GNN-based Graph Anomaly Detection with Graph Anomaly Loss

Tong Zhao\textsuperscript{1}, Chuchen Deng\textsuperscript{1}, Kaifeng Yu\textsuperscript{1}, Tianwen Jiang\textsuperscript{1,2}, Daheng Wang\textsuperscript{1}, Meng Jiang\textsuperscript{1}

\textsuperscript{1}Department of Computer Science and Engineering, University of Notre Dame, USA
\textsuperscript{2}Research 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.

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 GraphSAGE [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. BtNE [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. RV2 [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. rBox [47] located mini-scale attacks missed by spectral techniques. CatchSync [29] and CrossSpot [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 $\mathcal{R} = \{r_{uv}|u \in \mathcal{U}, v \in \mathcal{V}\}$, where $r_{uv}$ 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 $\{x_u \in \mathbb{R}^{d_x}, \forall u \in \mathcal{U}\}$ (where $d_x$ is the dimension of raw
features), find a mapping function of the representations of user nodes $f: u \in \mathcal{U} \rightarrow z_u \in \mathbb{R}^d$ 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 this paper, we use to aggregate embeddings from immediate neighbors, we aggregate 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, $z_u, \forall u \in \mathcal{U}$, and to tune the weight matrices $W^u, \forall k \in \{1, \ldots, K\}$ 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]:

$$
L_{rw}(u) = \mathbb{E}_{u \sim \mathcal{U}_u \cup \ldots \cup \mathcal{U}_u} \log \left( 0, z_{u}^Tw_{u} - z_{u}^Tw_{u} + \Delta \right),
$$

where $\Delta$ 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 $\mathcal{U}_u \cup \ldots \cup \mathcal{U}_u$ 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 mentioned before, it is neither proper nor effective when the task of anomaly detection by preserving both user embeddings’ 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 this paper, we use to aggregate embeddings from immediate neighbors, we aggregate 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, $z_u, \forall u \in \mathcal{U}$, and to tune the weight matrices $W^u, \forall k \in \{1, \ldots, K\}$ 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]:

$$
L_{rw}(u) = \mathbb{E}_{u \sim \mathcal{U}_u \cup \ldots \cup \mathcal{U}_u} \log \left( 0, z_{u}^Tw_{u} - z_{u}^Tw_{u} + \Delta \right),
$$

where $\Delta$ 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 $\mathcal{U}_u \cup \ldots \cup \mathcal{U}_u$ 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 characteristics. However, on a bipartite graph, this assumption does not hold as neighbors are different types of nodes. Hence in this paper, we use to aggregate embeddings from immediate neighbors, we aggregate 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, $z_u, \forall u \in \mathcal{U}$, and to tune the weight matrices $W^u, \forall k \in \{1, \ldots, K\}$ 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]:

$$
L_{rw}(u) = \mathbb{E}_{u \sim \mathcal{U}_u \cup \ldots \cup \mathcal{U}_u} \log \left( 0, z_{u}^Tw_{u} - z_{u}^Tw_{u} + \Delta \right),
$$

where $\Delta$ 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 $\mathcal{U}_u \cup \ldots \cup \mathcal{U}_u$ 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 $P(u|y_u)$ is the same at training and testing. Then let $P_{y_u}$ denote the class-conditional distribution, that is, $P_y = (u|y_u = j)$. 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 $u'$:

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

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

$$
L_{bal}[g] = \mathbb{E}_{(u, j) \sim P_{y_u}} \left[ \min_{y_u \neq j, y_u = j} \max g(u, u_j) \right].
$$

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

$$
L_j[g] = \mathbb{E}_{u \sim P_{y_u}} \left[ \min_{y_u \neq j} \max g(u, u_j) \right].
$$

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:

$$
y_j = \min_{u \in S_j} \left( \min_{y_u \neq j} \max g(u, u_j) - \max g(u, u_j) \right),
$$

where $y_{min} = \min\{y_1, \ldots, y_j\}$ is the widely used training margin in previous studies [35]. Then we let $L_j, y_j$ denote the margin for class $j$ when training:

$$
L_j[y], y_j = \mathbb{E}_{u \sim P_{y_u}} \left[ \min_{y_u \neq j} \max g(u, u_j) - \max g(u, u_j) \right],
$$

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

$$
\hat{R}_j(G) = \frac{1}{n_j} \mathbb{E}_y \sup_{g \in G} \sum_{u \in S_j} \sigma_u \left[ \min_{y_u \neq j} g(u, u_j) - \max g(u, u_j) \right],
$$

where $\sigma$ 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 $y_1, y_2, \ldots, y_k > 0$, all hypotheses $g \in G$ will have balanced-class generalization bounded by:

$$
L_{bal}[g] \leq \frac{1}{k} \sum_{j=1}^k \hat{L}_j(y), y_j + \frac{4}{y_j} \hat{R}_j(G), y_j + \epsilon_j(y_j),
$$

where $\epsilon_j(y_j) \leq \sqrt{\log \log \left( \frac{2 \max_{u \in S_j} \log \phi(u, y_j)|\epsilon_j(y_j)|}{n_j} \right) \log \frac{\delta}{n_j}} \leq \frac{\delta}{n_j}$ is typically a low-order term in $n_j$. Concretely, the Rademacher complexity $\hat{R}_j(G)$ will typically scale as $\sqrt{\log \frac{C(G)}{n_j}}$ for some complexity measure $C(G)$, in which case:

$$
L_{bal}[g] \leq \frac{1}{k} \sum_{j=1}^k \hat{L}_j(y), y_j + \frac{4}{y_j} \hat{R}_j(G), y_j + \epsilon_j(y_j),
$$

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 $C(G)$ and low order term $\epsilon_j(y_j)$) [7]

$$
\frac{1}{y_1 \sqrt{y_1}} + \frac{1}{y_2 \sqrt{y_2}}
$$

Although the it is hard to get the optimal margins with the above equation as they are complicate functions of the parameters in $g(x)$, we can figure out the relative scales between the two margins.
Suppose we have \( y_1^*, y_2^* > 0 \) that minimize the equation above, we observe that any \( y_1^* = y_1^* - \delta \) and \( y_2^* = y_2^* + \delta \) (where \( -2 < \delta < y_1^* \)) can be realized by the same parameters with a shifted bias term. Therefore, for \( y_1^*, y_2^* \) to be optimal, the following inequality must be satisfied [7],

\[
\frac{1}{y_1^* \sqrt{n_1}} + \frac{1}{y_2^* \sqrt{n_2}} \leq \frac{1}{(y_1^* - \delta) \sqrt{n_1}} + \frac{1}{(y_2^* + \delta) \sqrt{n_2}},
\]

which implies that

\[
y_1^* \propto n_1^{-1/4}, \quad \text{and} \quad y_2^* \propto n_2^{-1/4}.
\]

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

\[
\mathcal{L}(u) = \max \{0, \max_{y \neq y_u} g(u, y) - \min_{y \neq y_u} g(u, y) + \Delta_{y_u}\},
\]

where \( \Delta_{y_u} = \frac{C}{n_{y_u}} \).

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,

\[
\mathcal{L}(u) = \mathbb{E}_{u \sim \mathcal{U}, \beta \sim \mathbb{U}_u} \max \{0, g(u, \beta) - g(u, u_+) + \Delta_{y_u}\},
\]

where \( \Delta_{y_u} = \frac{C}{n_{y_u}} \).

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 \( \mathcal{U}_u \) and \( \mathcal{U}_u \) 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 outlier score of node \( u \) based on a pair of its particular features \( a_u \) and \( b_u \) [2, 29]:

\[
\text{outlieriness}(u) = |b_u - \hat{b}_u| \text{ or } \frac{\max(b_u, \hat{b}_u)}{\min(b_u, \hat{b}_u)} \cdot \log(|b_u - \hat{b}_u| + 1),
\]

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)” Akgulcu 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., \( b_u \propto a_u^\gamma \), \( \gamma \) 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., \( \text{sync} \propto \alpha \cdot \text{norm}^2 + \beta \), \( \alpha \) and \( \beta \) 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]:

\[
\text{outlierness}(u) = 1 - \frac{\sum_{v \neq u} \hat{R}(u,v) + \alpha_1 \mu_f + \alpha_2 \mu_i}{|N(u)| + |\mathcal{V}_
u(u)| + |\mathcal{V}_i(u)|},
\]

where \( \hat{R}(u,v) \) is the normality score of the rating from \( u \) to item \( v \), \( \mu_f \) is the prior belief of \( u \)'s normality score given by BirnNEst [24]. \( |N(u)| \) is \( u \)'s behavior normality score. \( \alpha_1 \) and \( \alpha_2 \) 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].

### 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 \( B_i \) dense subgraphs (i.e., “blocks”) \( \{B_i \subseteq \mathcal{U}, |V_i| \leq |\mathcal{V}_i|\}^{|B_i|}_{i=1} \), where \( \mathcal{U}_i \) and \( |V_i| \) denote the set of user nodes and item nodes in block \( B_i \), respectively. \( \mathcal{U}_i = \mathcal{U} \setminus \cup_{j \neq i}^{B_i} \mathcal{U}_j \) denotes the set of nodes that have never participated in dense blocks. For block \( B_i \), we denote the size by \( n_i = |\mathcal{U}_i| \) and \( m_i = |\mathcal{V}_i| \), we denote the number of ratings in the block by \( c_i = |\{(u,v) : u \in \mathcal{U}_i, v \in \mathcal{V}_i\}| \). The suspiciousness score is defined in different ways in different approaches:

- **Average Degree (AD)** [4, 5]: \( \text{adj}B_i = \frac{|E_i|}{m_i} \).
- **Singular Value (SV)** [44, 47]: \( \text{sv}B_i = \frac{\lambda_i}{m_i} \).
- **Kullback–Leibler divergence of Density (KL)** [28]:

\[
\text{KL}B_i = n_i \cdot m_i \cdot D_{KL}(\rho_i \parallel p).
\]
where $\rho_i = \frac{k_i}{m_i}$ is the block density, $p = \frac{|\{u \in U \mid u \notin V\}|}{|U||v|}$ is the data density, and $D_{KL}(\rho_i \parallel \rho) = p - \rho + \rho \log \frac{p}{\rho}$ is the KL divergence between $\rho_i$ and $\rho$.

**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 $u \in U$ in the loss function (Eq 14), we have

$$U_{u+} = \begin{cases} U_n \cup \bigcup_{i=1}^B U_b, & \text{if } u \in U_n \\ U_b, & \text{if } u \notin U_n \end{cases},$$

$$U_{u-} = \begin{cases} U_n \setminus \bigcup_{i=1}^B U_b, & \text{if } u \in U_n \\ U_b, & \text{if } u \notin U_n \end{cases}.$$  

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-post-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]: captures the synchronized behavior patterns in rating networks and social networks;
- **LockInfer** [30]: 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 only the 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 behave 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.**
- **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** [31]: 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;
- **DeepFD** [32]: it is a 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. DeepFD 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
outperforms the best anomaly detection method \texttt{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. \texttt{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. \texttt{Fraudar} achieved an F1 of 0.7540, while \texttt{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, \texttt{LINE} achieved an F1 of 0.8105, while \texttt{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-\texttt{LockInfer} performed the best: an F1 of 0.9042 and an AUC of 0.9843. It outperformed the best graph embedding method \texttt{LINE} relatively by +11.6% and +4.4% on the two metrics. And it outperformed \texttt{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