# Error-bounded Graph Anomaly Loss for GNNs

Tong Zhao, Chuchen Deng, Kaifeng Yu, Tianwen Jiang, Daheng Wang, Meng Jiang

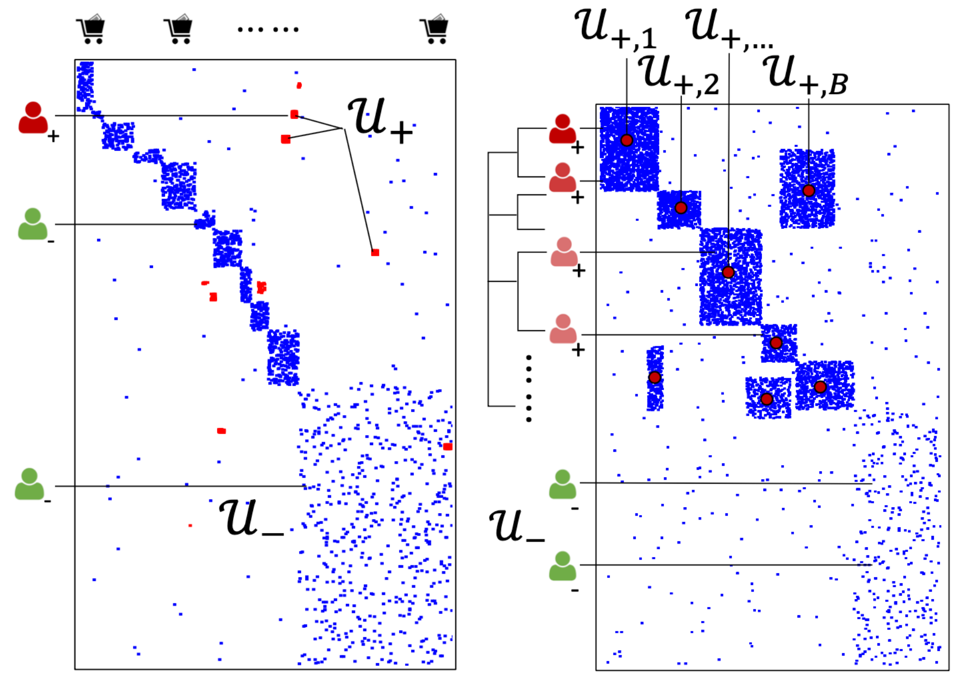
Department of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN 46556, USA

UNIVERSITY OF NOTRE DAME
College of Engineering

NSF

# Graph Anomaly Detection

- Graph Outliers.
  - Fake/unfair reviewers on reviewing networks, etc.
- Dense Blocks.
  - Spammers/bot nets on social networks, etc.




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

# How things were done

- Unsupervised graph mining methods:
  - Graph outlier detection:
    - Feature-based
    - Structure-based
    - Model-based
  - Dense block detection:
    - Average degree
    - Singular value
- Graph neural networks (GNNs) with unsupervised loss function for representation learning.

# Random walk-based loss function

- Basic assumption:
  - Nodes that are closer to each other in the graph should have similar representations.
- With positive/negative sampling, can be formulated as:

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

# Random walk-based loss function

- Basic assumption:
  - Nodes that are closer to each other in the graph should have similar representations.
- With positive/negative sampling, can be formulated as:

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

- Problem:
  - The assumption doesn't always hold true when the representations are used for anomaly detection.

# Graph Anomaly Loss (GAL)

- Goal: design an unsupervised loss function that can learn node representations that are optimized for anomaly detection.
- Desired abilities:
  - Give anomalous nodes similar representations.
  - Utilize the global information given by graph mining methods.
  - Capable of dealing with severely imbalanced data.

# Graph Anomaly Loss (GAL)

- Goal: design an unsupervised loss function that can learn node representations that are optimized for anomaly detection.
- Desired abilities:
  - Give anomalous nodes similar representations.
  - Utilize the global information given by graph mining methods.
  - Capable of dealing with severely imbalanced data.

$$\mathcal{L}(u) = \mathbb{E}_{u_+ \sim \mathcal{U}_{u+}, u_- \sim \mathcal{U}_{u-}} \max\{0, g(u, u_-) - g(u, u_+) + \Delta_{y_u}\}, \quad \text{where } \Delta_{y_u} = \frac{C}{n_{y_u}^{1/4}}. \tag{14}$$

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 $n_{y_u} = |\mathcal{U}_{u+}|$.

# GAL with graph outlier loss

- Existing outlier detection methods can give us:
  - $\mathcal{U}_u$: set of normal nodes; $\mathcal{U}_o$: set of outlier nodes:
- Goals:
  - Encourage pairs of outlier nodes to have similar representations.
  - Encourage pairs of normal nodes to have similar representations.
  - Enforce the representations of pairs of outlier and normal nodes are distinct.

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

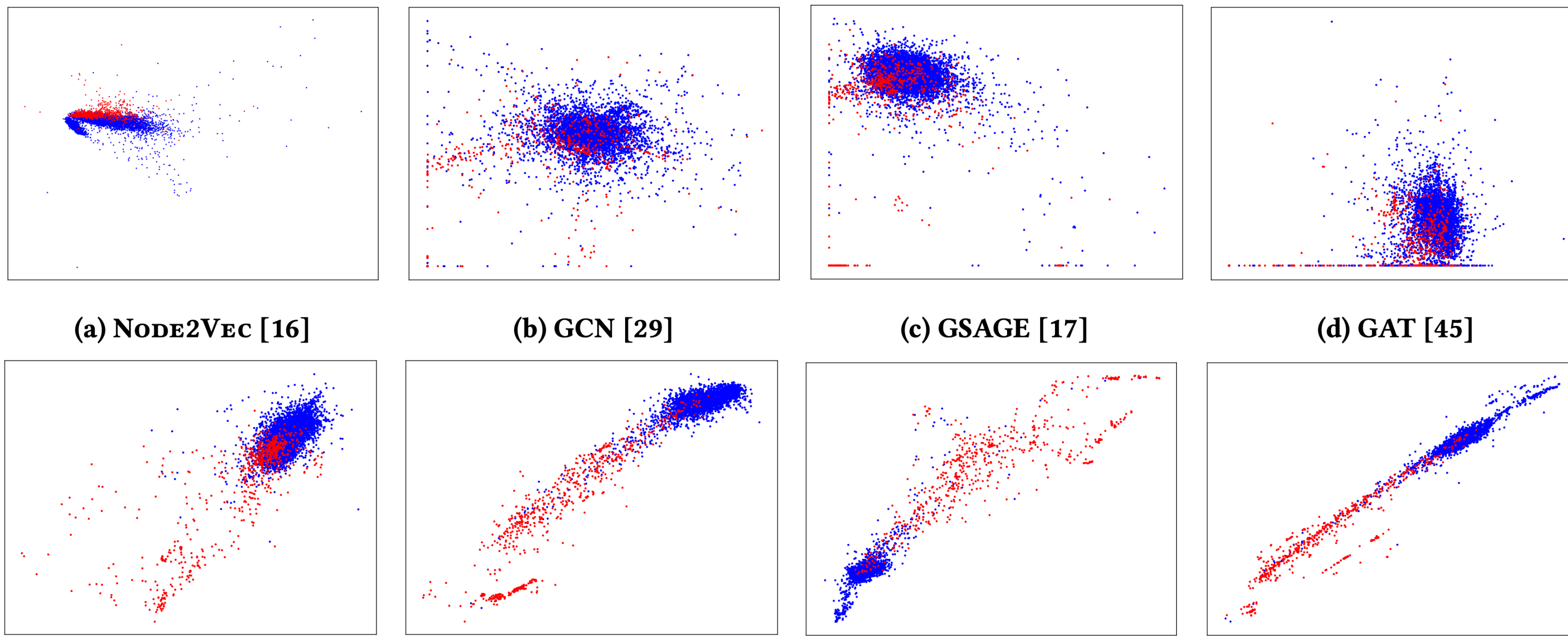
# GAL with dense block loss

- Existing dense block detection methods can give us:
  - $\mathcal{U}_n$: set of normal nodes;
  - Sets of nodes in different dense blocks: $\mathcal{U}_{b,1}, \mathcal{U}_{b,2}, \ldots$
- Goals:
  - Encourage node pairs in the same block to have similar representations.
  - Enforce nodes in different blocks to have distinct representations.

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

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

# Visualization of learned representations



(a) NODE2VEC [16] (b) GCN [29] (c) GSAGE [17] (d) GAT [45]

(e) GSAGE+FRAUDAR (GAL) (f) GSAGE+LOCKINFER (GAL) (g) GCN + LOCKINFER (GAL) (h) GAT + LOCKINFER (GAL)

**Figure 2: Visualizing user embeddings in Tencent-Weibo data. Blue dots represent benign users, and red dots represent anomalous users. Graph anomaly losses (e–h) are better than random-walk loss (b–d).**

Thanks for listening!

Funded by National Science Foundation IIS-1849816.

University of Notre Dame
College of Engineering