Federated Graph Neural Networks: Overview, Techniques and Challenges

来源：arXiv 2022

1. Abstract

In this paper, we bridge this gap by offering a comprehensive survey of this emerging field. We propose a unique 3-tiered taxonomy of the FedGNNs literature to provide a clear view into how GNNs work in the context of Federated Learning (FL).
本文提出了一种独特的 FedGNNs 文献的三层分类法，提供了一个关于GNN 在联邦学习(FL)环境中如何工作的清晰视角。

2.1. Terminology （专业术语）

- GNN
  - adjacency matrix: $\mathbf{A} \in \mathbb{R}^{N \times N}$
  - 节点特征 node features: $\mathbf{X} \in \mathbb{R}^{N \times f}$
- FL
  - *clients*: data owners with sensitive local data
  - *server*: 协调 clients

值得注意的是 GNN 和 FL 中都有 Aggregation 这个概念

GNN Aggregation

- 给定一个节点，通过聚合其邻居节点的信息来更新它的嵌入, Aggregation 操作可以是 **mean, weighted average, or max/min pooling methods**

FL Aggregation

- 服务器（用某种算法，**eg. FedAvg**）根据数据方的上传的本地模型参数去聚合更新全局模型的参数

2.2 The Proposed 3-Tiered FedGNN Taxonomy

Figure 1 展示了此篇论文提出的三层分类法

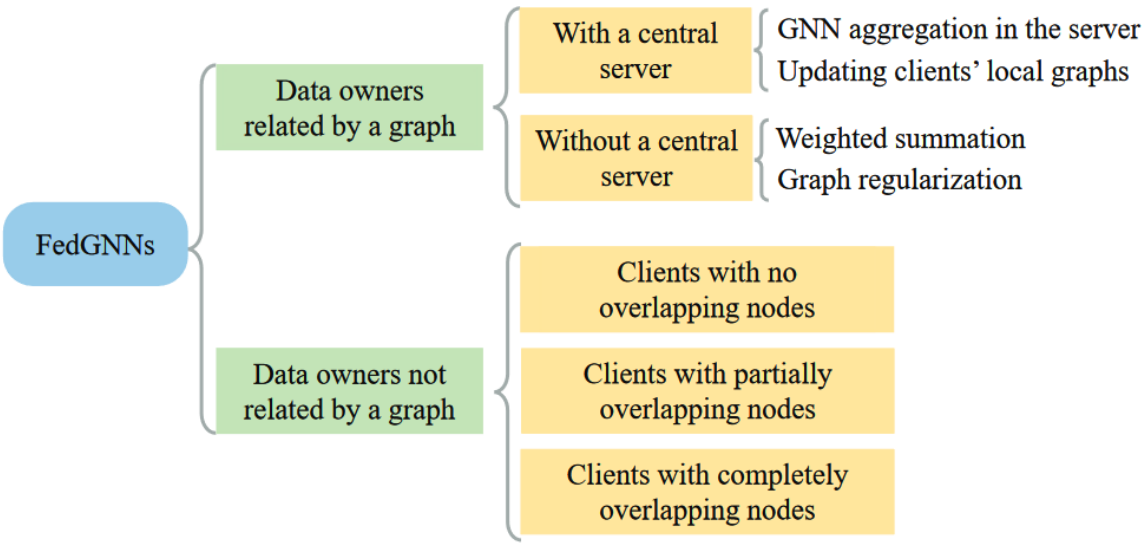

Figure 1: The Proposed 3-Tiered FedGNN Taxonomy.

data owners are related by a graph topology（第一大类）

- a bank can be a data owner which holds many accounts as the nodes in the graph and transactions between accounts as the edges. The bank can be related to other banks through transaction edges among accounts held by different banks. Note that as long as the data owners are related by a graph topology, the format of the local data does not necessarily need to be a graph.
- 一个银行可以是一个数据方，它掌握了许多账户（节点）和账户间的交易（边）信息。银行可以与其他银行通过交易产生关联。即数据方在图拓扑上产生了关联。

data owners are not related by a graph topology（第二大类）

- multiple e-commerce companies each having user-item browsing data represented as graphs collaborate to train a recommender system model leveraging FedGNN
- 我的理解是，多家电子商务公司，每家公司都有<用户, 物品>的购买信息，其可以表示成一张张图，但是这些购买信息之间并没有关联，也就是说多个数据方之间没有在图拓扑上产生关联。

3 data owners related by a graph(第一大类)

3.1 FedGNNs with a Central Server

Figure 2 是有中心服务器的 FedGNN 图示

- 论文中提到客户端的本地数据不一定要是图数据（这一点我不太明白为啥）。
- 服务器通过**客户端之间的在图上的关系（图中虚线）** 来**协调所有客户端（图中实线箭头所示）**
- 服务器主要做两方面的协调
  - firstly：根据图的拓扑结构来执行 Aggregation
  - secondly：帮助客户端更新它们的本地图，估计出在不同客户端之间的**缺失边**（不太理解）

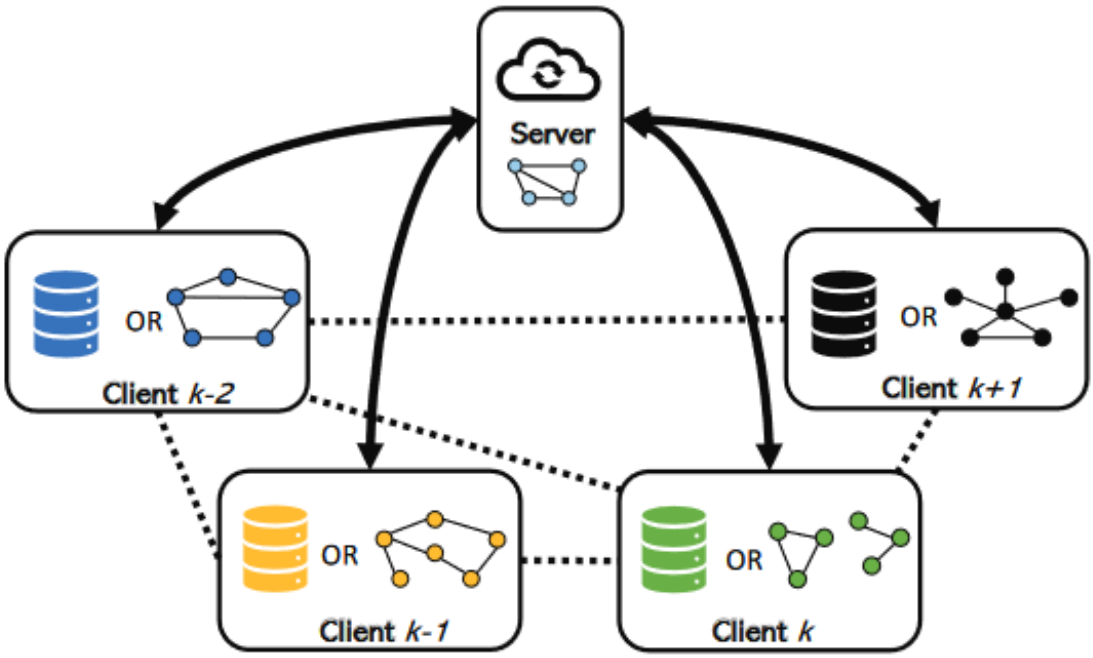

Figure 2: Illustration of a FedGNN with a central server.

GNN Aggregation by the FL Server

这一部分介绍了两篇论文的 GNN Aggregation 方式 ~~-(我看得懵懵的)-~~

1. [Meng et al., 2021]:
    - 提出了 inductive federated learning scheme
    - 利用一个交替优化过程来处理 **spatio-temporal data**
    - 具体实现(split learning):
        - 在客户端: 用 **Gated Recurrent Unit** 提取**时间特征**
        - 在服务器: encoding the connections among clients using their node embeddings with a GNN model in the server
2. [Xing et al., 2021]:
    - 用了不同的 **objective function**
    - 在内层循环中用 **local task objective functions** 训练客户端模型
    - 在外层循环中用 a separate **contrastive learning objective function** 来训练 GNN 模型
    - 服务器收集来自客户端的模型参数，作为 GCN 模型的的节点特征
    - 影响：第一个提出了 **bi-level optimization（双层优化）** 的方法，它提供了一个理论分析这个框架下的收敛

Updating Clients' Local Graphs

> Under this setting, it is assumed that each client's local data are in the form of a sub-graph of an entire graph. The relationship graph topology connecting clients is leveraged to help them update their local graphs, which benefits their local GNN model training and mitigates the non-IID data problem.
> 假设前提：
> 每个客户机的本地数据的形式是整个图的子图，根据**客户端之间的图拓扑结构**来帮助客户端更新它们的本地图。这有利于：
> （1）本地 GNN 模型的训练
> （2）减轻 Non-ID 数据问题

然后介绍了一些相关研究

3.2 FedGNNs without a Central Server

没有中央服务器的联邦图神经网络。

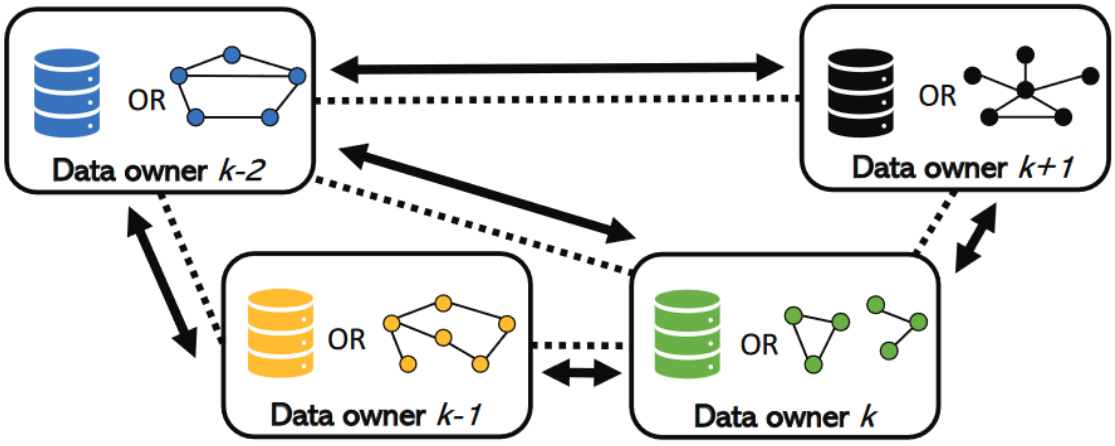

Figure 3: Illustration of a FedGNN without a central server.

- 系统中没有服务器来协调数据方
- 数据方之间直接通信（数据方之间的图拓扑结构要提前知道）
- 数据方有两种方法来更新本地模型参数：
  - 1. weighted summation 加权求和
  - 2. graph regularization 图正则化

weighted summation

$$w_i^{t+1} = \sum_{j \in N(i)} a_{ij} \cdot \left[\mathbf{w}_j^{(t)} \right]$$

where

(1) $\mathbf{w}_i^{(t+1)} \in \mathbb{R}^p$ denotes the model parameters of data owner i at iteration $(t+1)$.

(2) [.] is the encryption operation for data privacy protection.

(3) a_{ij} is the [i-th row, j-th column] element in the adjacency matrix A of the graph, which is assumed to reflect the local data distribution similarity between i and j. (数据方 i 和数据方 j 之间的本地数据分布相似度)

(4) $N(i)$ is neighborhood of i (including itself).

graph regularization（不太理解）

在此设置中，每个数据所有者将图拉普拉斯正则化纳入目标函数，**以使来自邻近客户端的模型参数相似，从而解决非IID数据问题**[Ortega et al., 2018]

公式：

$$R(\mathbf{W}, \mathbf{L}) = \mathrm{tr}\left(\mathbf{W}^T \mathbf{L} \mathbf{W}\right) = \frac{1}{2} \sum_{ij} a_{ij} ||\mathbf{w}_i - \mathbf{w}_j||^2$$

where
(1) $\mathbf{W} \in \mathbb{R}^{n \times p}$ denotes the model weights of neighboring clients.

(2) $\mathrm{tr}(\cdot)$ is the **trace operation**.
(3) $\mathbf{L} \in \mathbb{R}^{n \times n}$ is the Laplacian matrix of the graph topology between neighboring clients.
(4) a_{ij} is the edge weight in the adjacency matrix connecting data owner i and j .
(5) $\mathbf{w}_i \in \mathbb{R}^p$ denotes the model parameters in data owner i .

4 Data Owners not Related by a Graph(第二大类)

由于**来自不同领域的图数据由不同的数据参与方存储**，因此：

> *sharing graph information among data owners via the FL server can be beneficial_*
> *在不同数据参与方之间共享数据是有用的*

根据客户端间节点的重叠程度,又可以将其分为三类

1. clients with no overlapping nodes
2. clients with partially overlapping nodes
3. clients with completely overlapping nodes.

4.1 Clients with No Overlapping Nodes

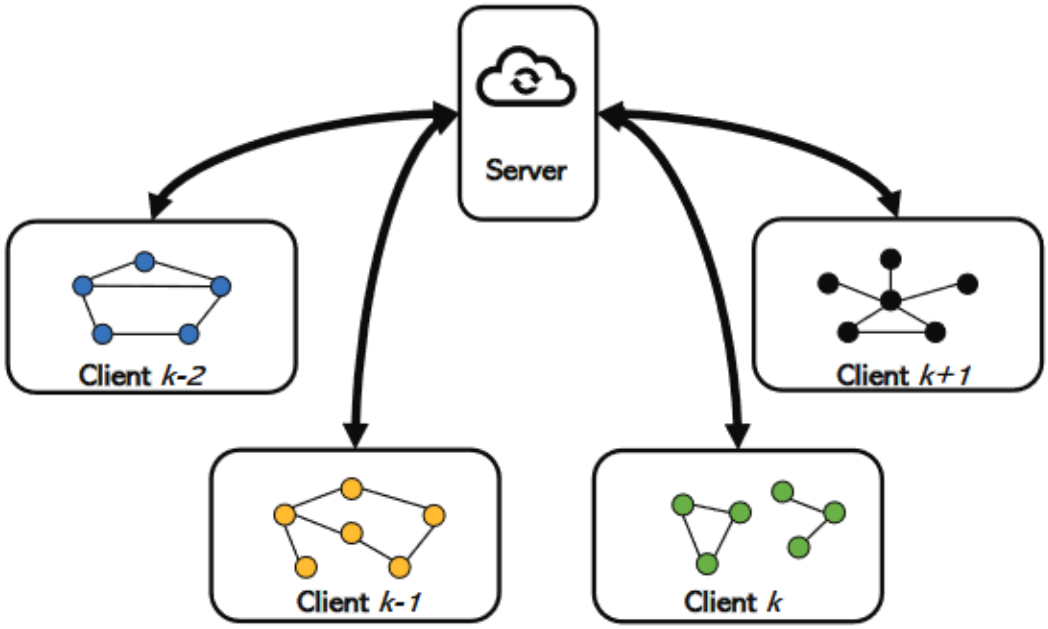

Figure 4: An illustration of clients with no overlapping nodes.

图4展示了参与方之间没有重叠节点的情况（节点由不同颜色标注），客户端用本地的图数据训练GNN，然后把模型参数上传到服务器来做 FL Aggregation。

4.2 Clients with Partially Overlapping Nodes

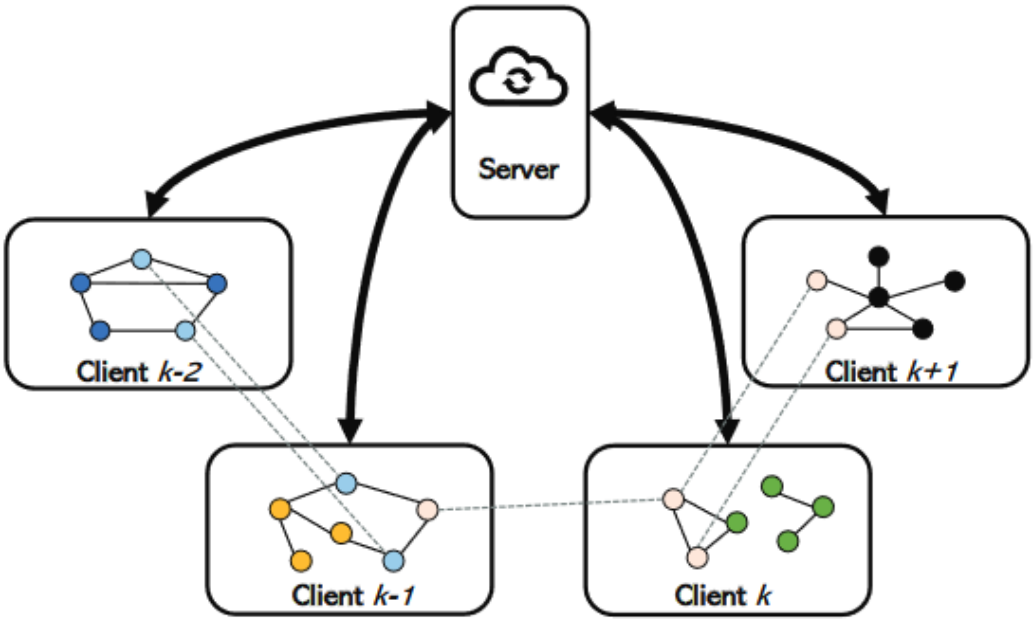

Figure 5: An illustration of clients with partially overlapping nodes.

图5展示了数据参与方的节点部分重叠的情况

4.3 Clients with Completely Overlapping Nodes

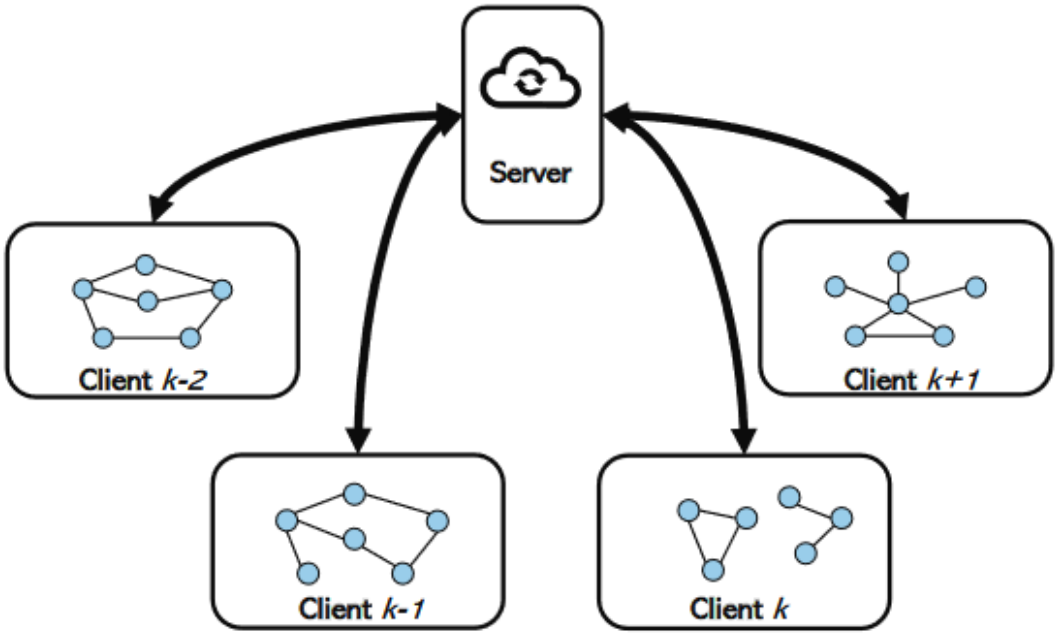

Figure 6: An illustration of clients with fully overlapping nodes.

图5展示了数据参与方的节点全部重叠的情况

每个客户端只持有部分节点特征，只有部分客户端持有学习任务的标签。所有客户端都持有相同的节点集，它们上传的是节点嵌入，而不是模型参数，到服务器做进一步的 FL aggregation

5 Promising Future Research Directions

6个有研究前景的问题

Robust FedGNNs against malicious attacks.

通过共享节点嵌入、图拓扑和模型参数，FedGNNs具有大型攻击面。

现有工作中利用的方法：

- differential privacy[Peng et al., 2021a; Zhang et al., 2021a; Wu et al., 2021]
- cryptographic methods
  - Secure Multi-Party Computation [Chen et al., 2021b; Zhou et al., 2020]
  - Homomorphic encryption [Ni et al., 2021]
  - Diffie-HellmanKey exchange [Pei et al., 2021]
  - Secret Sharing [Rizk and Sayed, 2021; Zheng et al., 2021]

they are designed to guard against only **semi-honest attackers**.

还需要探索怎么使 FebGNN 变得更加 **robust** in the face of malicious privacy attacks.

FedGNNs for dynamic graph data

动态图数据中的图拓扑或节点特征可以随时间变化。在这种情况下，在GNN训练过程中需要考虑 **temporal information(时间信息)**。

[Meng et al., 2021; Zhang et al., 2021a]. 从 从每个FL客户端内的动态图数据中提取 temporal features。然而在联邦学习的设置中，客户端之间的关系也可能随着时间的推移进化。因此还有一方面还可以探索，即能不能使得客户端之间的 connectivity 和 edge weights 是可学习的。

Efficient FedGNNs for large-scale graph data

现有的FedGNN通常使用小规模分布式数据集进行研究。因此，**通信效率**尚未得到充分考虑。然而，为了将FedGNNs扩展到大规模图数据(例如,知识图谱)，通信开销可能是一个重要的瓶颈，因为数据所有者通常采用多层GNN模型，需要传输大量的模型参数。

Explainable FedGNNs to improve interpretability

将可解释性纳入 FedGNN

Multi-hop neighborhood aggregation in decentralized FedGNNs

在现有的分布式FedGNN研究中，只有**1跳邻居**的模型参数被聚合以为每个数据所有者生成个性化的FL模型。尽管这种方法简化了模型结构，但它限制了FedGNN利用数据方之间图中丰富的邻域信息的能力。如何使FedGNN能够超越这一限制，同时保持模型结构和训练过程的简单性?

Realistic distributed graph datasets for benchmarking.

现有的 FedGNN 研究大部分是用合成的分布式图数据进行实验，这些数据是由 GNN 基准数据集生成，比如 Cora，PubMed，Citeseer，为了让数据符合 FL 的设置，现有的做法是将整张图划分为多个子图，再分配给不同的数据方，然而，这样**每个数据方分配到的图往往较小**

[He et al., 2021a] 提出了 an open-source platform，支持3个GNN模型和2个FL聚合方法。它收集了36个图形数据集，并将它们划分为分布式仓库，是一个有前景的FedGNN**基准测试工具**。

尽管如此，FedGNN领域的长期发展仍然需要建立**真实的大规模**联邦图形数据集，以支持在接近实际应用的环境下进行实验评估。

Real-world graph datasets, such as

- brain connectomic datasets
- molecule datasets

- recommender systems
- knowledge graphs

can be useful **starting points**.

--THANKS--