HE-GAD: a behavior-enhanced contrastive learning framework for graph anomaly detection

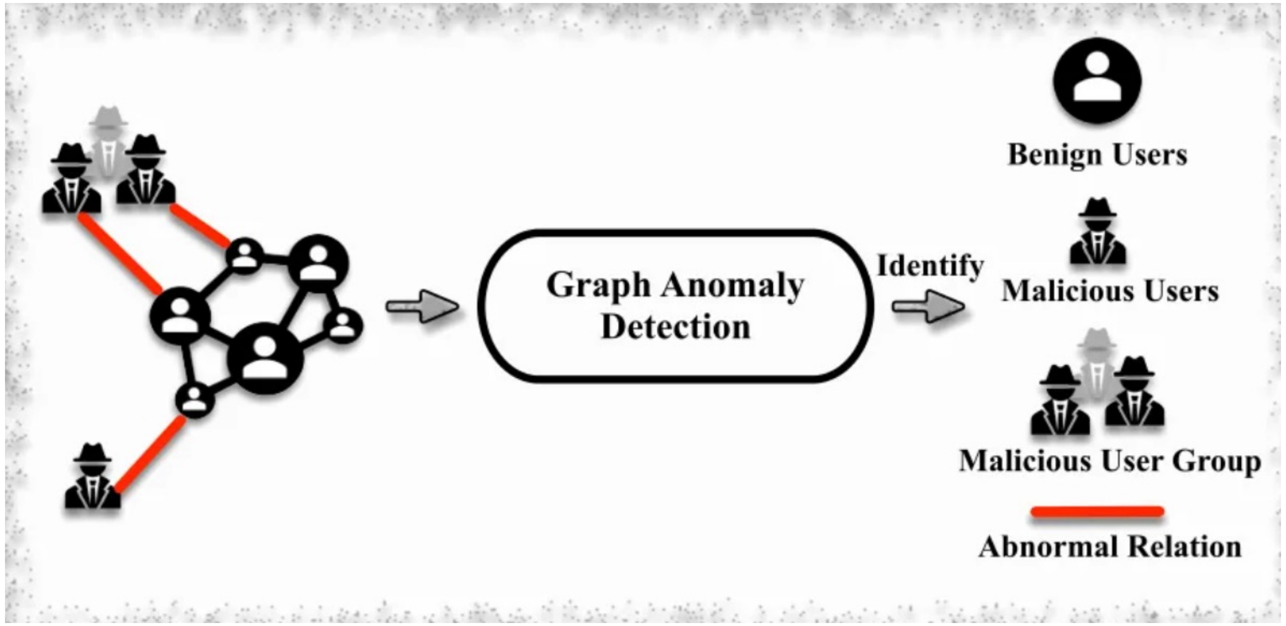
Ling Zheng[1], Qi Song[1,2*], Yihan Wang[1], Zhitao Wang[3], Xiangyang Li[1,2]

Contact: lingzheng@mail.ustc.edu.cn

中国科学技术大学
University of Science and Technology of China

Lab for intelligent network & knowledge engineering

Graph Anomaly Detection(GAD)



Node-Level GAD

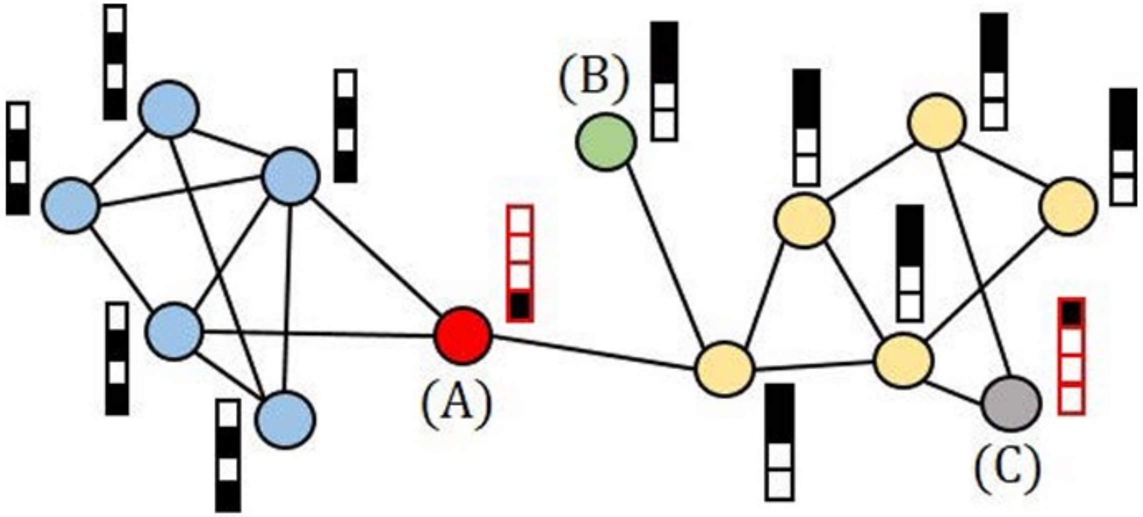
Input: $G = (V, E, X), L(Optional)$

Output: $f: V \rightarrow R$

Ma X, Wu J, Xue S, et al. A comprehensive survey on graph anomaly detection with deep learning[J]. IEEE transactions on knowledge and data engineering, 2021, 35(12): 12012-12038.

Challenges

Diversity of anomalies



Lack of labeled data

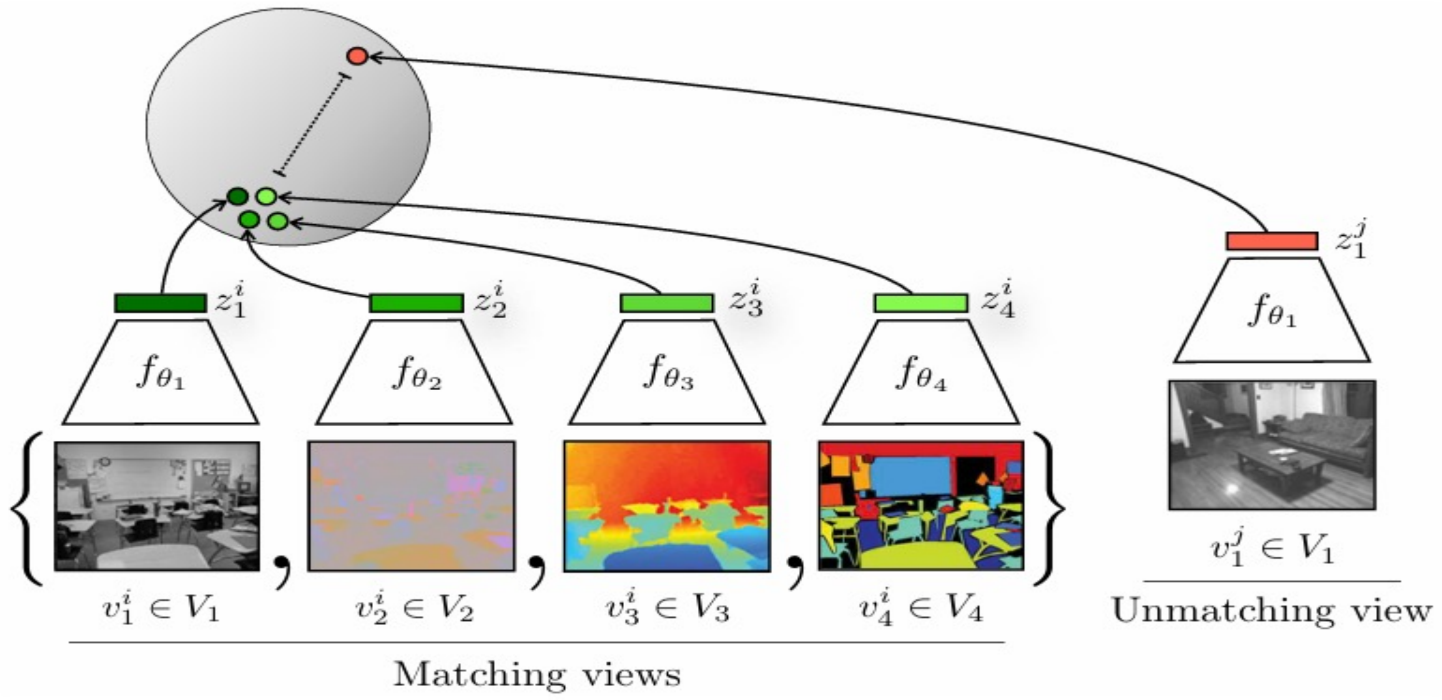
The high cost of manual annotation and the difficulty in ensuring label accuracy

Kim H, Lee B S, Shin W Y, et al. Graph anomaly detection with graph neural networks: Current status and challenges[J]. IEEe Access, 2022, 10: 111820-111829.

Existing Methods

- Supervised Methods:
  - Shallow Methods fail to handle complex graphs.
  - Deep Learning Methods are highly dependent on labeled data.
- Unsupervised Methods:
  - Reconstruction Methods: The consistency of reconstruction error and anomalous degree can not be guaranteed
  - Contrastive Learning Methods: The construction of contrastive pairs risk introducing additional noises.

Insight



Can we construct multi-view contrastive coding that is intrinsic to graphs ?

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Multiview Coding. In Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI. Springer-Verlag, Berlin, Heidelberg, 776–794. https://doi.org/10.1007/978-3-030-58621-8_45

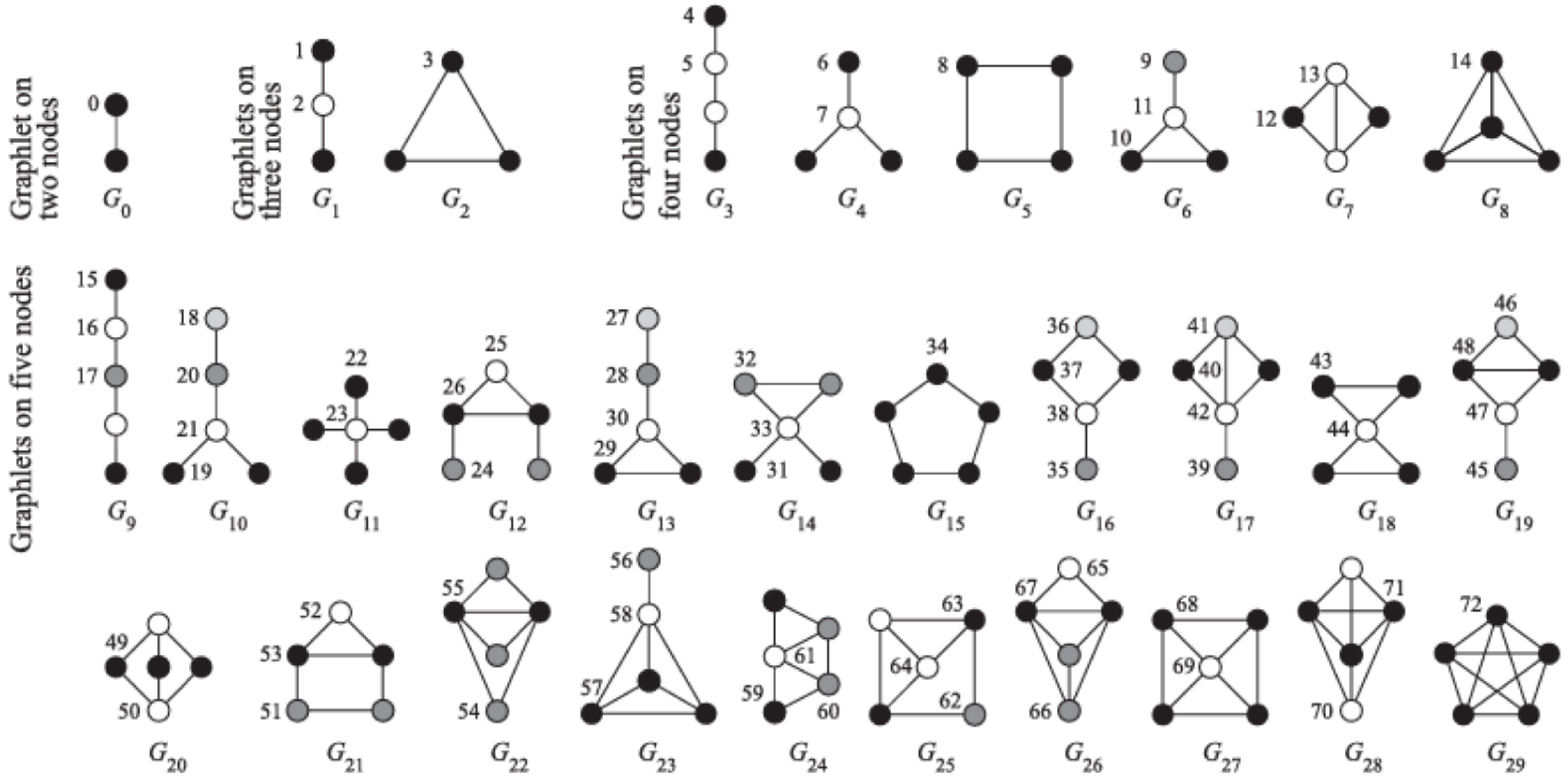
Behavior View

- Graphlet Degree Vector(GDV):
  - Graphlets: small, connected, induced, non-isomorphic subgraphs of a larger graph.
  - Graphlet Degree Vector for a particular node v is a vector that counts the number of each kind of graphlet that touches v.

- Orbit Degree Vector(ODV):
  - Orbits: the automorphism groups which nodes of every graphlet can be partitioned into.
  - Orbit Degree Vector count the number of nodes touching a particular graphlet at a node belonging to a particular orbit.

Przulj, N.: Biological network comparison using graphlet degree distribution. Bioinformatics 23(2), 177–183 (2007)
Przulj, N., Corneil, D.G., Jurisica, I.: Modeling interactome: scale-free or geomet ric? Bioinformatics. 20(18), 3508–3515 (2004)

Behavior View



Prˇzulj, N.: Biological network comparison using graphlet degree distribution. Bioinformatics 23(2), 177–183 (2007)
Przulj, N., Corneil, D.G., Jurisica, I.: Modeling interactome: scale-free or geomet ric? Bioinformatics. 20(18), 3508–3515 (2004)

1. Background

2. Motivation

3. Methodology

4. Experiments

Methodology

- Feature-based Embedding:

$$H_{(f)}^{(l)} = GNN(\mathbf{A}, H_{(f)}^{(l-1)}; W_{(f)}^{(l-1)})$$

- Behavior-based Embedding:
  - Similarity-based Graph Generation.
  - Behavior-based GNN.

Methodology

- Behavior-based Embedding:
  - Similarity-based Graph Generation.
  - Behavior-based GNN.

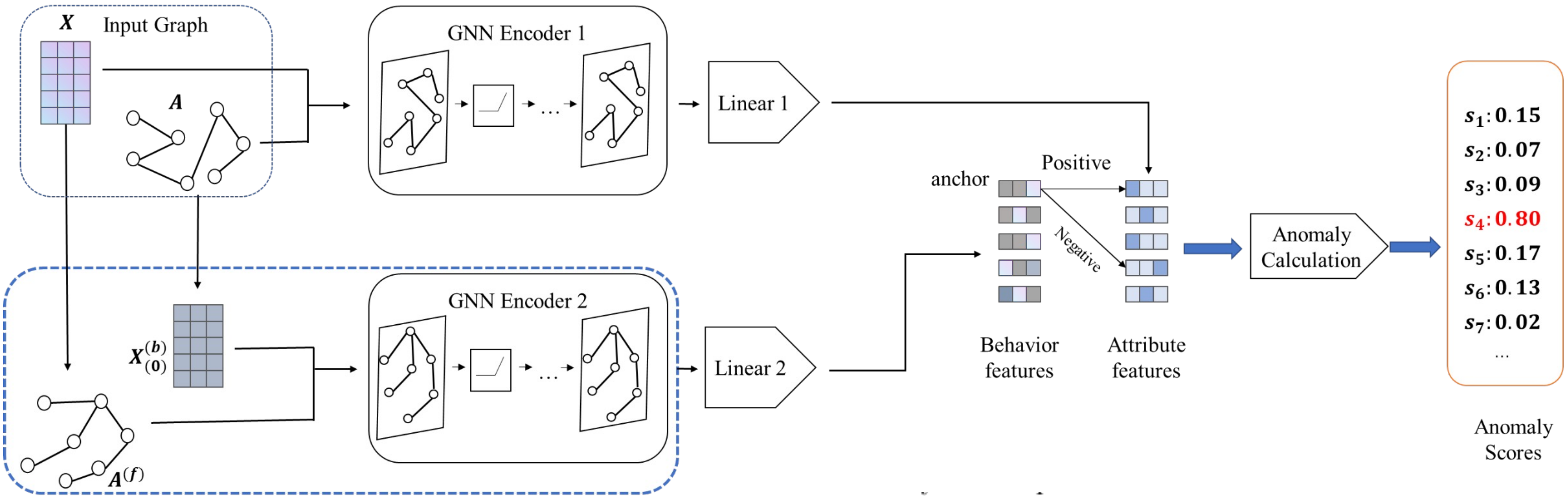
$$H_{(b)}^{(l)} = \sigma(\hat{\mathbf{D}}_{(f)}^{-1/2}\hat{\mathbf{A}}_{(f)}\hat{\mathbf{D}}_{(f)}^{-1/2}H_{(b)}^{l-1}W_{(b)}^{l-1})$$

Algorithm 1 Similarity-based Graph Generation(SGG)

Input: node set $\mathcal{V}$ with n nodes, node feature matrix $\mathbf{X}$, degree matrix $\mathbf{D}$
Output: The adjacency matrix of the generated graph $\mathbf{A_{(f)}}$

```
Initialize A' as an n × n matrix filled with zeros.
for v_i ∈ V do
    Compute the similarity between the feature vector of node v_i and all nodes: sim_i = [sim_{i,0}, ..., sim_{i,n}]
    Sort sim_i in descending order based on similarity values.
    Select the top k+1 nodes with the highest similarity values to form top_indices_i.
    for v_j ∈ top_indices_i[1 : 1 + D_ii] do
        Set A'_ij = 1
    end for
end for
for v_i ∈ V do
    for v_j ∈ V do
        Set A'_ij = max{A'_ij, A'_ji}
        Set A'_ji = max{A'_ij, A'_ji}
    end for
end for
A_(f) = A'
return A_(f)
```

Framework



$$\mathcal{L} = -\log \frac{exp(cos\_sim(h_i^{(f)}, h_i^{(b)})/\tau)}{exp(cos\_sim(h_i^{(f)}, h_i^{(b)})/\tau) + exp(cos\_sim(h_j^{(f)}, h_i^{(b)})/\tau)}$$

1. Background
2. Motivation
3. Methodology
4. Experiments

Experiments

Datasets

Table 1 Statistics of 3 real-world datasets, including the number of nodes and edges, the node feature dimension, the ratio of anomalous labels, and the concept of relations.

Dataset	**#Nodes**	**#Edges**	**#Feat.**	**Anomaly**	**Relation Concept**
Reddit	10,984	168,016	64	3.33%	Under Same Post
Tolokers	11,758	519,000	10	21.82%	Work Collaboration
Elliptic	203,769	234,355	166	9.76%	Payment Flow

Main Results

Table 3 AUC and AUPRC of HE-GAD and baselines. "-" indicates failed experiments due to memory constraint. The best result on each dataset is in bold while the second-best are underlined.

Datasets	Metrics	ARISE	GRADATE	NLGAD	PREM	HE-GAD
Reddit	AUC	0.5273	0.5261	0.5380	<u>0.5518</u>	**0.6328**
	AUPRC	0.0402	0.0393	<u>0.0415</u>	0.0413	**0.0514**
Tolokers	AUC	0.5514	0.5373	0.4825	<u>0.5654</u>	**0.6150**
	AUPRC	0.2505	0.2364	0.2025	<u>0.2590</u>	**0.2752**
Elliptic	AUC	-	-	0.4977	<u>0.4978</u>	**0.6518**
	AUPRC	-	-	<u>0.1009</u>	0.0905	**0.1061**

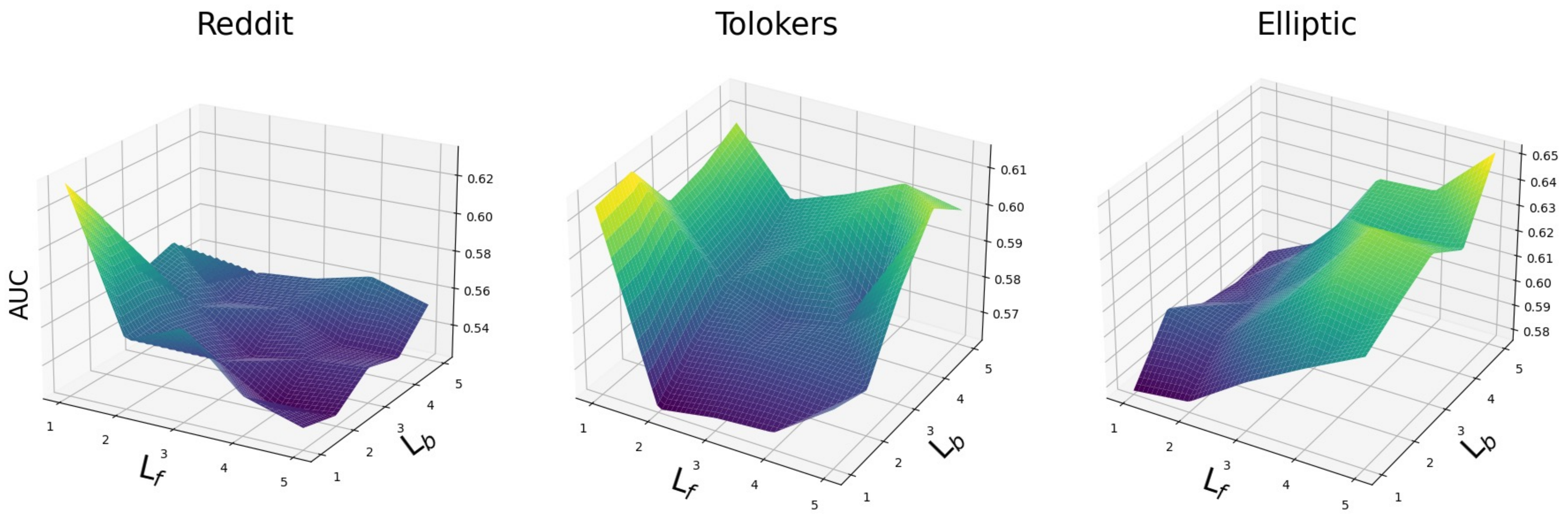
Ablation Study

- Whether it is reasonable to guide the aggregation of behavior features based on feature similarity?

Variant	Reddit	Tolokers	Elliptic
w/o similarity	0.4669	0.4982	0.5218
HE-GAD	**0.6328**	**0.6150**	**0.6518**

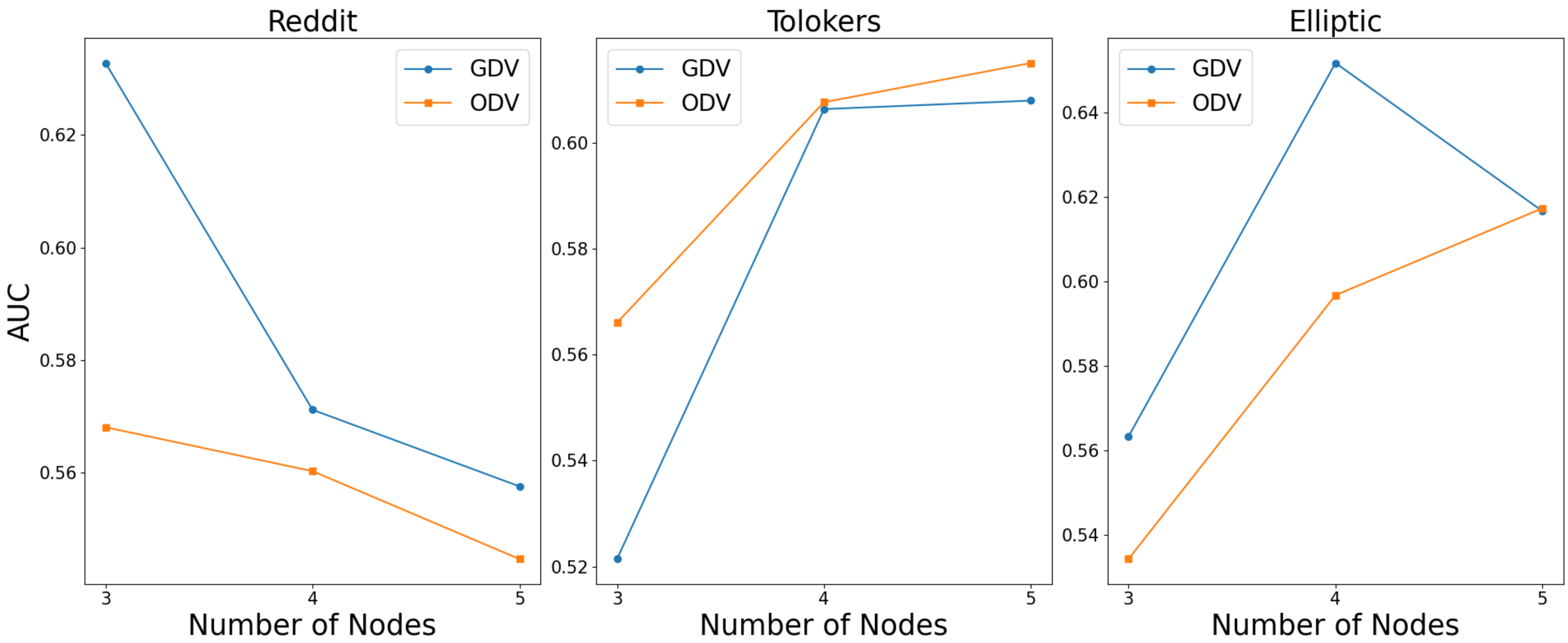
We choose to randomly select the same number of neighbors for each node for comparison.

Sensitivity Analysis



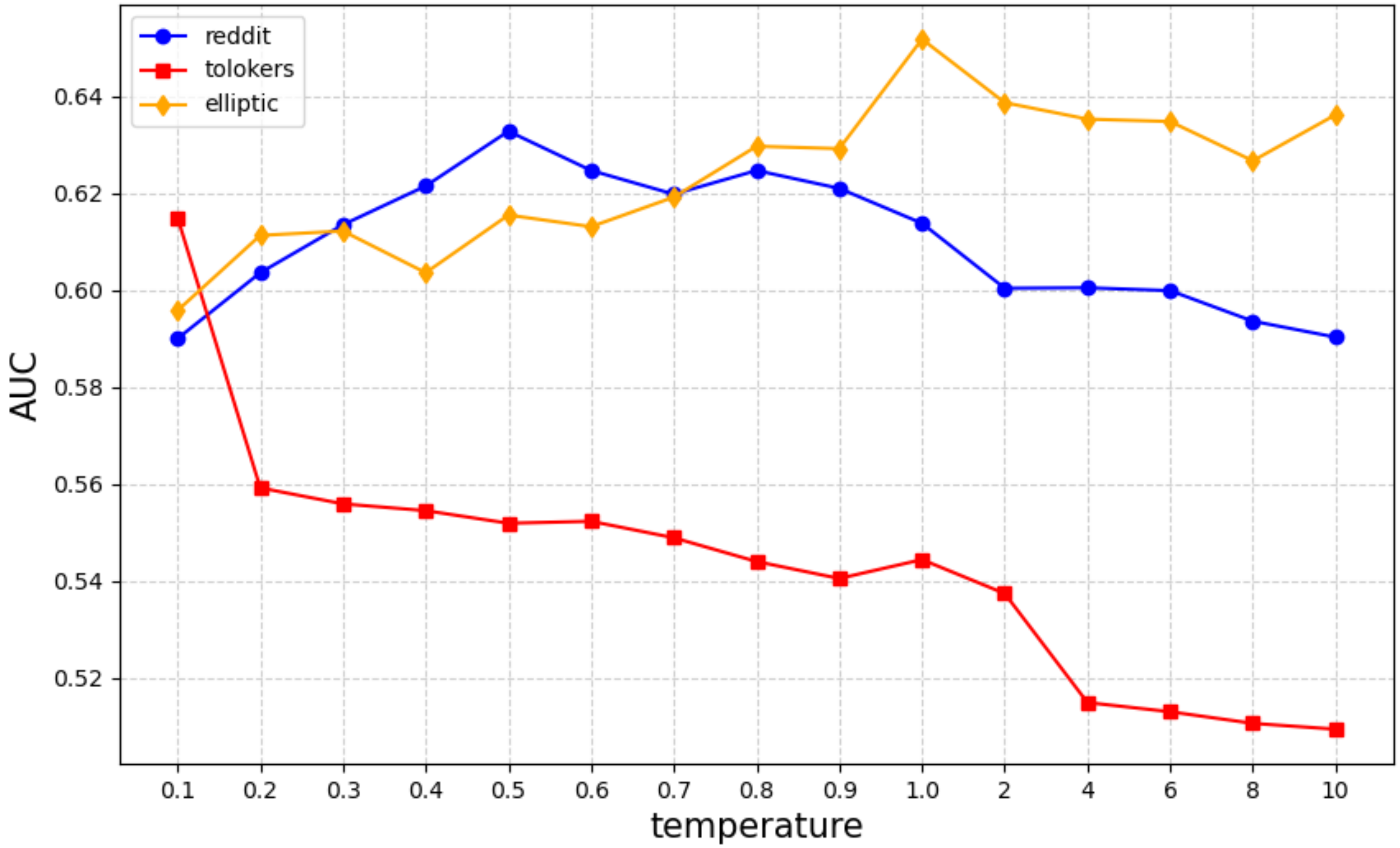
Number of layers of GNN encoders

Sensitivity Analysis



Type of Behavioral Features

Sensitivity Analysis



Temperature

Thanks for Listening!

Contact: lingzheng@mail.ustc.edu.cn