Appendix: Towards Effective Federated Graph Anomaly Detection via Self-boosted Knowledge Distillation

Jinyu Cai*
National University of Singapore
Singapore
jinyucai@nus.edu.sg

Yunhe Zhang*†
University of Macau
Macau, China
zhangyhannie@gmail.com

Zhoumin Lu
Northwest Polytechnical University
Xi'an, China
walker.zhoumin.lu@gmail.com

Wenzhong Guo
Fuzhou University
Fuzhou, China
guowenzhong@fzu.edu.cn

See-Kiong Ng
National University of Singapore
Singapore
seekiong@nus.edu.sg

1 Summary of Appendix

This appendix includes the following content:

(1) Detailed description of the graph benchmarks used in the experiment part.
(2) Detailed experimental settings.
(3) Theoretical and empirical complexity analysis.
(4) More Parameter analysis (number of GIN layers and latent dimensions).
(5) Justification of the backbone sharing strategy.
(6) Statistical significance of experimental results.

2 Detailed Description of Graph Benchmarks

In this section, we provide a detailed description of all the graph benchmarks used in the experiment, including the number of graphs, the average node numbers and edge numbers, and the classes. Table 1 summarizes the information of these graph benchmarks. Specifically, in the single-dataset experiment, we use three social network benchmarks, including IMDB-BINARY, COLLAB, and IMDB-MULTI. In the multi-dataset experiment, we construct four datasets by integrating different types of graph data, *e.g.*, molecules, biological, and social network data. The details are illustrated as follows:

- **MOLECULES:** This benchmark consists of multiple molecule datasets, including MUTAG, DHFR, PTC_MR, BZR, COX2, AIDS, and NCI1.
- **BIOCHEM:** This benchmark is a cross-domain dataset including datasets in MOLECULE, and additional biological datasets, including ENZYMES, PROTEINS, and DD.
- **SOCIALNET:** This benchmark contains multiple social network datasets, including IMDB-BINARY, COLLAB, and IMDB-MULTI.
- **MIX:** This benchmark contains all datasets from three domains, including molecular, biological, and social networks, in Table 1.

All graph benchmarks used in this paper are from TUDataset [2], a publicly available graph benchmark database[1].

3 Detailed Experimental Settings

In this section, we provide more details of the experimental settings in the paper, including the training details, trade-off parameter settings, and baseline settings.

- **Training Details:** We fix the batch size as 64 for all experiments and use Adam [1] as the optimizer with a fixed learning rate $\alpha = 0.001$. We first pre-train each local model, excluding the student network and knowledge distillation module for 10 epochs. Then, we jointly train the entire network with collaborative learning for 200 epochs.
- **Trade-off Parameter Settings:** The objective function of FGAD contains two trade-off parameters, *i.e.*, λ and γ, we vary their values within the range of $[1e^{-4}, 1e^{3}]$ and evaluate their impacts on performance in the Section 4.5.1 (in the main text). Regarding the number of clients C in a single-dataset, we vary it within the range of $[2, \ldots, 10]$ and evaluate its impact in Section 4.5.2 (in the main text), while for multi-dataset, the number of clients is set to the number of its sub-datasets. Besides that, for the number of GIN layers K, we also evaluate its impact under different values in Appendix 6.
- **Baseline Settings:** For the state-of-the-art baselines, including FedAvg, FedProx, GCFL, and FedStar, we integrate them with DeepSVDD [3] to build end-to-end GAD models. We also include the self-training strategy that removes collaborative learning, as one of the baselines. Note that we employ the same GIN backbone as FGAD to guarantee the fairness of the performance comparison. The objective of local models in each client is to minimize the distance from the projection of the training data in the latent space to the centroid, where the centroid is randomly initialized following the setting in DeepSVDD and fixed throughout the training phase. In the collaborative learning phase, we upload the learned decision boundaries in each client as part of the parameters and aggregate them in the server. Finally, we can calculate the anomaly scores by the distances between the graph representations and the centroid after training. The smaller the score, the more a graph tends to be considered normal.

4 Theoretical Complexity Analysis

Here, we provide a theoretical complexity analysis of the proposed FGAD method. Assume there are N graphs across all clients, and with maximal m nodes and $|E|_{\max}$ edges within a graph. In the

---

*Both authors contributed equally to this research.
†Corresponding author.
[1] https://chrsmrrs.github.io/datasets/docs/datasets/

Table 1: Detailed information of the datasets used in the experiment.

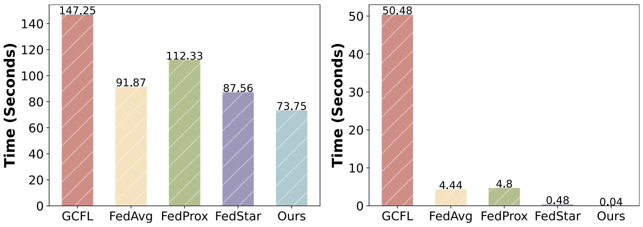
Dataset Name	#Graphs	#Average Nodes	#Average Edges	#Graph Classes	Data Type
IMDB-BINARY	1,000	19.77	96.53	2	Social Network
COLLAB	5,000	74.49	2,457.78	3	Social Network
IMDB-MULTI	1,500	13.00	65.94	3	Social Network
MUTAG	188	17.93	19.79	2	Molecule
DHFR	756	42.43	44.54	2	Molecule
PTC_MR	344	14.29	14.69	2	Molecule
BZR	405	35.75	38.36	2	Molecule
COX2	467	41.22	43.45	2	Molecule
AIDS	2,000	15.69	16.20	2	Molecule
NCI1	4,110	29.87	32.30	2	Molecule
ENZYMES	600	32.63	62.14	6	Biology
PROTEINS	1,113	39.06	72.82	2	Biology
DD	1,178	284.32	715.66	2	Biology

local model of each client, the maximal dimension among input and latent space of GIN is denoted by $\tilde{d}$, and the number of GIN layers is represented by L. In Addition, the maximal latent dimensions of the teacher and student heads are denoted by d_{t} and d_{s}, respectively, and the number of latent layers in the teacher and student heads is denoted by K_{t} and K_{s}. We can analyze the time and space complexity of FGAD within a single client, as well as the communication complexity in collaborative learning, as follows:

- **Time Complexity**: Since the teacher and student models share the same GIN backbone, the time complexity of the GIN backbone is $O(NL(m\tilde{d}^2+|E|_{\max}\tilde{d}))$. Similarly, the time complexity of the anomaly generator in the teacher model mainly comes from the GIN. For the teacher and student heads, the time complexities are $O(K_{\mathrm{t}}\tilde{d}d_{\mathrm{t}})$ and $O(K_{\mathrm{s}}\tilde{d}d_{\mathrm{s}})$, respectively. Consequently, the overall time complexity of FGAD framework is approximately $O(2NL(m\tilde{d}^2 + |E|_{\max}\tilde{d}) + (K_{\mathrm{t}}d_{\mathrm{t}} + K_{\mathrm{s}}d_{\mathrm{s}})\tilde{d})$, where includes the anomaly generator weight-shared GIN backbone, and the teacher and student heads.
- **Space Complexity**: For the space complexity of the GIN backbone, the space complexity mainly comes from the storage of weight and bias matrices in each layer, which can be denoted by $O(L\tilde{d}(1+\tilde{d})$. For the teacher and student heads, their space complexities can be derived similarly, *i.e.*, $O(K_{\mathrm{t}}\tilde{d}(1+d_{\mathrm{t}})+K_{\mathrm{s}}\tilde{d}(1+d_{\mathrm{s}}))$. Consequently, the overall space complexity of FGAD framework is approximately $O(L\tilde{d}(1+\tilde{d}) + K_{\mathrm{t}}\tilde{d}(1+d_{\mathrm{t}}) + K_{\mathrm{s}}\tilde{d}(1+d_{\mathrm{s}}))$.
- **Communication Complexity**: Since the teacher model in FGAD is used for the personalization of local clients, only the student head engages in collaboration. Consequently, the communication complexity is approximately $O(K_{\mathrm{s}}\tilde{d}d_{\mathrm{s}})$ and $O(K_{\mathrm{s}}\tilde{d}(1+d_{\mathrm{s}}))$.

5 Empirical Complexity Analysis

To more comprehensively analyze the complexity of FGAD, we further provide empirical complexity analysis. Specifically, we compare the running time (in local training) and communication time (in collaboration learning) of FGAD with other baselines. Note that the experiment is conducted under uniform device settings to ensure fairness. The experimental results are presented in Figure 1.



(a) Running Time (b) Communication Cost

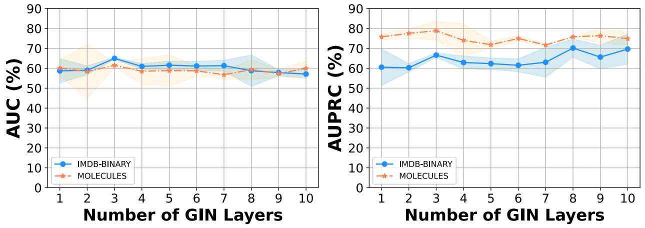
Figure 1: Running time and communication cost comparison in 200 epochs.

It can be observed from Figure 1(a) that the time complexity of FGAD is competitive with several baselines, *e.g.*, that of FedStar, and significantly better than that of GCFL. Combined with the performance comparison in Tables 1-2 (in the main text), the overall experimental results demonstrate that FGAD not only significantly improves anomaly detection performance but also possesses promising time efficiency compared to other baselines.

Additionally, communication cost (time) is also an important evaluation metric in federated learning. Therefore, we further conduct the comparative experiment to demonstrate the efficiency of FGAD. As shown in Figure 1(b), FGAD has the lowest communication time compared with other baselines, which aligns with the comparison of exchanging amount of network parameters in Table 1 (in the main text). It should be noted that this is the analog communication time without considering the network bandwidth. When it comes to real-world collaboration, the network bandwidth will significantly impact the efficiency of model parameter transmission. Consequently, in cases of models with large parameter sizes, communication time will become a pivotal factor that influences the time complexity of collaborative learning.

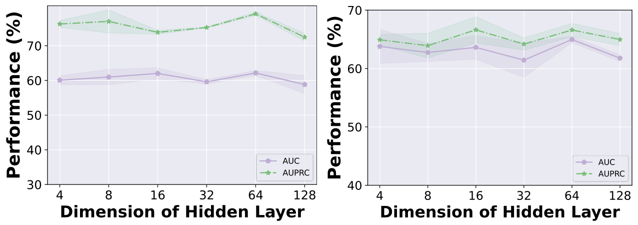
6 Impact of GIN Layers

We delve into the impact of the number of GIN layers K on the anomaly detection performance of FGAD. The parameter K plays a pivotal role in determining the extent to which the model explores neighborhood information and the overall complexity of FGAD. We



(a) (b)

Figure 2: Average performance with standard deviation under different numbers of GIN layers on IMDB-BINARY and MOLECULES datasets. Note that the number of GIN layers is set to $[1, \ldots, 10]$.



(a) MOLECULES (b) IMDB-BINARY

Figure 3: Parameter sensitivity of different dimensions in the hidden layer.

systematically analyze its impact by varying the K within the range of $[1, \ldots, 10]$ and conduct a series of experiments. Figure 2 reports the experimental results on the IMDB-BINARY and MOLECULES datasets, from which we have the following observations. First, a certain depth of GIN is beneficial to fully leverage the structural information of graph data for learning powerful GAD models, which could be verified from the performance improvement in both datasets. Second, when the number of GIN layers continues to increase, the performance improvements become increasingly marginal or even exhibit slight degradation. This trend indicates that a moderate number of GIN layers (*e.g.*, 3) is sufficient to effectively leverage the neighborhood information within graphs. Third, we can observe from the overall experimental results that the performance stays relatively stable under the variation of K, which demonstrates the robustness of FGAD.

7 Impact of Latent Dimensions

Here, we further conduct additional parameter analysis to evaluate the impact of the latent dimension in the GIN layer. Specifically, we set the latent dimension from $[4, \ldots, 128]$, and the experimental results on MOLECULES and IMDB-BINARY are shown in Figure 3. The results suggest that FGAD exhibits relatively stable performance across a wide range of latent layer dimensions, demonstrating its robustness. Nevertheless, we can observe that excessively high dimensions (*e.g.*, 128) might adversely affect performance, which is probably due to the redundant information it brings.

8 Justification of the Backbone Sharing

To justify the rationale for sharing the backbone network between the teacher and student models, we conduct additional experiments by comparing the performance of FGAD with and without sharing the GIN backbone. As shown in Table 2, we can observe there is only a marginal difference in performance between these two strategies. This observation suggests that sharing the GIN backbone would not decrease the effectiveness of knowledge distillation in FGAD. More importantly, the significant benefit of sharing the GIN backbone is the substantial reduction in model complexity. This streamlined architecture leads to a more efficient model in terms of computational and memory resource usage.

Table 2: Performance (mean(%) ± std(%)) of FGAD under shared/separated GIN backbone.

Backbone	IMDB-BINARY		IMDB-MULTI	
	AUC	AUPRC	AUC	AUPRC
Shared GIN	64.97±0.52	66.60±1.12	60.51±1.18	66.82±0.14
w/o Shared GIN	63.13±1.19	66.43±2.23	58.13±0.84	66.67±0.00

9 Statistical Significance of Results

To verify the statistical significance of the experimental results comparing FGAD with other baseline methods, we conduct a Student’s t-test between the proposed FGAD with several state-of-the-art baselines. Note that a difference is considered statistically significant if the p-value from the t-test is less than 0.05. Table 3 presents the p-values (with 10 runs) for FGAD compared to various baselines across two datasets (IMDB-Binary and MIX), demonstrating that our results are statistically significant.

Table 3: p-value (t-test) of FGAD v.s. several baselines. Note that IMDB-B denotes the IMDB-Binary dataset.

Method	v.s. Self-train	v.s. FedAvg	v.s. GCFL	v.s. FedStar
IMDB-B(AUC)	1.0×10^{-8}	2.9×10^{-14}	1.3×10^{-20}	1.7×10^{-12}
IMDB-B(AUPRC)	2.2×10^{-9}	5.4×10^{-13}	3.6×10^{-12}	1.7×10^{-10}
MIX(AUC)	5.5×10^{-6}	2.7×10^{-14}	6.0×10^{-14}	3.2×10^{-6}
MIX(AUPRC)	4.9×10^{-12}	5.3×10^{-20}	1.9×10^{-15}	1.0×10^{-9}

References

[1] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[2] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. 2020. TUDataset: A collection of benchmark datasets for learning with graphs. In *Proceedings of the ICML Workshop on Graph Representation Learning and Beyond*. arXiv:2007.08663 www.graphlearning.io

[3] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4393–4402.