

Unsupervised graph outlier detection methods.

- **FRAUDAR_R**: it is the reverse of **FRAUDAR**. We use **FRAUDAR** to detect first several dense blocks until the remaining density is very low and report the remaining users as outliers;
- **LOCKINFER_R**: it is the reverse of **LOCKINFER**. We use **LOCKINFER** to detect all user groups in spectral subspace and report users that are not contained in any group as outliers;
- **DBSCAN** [36]: it groups together points that are close in space and then finds the outliers that lie alone in low-density regions;
- **FRAUDEAGLE** [1]: it uses a belief propagation-based algorithm to give a fraud score to each user. Outlying users who behaves more different than the majority are given higher fraud scores;
- **BIRDNEST** [24]: it ranks users by a Bayesian model with the users' rating timestamps and rating score distribution;
- **REV2** [37]: it uses an iterative algorithm to find unfair users whose behaviors can be considered outlier compared with the majority.

Unsupervised graph embedding methods.

- **NODE2VEC** [17]: it uses biased random walks to capture the homophily and local structure information of the network. Hyperparameters $p, q \in \{0.25, 0.5, 1, 2, 4\}$ control the search bias;
- **LINE** [50]: it uses 1st order and 2nd order proximity to capture the local and 2-hop structure of the network via edge sampling;
- **BiNE** [16]: it learns the representations of vertices in a bipartite network. Biased random walks are conducted to preserve the long-tail distribution of vertices.

Unsupervised GNN methods.

- **GCN** [33]: it is a spectral-based GNN that learns node embeddings via a localized first-order approximation of spectral graph convolutions. It uses a unsupervised RW-based loss function;
- **GRAPHSAGE** [19]: it is a GNN that enables specifying different weights to different nodes in a neighborhood. It uses a unsupervised RW-based loss function;
- **GAT** [51]: it is a GNN that learns node embeddings inductively from its own feature and the aggregated features of its neighbors. It uses a unsupervised RW-based loss function;
- **DEEFD** [52]: it is an encoder-decoder structured deep neural network model for fraud detection. It learns user embeddings by minimizing the difference of pairwise distance between user embeddings and a modified Jaccard similarity;
- **DOMINANT** [11]: it is a graph auto-encoder based deep neural network model for graph anomaly detection. It reconstructs the graph structure and node attributes to find the anomaly nodes.

In summary, we will compare among **17** methods, including 3 dense block detection methods, 6 graph outlier detection methods, 3 graph embedding methods, and 5 unsupervised GNN methods. We also compare GAL itself among **10** variants: 3 with density losses (+FRAUDAR, +CATCHSYNC, +LOCKINFER), 5 with outlier losses (+FRAUDAR_R, +LOCKINFER_R, +FRAUDEAGLE, +BIRDNEST, +REV2).

4.2 Results on Bitcoin-Alpha

Table 1 presents the performance of our GAL with different anomaly losses and all baselines for the task of anomaly detection on the Bitcoin-Alpha dataset.

- *The suspicious behaviors on Bitcoin-Alpha are more likely to form outliers than dense subgraphs.* Most of the graph outlier detection methods perform better than dense subgraph detection methods. For example, **FRAUDAR_R** achieved an F1 of 0.6733, while **CATCHSYNC** had an F1 of 0.5616.
- *Graph representation learning models, including GNNs, combine both node feature and graph structural information, but the improvement by those with RW-based loss is not significant.* For example, **NODE2VEC** achieved an F1 of 0.6977 and **FRAUDAR_R** achieved an F1 of 0.6733. The reason is that the outlier's local neighbors may not be outliers. **DEEFD** only obtained an F1 of 0.5479, because it relied heavily on the Jaccard similarity to optimize node embeddings, but the Jaccard similarity only used neighbor's information.
- *GNNs trained by our GAL, graph outlier losses, outperform all baseline methods.* GAL-FRAUDAR_R achieved the best performance: an F1 of 0.7568, an AP of 0.8221, and an AUC of 0.8556. It outperformed GCN relatively by **+8.1%** and **+14.6%** on the two metrics. And it

Table 1: GNNs with graph outlier losses perform the best on Bitcoin-Alpha data.

	Precision	Recall	F1	AUC
Graph outlier detection algorithms				
DBSCAN [36]	0.3879	1.0000	0.5589	0.5016
FRAUDAR_R	0.5714	0.8193	0.6733	0.7659
LOCKINFER_R	0.5512	0.8434	0.6667	0.7593
FRAUDEAGLE [1]	0.4205	0.8916	0.5714	0.5149
BIRDNEST [24]	0.3971	1.0000	0.5685	0.3756
REV2 [37]	0.6471	0.6627	0.6548	0.7094
Raw node features				
MLP	0.8667	0.5652	0.6842	0.7239
Graph embedding methods				
NODE2VEC [17]	0.6522	0.7500	0.6977	0.7606
LINE [50]	0.4722	0.8500	0.6071	0.6455
BiNE [16]	0.5143	0.9000	0.6545	0.7409
Graph neural network methods (GNNs)				
GCN [33]	0.5526	0.9545	0.7000	0.7463
GRAPHSAGE [19]	0.6471	0.6111	0.6286	0.7238
GAT [51]	0.6071	0.7727	0.6800	0.7199
DEEPFD [52]	0.3774	1.0000	0.5479	0.4652
DOMINANT [11]	0.4231	1.0000	0.5946	0.3966
GRAPHSAGE + dense block losses (GAL)				
+FRAUDAR	0.7000	0.7368	0.7179	0.8282
+CATCHSYNC	0.7895	0.6818	0.7317	0.7742
+LOCKINFER	0.7143	0.7500	0.7317	0.7970
GRAPHSAGE + graph outlier losses (GAL)				
+FRAUDAR_R	0.7368	0.7778	0.7568	0.8556
+LOCKINFER_R	0.6800	0.7727	0.7234	0.7874
+FRAUDEAGLE	0.9231	0.6000	0.7273	0.8303
+BIRDNEST	0.7368	0.6364	0.6829	0.7683
+REV2	0.7500	0.7500	0.7500	0.8030
Other GNNs + FRAUDAR_R (GAL)				
GCN+GAL	0.7143	0.7895	0.7500	0.8498
GAT+GAL	0.6333	0.8636	0.7308	0.7273

outperforms the best anomaly detection method FRAUDAR_R relatively by **+12.4%** and **+11.7%**. GAL with graph outlier losses can train the GNNs more effectively.

4.3 Results on Tencent-Weibo

Table 2 presents the performances on Tencent-Weibo dataset. Since the labels are seriously biased, F1 is more representative than AUC. BIRDNEST could not work due to the absence of timestamps.

- *The suspicious behaviors on Tencent-Weibo are more likely to form dense subgraphs than outliers.* Dense subgraph detection methods perform much better than outlier detection methods. FRAUDAR achieved an F1 of 0.7540, while FRAUDEAGLE only made an F1 of 0.4102. The reason is that fraudsters had to post a large number of messages in group to inflate popularity of hashtag.
- *Local neighborhood is more informative than pure global structure, as graph embedding models perform better than the dense subgraph detection algorithms.* For example, LINE achieved an F1 of 0.8105,

Table 2: GNNs with dense block losses perform the best on Tencent-Weibo data.

	Precision	Recall	F1	AUC
Dense block detection algorithms				
FRAUDAR [26]	0.9624	0.6198	0.7540	0.9079
CATCHSYNC [29]	0.3200	0.8111	0.4589	0.7752
LOCKINFER [30]	0.9318	0.4562	0.6125	0.8674
Raw node features				
MLP	0.5105	0.7711	0.6143	0.8907
Graph embedding methods				
NODE2VEC [17]	0.9111	0.7218	0.8055	0.9642
LINE [50]	0.8675	0.7606	0.8105	0.9432
BiNE [16]	0.3678	0.7007	0.4824	0.8774
Graph neural network methods (GNNs)				
GCN [33]	0.8373	0.8697	0.8532	0.9690
GRAPHSAGE [19]	0.8200	0.8662	0.8425	0.9780
GAT [51]	0.8586	0.8768	0.8676	0.9706
DEEPFD [52]	0.1080	0.9562	0.1941	0.4980
DOMINANT [11]	0.3006	0.8627	0.4524	0.8167
GRAPHSAGE + dense block losses (GAL)				
+FRAUDAR	0.8669	0.8944	0.8804	0.9777
+CATCHSYNC	0.9067	0.8556	0.8804	0.9685
+LOCKINFER	0.9294	0.8803	0.9042	0.9843
GRAPHSAGE + graph outlier losses (GAL)				
+FRAUDAR_R	0.8547	0.8697	0.8621	0.9697
+LOCKINFER_R	0.8591	0.8803	0.8696	0.9757
+FRAUDEAGLE	0.8581	0.8732	0.8656	0.9730
+BIRDNEST	n/a	n/a	n/a	n/a
+REV2	0.8759	0.8944	0.8850	0.9677
Other GNNs + LOCKINFER (GAL)				
GCN+GAL	0.9459	0.8627	0.9024	0.9853
GAT+GAL	0.9097	0.8873	0.8984	0.9838

while FRAUDAR achieved an F1 of 0.7540. The models that use local structures for embedding aggregation can preserve the user similarity of being in the same blocks.

- *GNNs trained by our GAL, dense block losses, outperform all baseline methods.* GAL-LOCKINFER performed the best: an F1 of 0.9042 and an AUC of 0.9843. It outperformed the best graph embedding method LINE relatively by **+11.6%** and **+4.4%** on the two metrics. And it outperformed GRAPHsAGE relatively by **+7.3%** and **+0.6%**. GAL with dense block losses can train the GNN models more effectively.

5 CONCLUSIONS

In this work, we presented a novel Graph Anomaly Losses (GAL) that is able to unsupervisedly train GNNs for anomaly-detectable node representations. GAL has a bounded test error and GAL with *graph outlier losses* and *dense block losses* evaluates node similarity based on the global properties discovered by graph mining algorithms. Experiments on two real-world datasets demonstrated that (i) GNNs with GAL significantly outperformed 17 baseline methods and (ii) GAL trained the models more effectively than traditional RW-based loss on various state-of-the-art GNN frameworks.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation IIS-1849816.

REFERENCES

- [1] Leman Akoglu, Rishi Chandy, and Christos Faloutsos. 2013. Opinion fraud detection in online reviews by network effects. In *ICWSM*.
- [2] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. 2010. Oddball: Spotting anomalies in weighted graphs. In *PAKDD*. 410–421.
- [3] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: a survey. *DMKD* 29, 3 (2015), 626–688.
- [4] Reid Andersen. 2010. A local algorithm for finding dense subgraphs. *TALG* 6, 4 (2010), 60.
- [5] Yuichi Asahiro, Kazuo Iwama, Hisao Tamaki, and Takeshi Tokuyama. 2000. Greedily finding a dense subgraph. *Journal of Algorithms* 34, 2 (2000), 203–221.
- [6] Sambaran Bandyopadhyay, Saley Vishal Vivek, and MN Murty. 2020. Outlier Resistant Unsupervised Deep Architectures for Attributed Network Embedding. In *WSDM*. 25–33.
- [7] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. *arXiv preprint arXiv:1906.07413* (2019).
- [8] Deepayan Chakrabarti. 2004. Autopart: Parameter-free graph partitioning and outlier detection. In *EC-DMKD*. 112–124.
- [9] Nitesh V Chawla. 2003. C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *ICML*, Vol. 3. 66.
- [10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *JAIR* 16 (2002), 321–357.
- [11] Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. 2019. Deep anomaly detection on attributed networks. In *SDM*. SIAM, 594–602.
- [12] Kaize Ding, Yichuan Li, Jundong Li, Chenghao Liu, and Huan Liu. 2019. Graph Neural Networks with High-order Feature Interactions. *arXiv preprint arXiv:1908.07110* (2019).
- [13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *KDD*, Vol. 96. 226–231.
- [14] Dhivyaa Eswaran, Christos Faloutsos, Sudipto Guha, and Nina Mishra. 2018. Spotlight: Detecting anomalies in streaming graphs. In *KDD*. ACM, 1378–1386.
- [15] Jing Gao, Feng Liang, Wei Fan, Chi Wang, Yizhou Sun, and Jiawei Han. 2010. On community outliers and their efficient detection in information networks. In *KDD*. 813–822.
- [16] Ming Gao, Leihui Chen, Xiangnan He, and Aoying Zhou. 2018. BiNE: Bipartite Network Embedding.. In *SIGIR*. 715–724.
- [17] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *KDD*. ACM, 855–864.
- [18] Hongyu Guo and Henna L Viktor. 2004. Learning from imbalanced data sets with boosting and data generation. *KDD Explorations* (2004).
- [19] Will Hamilton, Zitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeurIPS*. 1024–1034.
- [20] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584* (2017).
- [21] Ville Hautamaki, Ismo Karkkainen, and Pasi Franti. 2004. Outlier detection using k-nearest neighbour graph. In *ICPR*, Vol. 3. 430–433.
- [22] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IJNN*. 1322–1328.
- [23] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *TKDE* 21, 9 (2009), 1263–1284.
- [24] Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. 2016. Birdnest: Bayesian inference for ratings-fraud detection. In *SDM*. 495–503.
- [25] Bryan Hooi, Kijung Shin, Hyun Ah Song, Alex Beutel, Neil Shah, and Christos Faloutsos. 2017. Graph-based fraud detection in the face of camouflage. *TKDD* 11, 4 (2017), 44.
- [26] Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. 2016. Fraudar: Bounding graph fraud in the face of camouflage. In *KDD*. 895–904.
- [27] Meng Jiang. 2016. Catching Social Media Advertisers with Strategy Analysis. In *Workshop on CyberSafety*. 5–10.
- [28] Meng Jiang, Alex Beutel, Peng Cui, Bryan Hooi, Shiqiang Yang, and Christos Faloutsos. 2016. Spotting suspicious behaviors in multimodal data: A general metric and algorithms. *TKDE* 28, 8 (2016), 2187–2200.
- [29] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. 2014. Catchsync: catching synchronized behavior in large directed graphs. In *KDD*. 941–950.
- [30] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. 2014. Inferring strange behavior from connectivity pattern in social networks. In *PAKDD*. 126–138.
- [31] Parisa Kaghazgaran, James Caverlee, and Anna Squicciarini. 2018. Combating crowdsourced review manipulators: A neighborhood-based approach. In *WSDM*. 306–314.
- [32] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. 2009. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NeurIPS*. 793–800.
- [33] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [34] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).
- [35] Vladimir Koltchinskii, Dmitry Panchenko, et al. 2002. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics* 30, 1 (2002), 1–50.
- [36] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. 2011. Density-based clustering. *DMKD* 1, 3 (2011), 231–240.
- [37] Srikanth Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and VS Subrahmanian. 2018. Rev2: Fraudulent user prediction in rating platforms. In *WSDM*. ACM, 333–341.
- [38] Srikanth Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. 2016. Edge weight prediction in weighted signed networks. In *ICDM*. 221–230.
- [39] Chong Li, Kunyang Jia, Dan Shen, CJ Shi, and Hongxia Yang. 2019. Hierarchical representation learning for bipartite graphs. In *IJCAI*. AAAI Press, 2873–2879.
- [40] David Mease, Abraham J Wyner, and Andreas Buja. 2007. Boosted classification trees and class probability/quantile estimation. *JMLR* 8, Mar (2007), 409–439.
- [41] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutikov. 2016. Learning convolutional neural networks for graphs. In *ICML*. 2014–2023.
- [42] Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. 2018. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *KDD*. 2041–2050.
- [43] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *KDD*. ACM, 701–710.
- [44] B Aditya Prakash, Ashwin Sridharan, Mukund Seshadri, Sridhar Machiraju, and Christos Faloutsos. 2010. Eigenspokes: Surprising patterns and scalable community chipping in large graphs. In *PAKDD*. 435–448.
- [45] Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *KDD*. ACM, 985–994.
- [46] Neil Shah. 2017. FLOCK: Combating astroturfing on livestreaming platforms. In *WWW*. 1083–1091.
- [47] Neil Shah, Alex Beutel, Brian Gallagher, and Christos Faloutsos. 2014. Spotting suspicious link behavior with fbox: An adversarial perspective. In *ICDM*. 959–964.
- [48] Kijung Shin, Bryan Hooi, and Christos Faloutsos. 2016. M-zoom: Fast dense-block detection in tensors with quality guarantees. In *ECML-PKDD*. 264–280.
- [49] Jimeng Sun, Huiming Qu, Deepayan Chakrabarti, and Christos Faloutsos. 2005. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM*. IEEE, 8.
- [50] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *WWW*. 1067–1077.
- [51] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [52] Haibo Wang, Chuan Zhou, Jia Wu, Weizhen Dang, Xingquan Zhu, and Jilong Wang. 2018. Deep structure learning for fraud detection. In *ICDM*. 567–576.
- [53] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How Powerful are Graph Neural Networks? *arXiv preprint arXiv:1810.00826* (2018).
- [54] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *KDD*. ACM, 974–983.
- [55] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *KDD*. 793–803.
- [56] Tong Zhao, Matthew Malir, and Meng Jiang. 2018. Actionable objective optimization for suspicious behavior detection on large bipartite graphs. In *BigData*. IEEE, 1248–1257.