FTEC5920 Financial Technology Project II
Federated Graph Link Prediction Implementation

Student Name: DU Jiaxing
Student ID: 1155204280

Supervisor: Professor XU Hong, CUHK
Cobi XU, A.I. Phoenix Technology Co

April 2024

Table of Contents

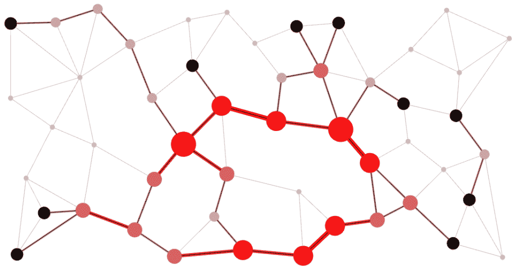
Table of Contents

Graph Neural Networks (GNNs)

- Capture intricate connections and dependencies between nodes and edges
- Message Passing Neural Network (MPNN) framework

$$\text{Aggregate neighbors' info}: \quad a^{(l)}(v) = \mathsf{AGGREGATE}\left(\{h^{(l-1)}(u) | u \in \mathcal{N}(v)\}\right)$$

$$\text{Feed forward}: \quad h^{(l)}(v) = \mathsf{UPDATE}\left(h^{(l-1)}(v), a^{(l)}(v)\right)$$



Link Prediction Task

- Encoder-decoder framework
  - GNN for deep embedding

$$\text{ENC} : \mathbf{Z} = GNN(\mathcal{V}, \mathcal{E})$$

  - Inner-product decoder by looking up the embedding matrix $\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}| \times d}$ For node pairs $(\mathbf{z}_u, \mathbf{z}_v)$

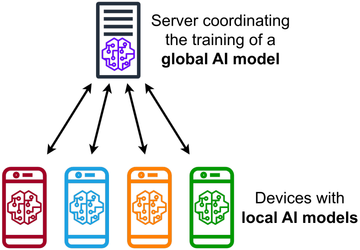
$$\text{DEC} : \hat{y}_i = DEC(\mathbf{z}_u, \mathbf{z}_v)$$

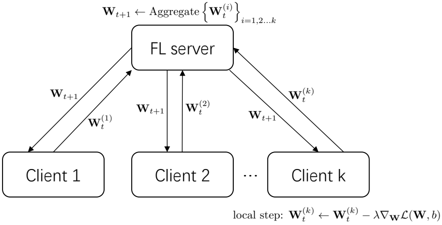
  - Loss function: mean square error (MSE) for regression and cross-entropy for classification

$$\mathcal{L} = \sum_{(u,v) \in \mathcal{D}} loss(\mathbf{z}_u^{\mathrm{T}} \mathbf{z}_v, y_i)$$

Federated Learning (FL)

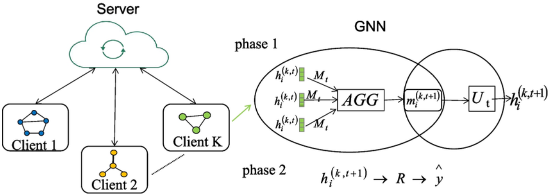
- Multiple parties collaboratively construct a machine learning model while keeping their private data locally
- Handle privacy concerns of sensitive data sources, such as bank transaction systems, healthcare industries, and social networks





FedAvg algorithm (McMahan et al., 2017)

Our Implementation: FedGNNs



Federated Learning + GNNs

- Applications
  - Recommendation system prediction: link regression
  - Anti-Money Laundering (AML) detection: link classification
- GitHub: `https://github.com/Worshipper6/Industrial-project`

Table of Contents

Data Sources

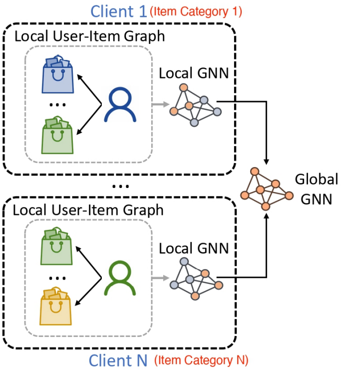
- Epinions
  - User ratings for various products on shopping website `Epinions.com`
  - 22166 users, 53717 ratings on a numerical scale ranging from 1 to 5
  - Products belongs to 27 categories, such as Electronics, Kids & Family, Restaurants
- MovieLens
  - User ratings for movies colledted by GroupLens Research Group
  - 610 users, 100836 5-star ratings
  - Each movie has 3 features, including movie names, genres and release years

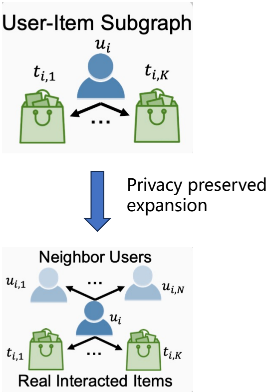
Federated Settings

- We adopt GCN, GAT, GraphSage as GNN models

Epinions: Items that belong to the same category form a client, in total 27 clients

MovieLens: Each user acts as a client, in total 610 clients



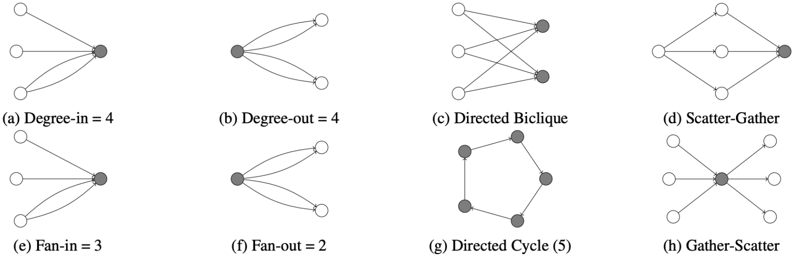


(He et al., 2021; Wu et al., 2021)

Table of Contents

Motivation

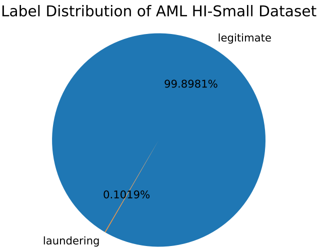
- Complex structure of criminals' money laundering patterns
- Legal and privacy concerns of bank transaction data
- Highly imbalance of data labels (legitimate or laundering)

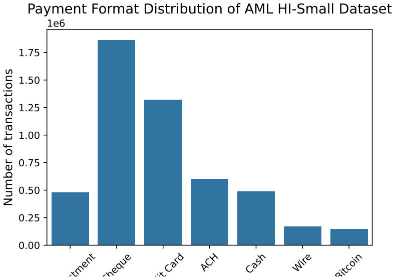


Graph-structure Money Laundering Patterns (Egressy et al., 2023)

Data Sources

- Synthetic transaction data from IBM, available at Kaggle
- 4268525 transaction records from 336889 bank accounts, 970 banks
  Each record has four features
- Our task is to predict whether a transaction is laundering or not





Models for Directed Graphs

- We adopt GAT, RGCN, PNA as GNN models
- We add the Reverse Message-Passing Mechanism for directed graphs

  Aggregate incoming neighbors

  $$a_{in}^{(l)}(v) = \mathsf{AGGREGATE}\left(\{h^{(l-1)}(u) | u \in \mathcal{N}_{in}(v)\}\right)$$

  Aggregate outgoing neighbors

  $$a_{out}^{(l)}(v) = \mathsf{AGGREGATE}\left(\{h^{(l-1)}(u) | u \in \mathcal{N}_{out}(v)\}\right)$$

  Feed forward

  $$h^{(l)}(v) = \mathsf{UPDATE}\left(h^{(l-1)}(v), a_{in}^{(l)}(v), a_{out}^{(l)}(v)\right)$$

  (Egressy et al., 2023)
- Federated setting: 5 uniform partitions of original transactions

Table of Contents

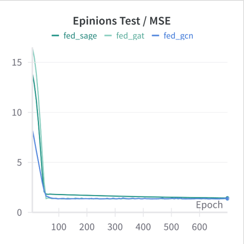
Link Regression

- FL performance is close to centralized training
- GCN tends to be robust in FL and local training
- Similar performance of three models

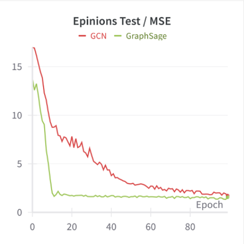
Recommendation system prediction: Test mean square error (MSE)

Datasets	**FL**			**Centralized**		
(MSE)	GCN	GAT	SAGE	GCN	GAT	SAGE
Epinions	1.374	**1.359**	1.443	**1.317**		1.360
MovieLens	**1.052**	1.088	1.133	**0.822**	0.880	

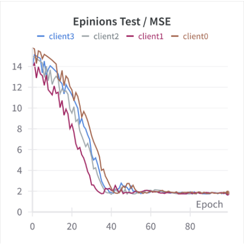
Link Regression



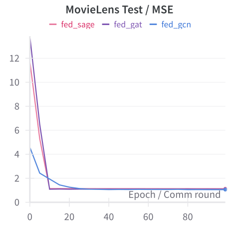
Epinions FL



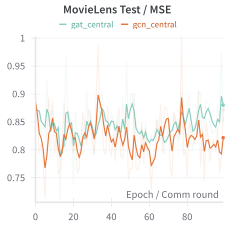
Epinions central



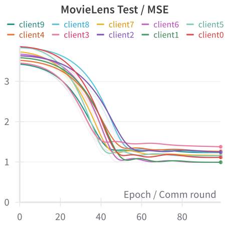
Epinions local



MovieLens FL



MovieLens central



MovieLens local

Link Classification

- FL performance is undesirable in all models under Message Passing Neural Network (MPNN) framework
- We observe dramatic improvements in FL after adding Reverse Message Passing

AML Minority class F1 score

F1 score	GAT	RGCN	PNA	r-PNA [1]
FL	0	0.1187	0.1679	**0.3766**
Centralized	0	0.3042	0.4181	**0.4814**
Local (avg)	0	0.0823	0.1015	**0.2345**

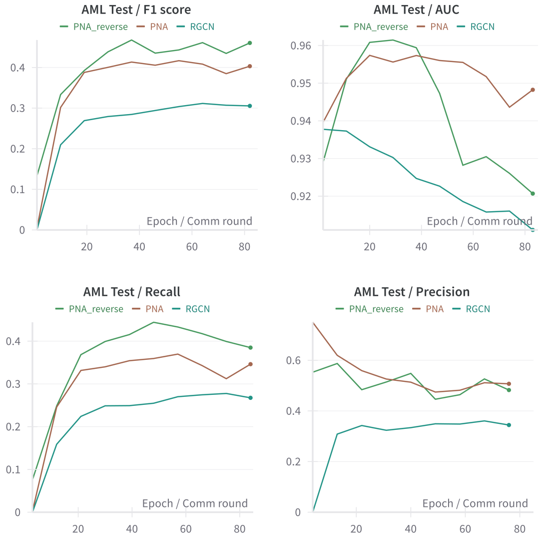
[1] With Reverse Message Passing

Link Classification

- AML detection Federated Learning (FL) performance

Link Classification

- AML detection Centralized Training performance



Link Classification

- AML detection Client Local Training performance (r-PNA)

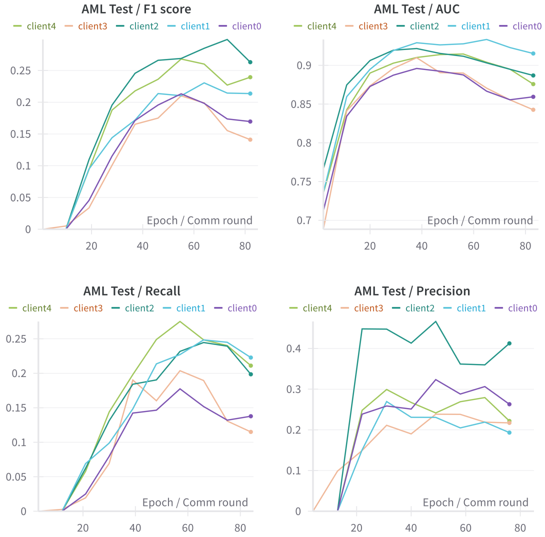


Table of Contents

Discussion

- Summary
  - The effectiveness of FL varies in different scenarios
  - GNN model selection does not dominate the performance of two recommendation system prediction tasks
  - Reverse Message Passing is important in Federated AML detection

Discussion

- Summary
  - The effectiveness of FL varies in different scenarios
  - GNN model selection does not dominate the performance of two recommendation system prediction tasks
  - Reverse Message Passing is important in Federated AML detection
- Future study
  - Add port numbering and ego IDs to improve AML detection
  - Simulate more patterns in FL settings, such as non-IID distributions

Thanks for listening!

References

Egressy, B., Von Niederhäusern, L., Blanusa, J., Altman, E., Wattenhofer, R., & Atasu, K. (2023). Provably powerful graph neural networks for directed multigraphs. *arXiv preprint arXiv:2306.11586*.

He, C., Balasubramanian, K., Ceyani, E., Yang, C., Xie, H., Sun, L., … others (2021). Fedgraphnn: A federated learning benchmark system for graph neural networks. In *Iclr 2021 workshop on distributed and private machine learning (dpml).*

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273–1282).

Wu, C., Wu, F., Cao, Y., Huang, Y., & Xie, X. (2021). Fedgnn: Federated graph neural network for privacy-preserving recommendation. *arXiv preprint arXiv:2102.04925*.