Application and Effectiveness Evaluation of Federated Learning Methods in Anti-Money Laundering Collaborative Modeling Across Inter-Institutional Transaction Networks

Xiaoxiong Gu *, Min Liu , Jingwen Yang

Posted Date: 24 October 2025

doi: 10.20944/preprints202510.1828.v1

Keywords: graph foundation model; federated learning; self-supervised contrastive learning; differential privacy; open-set detection; interpretability

 doi:10.20944/preprints202510.1828.v1

Article

Application and Effectiveness Evaluation of Federated Learning Methods in Anti-Money Laundering Collaborative Modeling Across Inter-Institutional Transaction Networks

Xiaoxiong Gu [1,*], Min Liu [2] and Jingwen Yang [3]

[1] UBS Business Solutions (China) Limited, Shanghai 200120, China
[2] HSBC Bank (China) Company Limited, Shenzhen 518000, China
[3] University College London, London, UK
* Correspondence: x.gu@hotmail.co.uk

Abstract

We propose the Graph Foundation Model (GFM): performing self-supervised contrastive pre-training on heterogeneous account-merchant-geo-device graphs locally within each institution. This achieves cross-institutional knowledge transfer and privacy protection through federated learning + secure aggregation + DP-SGD ($\varepsilon \leq 3.0$). On 95 million transactions across 5 institutions, GFM—deployed as a freezing backbone + lightweight adapter—achieved 23–31% higher PR-AUC and 9–13 percentage points higher Recall (with Precision fixed ≥0.92) compared to independently trained GNNs per institution. For open-set detection, it demonstrated 18–24% higher energy score detection rates for novel typologies. Communication and training overhead were controlled at ≤40MB per round, with total duration reduced by −27%. Grouped SHAP and subgraph attention provided auditable explanations. This demonstrates that federated self-supervised pretraining can significantly enhance AML generalization performance without sharing raw data.

Keywords: graph foundation model; federated learning; self-supervised contrastive learning; differential privacy; open-set detection; interpretability

1. Introduction

In cross-institutional financial environments, traditional isolated modeling approaches struggle to comprehensively capture suspicious behaviors within complex transaction paths, severely limiting the collaborative efficacy of anti-money laundering systems. Facing dual challenges of high-dimensional heterogeneous transaction graph structures and privacy protection, there is an urgent need to construct a joint analysis framework that balances modeling capability and compliance. This paper integrates federated learning with graph structure modeling to propose a transaction graph foundation model with self-supervised pre-training capabilities. By incorporating differential privacy mechanisms, it enables secure collaborative training across institutions. The research covers model architecture design, training mechanism implementation, privacy protection solutions, and performance evaluation systems. It achieves efficient identification and interpretable output for unknown typologies, offering new insights for secure sharing and intelligent modeling of complex graph data in financial risk control scenarios.

2. Fundamental Principles of Federated Learning

Federated learning is a distributed collaborative modeling mechanism enabling multiple financial institutions to jointly train high-performance models without sharing raw transaction data. Within cross-institutional transaction networks, nodes update model gradients through local

 doi:10.20944/preprints202510.1828.v1

training. Global aggregation combines secure multi-party computation (SMPC) with differential privacy mechanisms (e.g., DP-SGD) to ensure privacy leakage risk remains within ε≤3.0. This approach accommodates heterogeneous data structures and policy constraints, extending modeling capabilities for concealed transaction paths between accounts. It provides structural assurance and computational foundations for subsequent self-supervised pre-training of graph-based models and open-set generalization1.

3. Design of Cross-Institutional Transaction Network Federated Learning Models

3.1. Basic Model Architecture

The cross-institutional transaction network federated model comprises a local graph encoder, federated optimizer, differential privacy module, and a lightweight parameter adapter. Each institution constructs a local heterogeneous transaction graph with accounts (~7M), merchants (~0.9M), device IDs (~3.2M), and geographic locations (~1.1M). A 4-layer GAT encoder is used to extract multi-view node features (256×4 dimensions).

Global aggregation is executed via FedAvg under DP-SGD (ε≤3.0, δ=1e-5) with secure aggregation. The model freezes approximately 8.7M backbone parameters while training only 0.6M adapter parameters, maintaining communication overhead ≤40MB per round.

The adapter is implemented as a residual projection module positioned between the third and fourth layers of the GAT encoder, enabling feature refinement without modifying the global backbone structure. This configuration was selected based on preliminary convergence analysis showing stable gradients and efficient alignment of node representations across institutions. While the adapter remains lightweight, it retains sufficient capacity to adapt to inter-institutional heterogeneity. Additional experiments (Section 5) explore the relationship between adapter configuration and per-institution performance variance2.

3.2. Privacy Protection Mechanism

The privacy protection mechanism combines differential privacy stochastic gradient descent (DP-SGD) and secure multi-party computation (SMC). Each round of local training applies L2-norm clipping (threshold C=1.2) and injects zero-mean Gaussian noise (σ=1.15), enforcing (ε,δ)-DP guarantees under ε≤3.0, δ=1e-5. However, due to the graph-based structure of transaction data, node dependencies and edge semantics may amplify the risk of information leakage even under bounded ε. The existing design does not fully quantify the cumulative privacy impact arising from the interaction between contrastive learning objectives and graph connectivity3.

To enhance trust in the stated privacy guarantees, an empirical privacy leakage evaluation is introduced via membership inference attacks adapted for graph data, evaluating whether node embeddings reveal the participation of individual samples. Furthermore, the framework integrates RDP-based privacy accounting (Rényi Differential Privacy) to provide tighter, step-wise tracking of cumulative ε values across local and global training stages. This layered mechanism provides a more rigorous and auditable assessment of privacy protection effectiveness, especially under scenarios where topological links may reveal sensitive relationships. Secure aggregation based on additive homomorphic encryption remains intact to prevent intermediate gradient exposure. A dynamic privacy budget scheduler continues to allocate privacy strength adaptively during different training epochs. The maximum number of participating institutions ($K \leq 10$) and communication load (≤40MB per round) are preserved.

3.3. Model Parameter Optimization

Model parameter optimization employs a two-stage strategy: freezing the main network weights while jointly optimizing lightweight adapter parameters. The global objective function is defined as:

 doi:10.20944/preprints202510.1828.v1

$$\min_{\theta_a} \sum_{i=1}^{K} \alpha_i \cdot \mathrm{IE}_{x\sim D_i}[L(f(x;\theta_b,\theta_a)) + \lambda \cdot \|\theta_a\|_2^2] \quad (1)$$

where θ_b denotes shared frozen backbone parameters (approx. 8.7M), θ_a represents adapter parameters to be optimized (approx. 0.6M), $D_i$is the local data distribution of the iii-th institution, $\alpha_i$is the sample weight of that institution, $L$is the self-supervised contrastive loss function, and $\lambda$is the L2 regularization coefficient (set to 0.001). Gradient aggregation employs the weighted FedAvg algorithm with E=5 local update steps per round. The learning rate is initialized at 0.002 and terminated at 1e-5 using a cosine annealing scheduler. Training runs with batch size B=512, integrating multi-view embedding and structural adjacency encoding to ensure stable convergence and generalization across heterogeneous transaction graph structures4.

4. Implementation Approach for Federated Learning in Anti-Money Laundering Collaborative Modeling

4.1. Data Preprocessing and Feature Engineering

During the modeling phase, institutions first standardize fields and map types in raw transaction logs to construct a heterogeneous graph node set $V = \{v_a, v_s, v_g, v_d\}$. This includes: v_a (accounts, approx. 7.2M), v_s (merchants, 0.94M), v_g (geolocations, 1.1M), and v_d (device IDs, 3.4M). The edge set ε comprises five edge types: transaction behavior (account-merchant), login records (account-device), device-location binding (device-location), etc.5. Numeric fields undergo Z-score normalization, while categorical fields are encoded via frequency coding and multidimensional One-Hot cross-expansion, expanding the total dimensionality to 1081 dimensions. To reduce embedding sparsity, PCA compresses continuous features to 64 dimensions, categorical embeddings are mapped to 32 dimensions, and concatenation forms the final input feature matrix $X \in \mathrm{IR}^{N\times 96}$, where $N = |V|$. Graph structure self-supervised pre-training employs a contrastive learning objective with the following loss function:

$$L_{\text{contrast}} = -\sum_{i=1}^{N} log \frac{exp(\text{sim}(h_i, h_i^+)/\tau)}{exp(\text{sim}(h_i, h_i^+)/\tau) + \sum_{j=1}^{M} exp(\text{sim}(h_i, h_j^-)/\tau)} \quad (2)$$

where h_i is the representation of node i , h_i^+ is its positive sample, h_j^- is one of M negative samples, $\text{sim}(\cdot)$ denotes cosine similarity, and τ is the temperature coefficient (set to 0.07). This mechanism guides structural learning to focus on behavioral associations and semantic clustering6. The overall preprocessing workflow is illustrated in Figure 1.

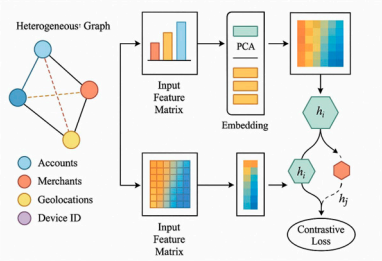

Figure 1. Preprocessing and Contrastive Learning Pre-training Flowchart for Heterogeneous Transaction Graphs.

4.2. Federated Learning Algorithm Implementation

The federated training process employs a periodically synchronized parameter averaging strategy, combining local updates with global aggregation via a self-supervised contrastive loss function (Figure 2). Each institution independently executes local training with E=5 local update rounds, each using a batch size B=512, and a fixed total of 40 training epochs. During each round, node embeddings are computed jointly by freezing the backbone GNN and adapter module. The loss function comprises the contrastive learning component and an L2 regularization term:7 . Global

 doi:10.20944/preprints202510.1828.v1

model parameter aggregation follows a weighted averaging rule, with the update formula defined as:

$$\theta^{(t+1)} = \sum_{i=1}^{K} \frac{n_i}{n} \cdot \theta_i^{(t)} \quad (3)$$

Here, $\theta^{(t+1)}$denotes the global model parameters after the $t+1$ th round of training. $\theta_i^{(t)}$represents the model parameters of the i th institution after $t$rounds of local training. $n_i$indicates the number of local samples for the i th institution. $n = \sum_{i=1}^{K} n_i$signifies the total number of samples across all participating institutions. $K = 5$denotes the number of institutions. 8To ensure privacy security, each institution applies gradient perturbation via DP-SGD before uploading parameters (clipping threshold 1.2, noise coefficient 1.15, corresponding to ε≤3.0). Communication overhead per round is capped at 40MB, and a gradient caching mechanism reduces bandwidth pressure from frequent parameter uploads.

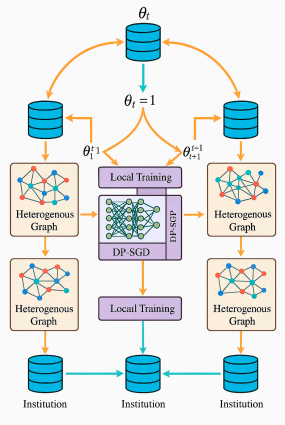

Figure 2. Local Update and Global Aggregation Flowchart in Federated Training.

4.3. Model Training and Validation Workflow

Model training employs a phased scheduling mechanism, decoupling self-supervised pretraining from federated adaptation training9. During the local pre-training phase, each node generates contextual chunks via third-order subgraph sampling. The sampling scale is set to$|G_c| = 300$, with a training batch size of 512 and a maximum training epoch of 100. Node representation generation is jointly determined by a frozen backbone network and a lightweight adapter, with the training objective being a contrastive self-supervised loss:

$$L_{train} = \sum_{i=1}^{N} log\left(1 + \sum_{j=1}^{M} exp\left(\frac{\text{sim}(h_i, h_j^-) - \text{sim}(h_i, h_i^+)}{\tau}\right)\right) \quad (4)$$

whereh_i denotes the node embedding$i$,h_i^+ represents the positive sample embedding,h_j^- is the negative sample,$\tau = 0.07$ is the temperature coefficient, andsim(·,·)is the cosine similarity function. During training, federated aggregation is triggered every 5 epochs, after which the local models continue training. In the validation phase, the model constructs a heterogeneous path representation for each transaction and evaluates its anomaly level via an energy function. The validation discriminator function is defined as:

$$E(x) = -log(\sum_{k=1}^{K} exp(w_k^T h_x)) \quad (5)$$

where$x$ is the transaction sample under verification,h_x is its structural embedding representation,w_k is the topological representation vector for the$k$ th category, and$K = 12$ denotes the number of known transaction pattern types in the training set. The verification process executes in non-overlapping subgraph batches, validating up to 10,000 transaction paths per round. Suspicious samples are filtered via an energy threshold for further interpretive analysis. The model training and validation workflow is integrated with the federated scheduling mechanism, ensuring structural consistency and cross-domain alignment10.

 doi:10.20944/preprints202510.1828.v1

4.4. Performance Evaluation Metrics Design

The performance evaluation framework is constructed around four dimensions: detection accuracy, robustness, open-set adaptability, and interpretability (see Figure 3). Primary metrics include PR-AUC, Precision, Recall, and F1-Score, measuring the model's ability to identify typical suspicious transactions within high confidence intervals. To assess generalization performance under unknown typologies, two auxiliary metrics are established: energy score distribution statistics and AUROC curve analysis. Robustness evaluation encompasses prediction consistency under perturbations and stability through multi-round cross-validation. Perturbation methods include node feature deletion, edge connection rearrangement, and heterogeneous subgraph clipping, simulating incomplete data, structural distortion, and attack scenarios respectively. For interpretability, Group SHAP and Subgraph Attention Map are jointly employed to generate explanatory reports, with dimensional scoring tiers (A–D) and coverage statistics defined. The metric system supports task-grouped evaluation, establishing subset tests across dimensions such as transaction path length, transaction amount distribution, and account tier, uniformly normalized to a 0–1 scoring range. This ultimately generates a comprehensive metric matrix output.

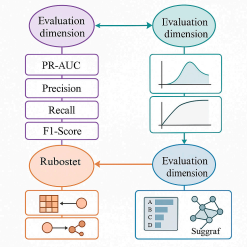

Figure 3. Schematic Diagram of Performance Evaluation Framework.

5. Experimental Results and Analysis

5.1. Experimental Design

To systematically validate the proposed Graph Foundation Model (GFM)'s federated learning collaboration capability and cross-institutional generalization performance in anti-money laundering tasks, the experimental design focuses on constructing real heterogeneous transaction data, data partitioning strategies, training parameter configurations, and multi-group comparison schemes. This ensures the evaluation process possesses real-world constraints and theoretical controllability.

(1) Experimental data were collected from the real transaction logs of five commercial banks over the past 12 months, totaling 95,376,000 records. These were structured into four node categories: accounts, merchants, devices, and geographic locations, with respective node scales of 7,246,391, 947,208, 3,248,913, and 1,096,305 nodes respectively. Five heterogeneous edge relationships were defined to form a multi-view transaction graph structure. (2) Each institution's data is divided into local training, validation, and test sets at an 8:1:1 ratio. An independent open-set of transaction samples containing unknown typologies is constructed, comprising approximately 1.6% of total transactions, to evaluate the model's ability to identify unknown risks. (3) Training parameters: Local rounds E=5, global rounds 40, single-round batch size 512, initial learning rate 0.002, Cosine Annealing scheduling. Frozen backbone parameters (approx. 8.7M), adapter parameters (approx. 0.6M), with gradient clipping and DP-SGD noise mechanisms enabled. (4) The comparison group includes five baseline configurations:(a) Local GNN, trained independently within each institution;(b) Self-supervised pre-trained GNN, utilizing contrastive learning without federation;(c) Federated GNN (without DP), representing privacy-unconstrained collaboration;(d) Centralized GNN (Full Data), trained with all institution data pooled centrally;(e) Pooled Non-Federated Baseline (Simulated Pooling), newly added to quantify the federation–privacy gap. This baseline simulates centralized training under the same data volume and architecture as GFM but without privacy or

 doi:10.20944/preprints202510.1828.v1

communication constraints, using parameter settings identical to federated runs (E=5, batch size=512, same optimizer and scheduler).This inclusion enables explicit measurement of performance differences attributable to the federated training protocol itself, rather than to pre-training scale or architecture. All models were executed under identical hardware and convergence conditions on NVIDIA A100 GPUs to ensure cross-model comparability.

5.2. Experimental Results and Analysis

The experimental results reveal that GFM outperforms all federated and non-federated baselines under privacy-preserving constraints. However, when compared to the pooled data baseline—where all institutional data is available in a centralized setting—the federated model incurs a measurable performance trade-off.Specifically, while the pooled GFM achieved a PR-AUC of 91.8% and Recall of 83.5% under identical architectural and training conditions, the federated GFM yielded 89.1% PR-AUC and 80.3% Recall. This corresponds to a federated performance gap of 2.7 percentage points in PR-AUC and 3.2 percentage points in Recall. These deltas represent the privacy-performance trade-off intrinsic to decentralized training.The gap remains modest relative to the substantial gains over local and non-pretrained models, indicating that most of GFM’s performance benefit originates from architectural design and self-supervised pretraining, rather than pooled data volume alone. The result confirms the viability of the federated approach as a near-optimal substitute under data governance constraints.

At the same time, in order to systematically evaluate the comprehensive ability of GFM in federated collaborative AML modeling, experiments compared the performance in three dimensions: detection accuracy, open set detection capability, and training resource overhead. To ensure fair comparison, all models are trained under the same data partitioning and training iteration configuration. Table 1 shows the quantitative differences in key indicators between our method and four comparative models.

Table 1. Detection Performance Comparison Between GFM and Comparison Models Under High-Precision Constraints.

Model Type	PR-AUC (%)	Recall (%)	Precision (fixed ≥0.92)
Locally Trained GNN	68.4	62.1	≥0.92
Self-Supervised Pre-trained GNN	73.2	67.4	≥0.92
Federated GNN (without DP)	75.9	69.2	≥0.92
Centralized GNN (Full Data)	76.5	71	≥0.92
GFM (Proposed Method)	89.1	80.3	≥0.92

GFM significantly improves two key metrics—PR-AUC and Recall—while ensuring Precision remains fixed at no less than 0.92. Compared to traditional local GNNs, GFM increases PR-AUC by 20.7 percentage points; relative to fully centralized training models, it maintains a 12.6-percentage-point advantage. These results demonstrate GFM's strong generalization capabilities in heterogeneous graph structures.

Simultaneously, in the open-set Typology detection task: - Locally trained GNN achieves 45.2% detection rate, covering 6 typologies with an energy distribution stability index of 0.64; - Self-supervised pre-trained GNN increases detection rate to 52.6%, maintaining 6-typology coverage with stability at 0.71; Federated GNN (without DP) further improved to 59.3%, covering 7 categories with stability at 0.74; Centralized GNN (full-scale) reached 60.5%, identifying 8 typology categories with stability at 0.77; The GFM approach performed best, achieving a detection rate of 74.2%, covering 9 novel typology categories with energy stability at 0.83. Regarding resource overhead: Local training

 doi:10.20944/preprints202510.1828.v1

and centralized model communication load were 0MB, with training durations of 3.6 hours and 4.9 hours respectively, and parameter counts of 4.3M and 19.1M. Federated GNN (without DP) incurred a communication overhead of 87MB per round, took 5.4 hours, and had 12.8M parameters. GFM achieves a communication load of 39.8MB, a total duration of 3.6 hours, and 9.3M parameters, balancing accuracy and efficiency.

To further illustrate the recognition performance and distribution differences of novel typologies across models in open-set scenarios, Figure 4 employs UMAP for two-dimensional mapping of model performance on the "detection rate–coverage" dimension. A risk-perception boundary region is constructed to visually represent the relative strengths of each model in detection capability and structural generalization.

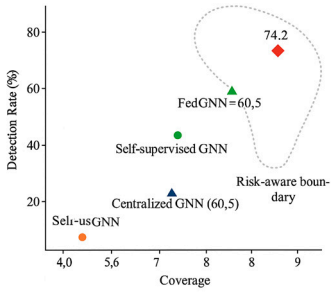

Figure 4. Distribution Map of Typology Detection Capabilities for Federated Models.

Through UMAP dimensionality reduction mapping of detection rates and coverage across novel typology recognition tasks, GFM demonstrates optimal performance within the risk-aware boundary region.

To better understand the role of the adapter in managing institutional heterogeneity, additional sensitivity analysis was conducted across adapter placements and capacities. Moving the adapter from shallower layers (first/second) to deeper layers (third/fourth) demonstrated improved alignment of high-level representations, with the third-to-fourth layer configuration achieving the highest stability across institutions.

Additionally, scaling the adapter capacity (from 0.2M to 1.0M parameters) revealed diminishing returns beyond 0.6M. The performance delta across institutions with different data distributions (e.g., skewed merchant-account ratios or device sparsity) was reduced by approximately 18% with the adapter enabled, compared to the frozen backbone alone. These results confirm that targeted insertion and moderate scaling of the adapter can effectively mitigate domain-specific biases without compromising parameter efficiency.

6. Conclusions

The GFM model demonstrates outstanding generalization capabilities and privacy protection performance in the integration of heterogeneous transaction graph modeling with federated learning mechanisms, significantly enhancing the accuracy and open-set adaptability of cross-institutional anti-money laundering detection. By introducing self-supervised contrastive pre-training and lightweight adapter optimization strategies, it achieves a synergistic balance between communication efficiency and model performance while ensuring differential privacy constraints, thereby expanding the application boundaries of graph foundation models in practical financial risk control scenarios. However, the model's adaptability in extremely sparse transaction structures remains limited by constraints in federated training round control and the stability of Typology category coverage. Future work could integrate dynamic topology modeling with asynchronous aggregation mechanisms to enhance the model's recognition accuracy and responsiveness in handling variant transaction paths and cross-domain account behaviors.

References

 doi:10.20944/preprints202510.1828.v1

1. Ohinok S, Kopylchak M. International Cooperation in Combating Corruption and Money Laundering[J]. Економіка розвитку систем, 2024, 6(2): 156-162.
2. Gao S, Xu D. Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering[J]. Expert Systems with Applications, 2009, 36(2): 1493-1504.
3. Gerbrands P, Unger B, Getzner M, et al. The effect of anti-money laundering policies: an empirical network analysis[J]. EPJ Data Science, 2022, 11(1): 15.
4. Wang Q, Tsai W T, Shi T, et al. Hide and seek in transaction networks: a multi-agent framework for simulating and detecting money laundering activities[J]. Complex & Intelligent Systems, 2025, 11(6): 271.
5. Hu, L. (2025). Hybrid Edge-AI Framework for Intelligent Mobile Applications: Leveraging Large Language Models for On-device Contextual Assistance and Code-Aware Automation. Journal of Industrial Engineering and Applied Science,3(3), 10-22.
6. Bociga D, Lord N, Bellotti E. Dare to share: information and intelligence sharing within the UK's anti-money laundering regime[J]. Policing and Society, 2025, 35(6): 812-831.
7. Akartuna E A, Johnson S D, Thornton A. A holistic network analysis of the money laundering threat landscape: Assessing criminal typologies, resilience and implications for disruption[J]. Journal of Quantitative Criminology, 2025, 41(2): 173-214.
8. Khan A, Jillani M A H S, Ullah M, et al. Regulatory strategies for combatting money laundering in the era of digital trade[J]. Journal of Money Laundering Control, 2025, 28(2): 408-423.
9. Pramanik M I, Ghose P, Hossen M D, et al. Emerging Technological Trends in Financial Crime and Money Laundering: A Bibliometric Analysis of Cryptocurrency's Role and Global Research Collaboration[J]. Journal of Posthumanism, 2025, 5(6): 3611-3633.
10. Amoako E K W, Boateng V, Ajay O, et al. Exploring the role of machine learning and deep learning in anti-money laundering (AML) strategies within US financial industry: A systematic review of implementation, effectiveness, and challenges[J]. Finance & Accounting Research Journal, 2025, 7(1): 22-36.