東南大學
SOUTHEAST UNIVERSITY

Federated Node Classification over Graphs with Latent Link-type Heterogeneity

Han Xie, Li Xiong, Carl Yang

WWW'23

李自超

提纲

- 研究背景
- 框架设计
- 实验设计
- 总结与讨论

提纲

研究背景

图神经网络及应用

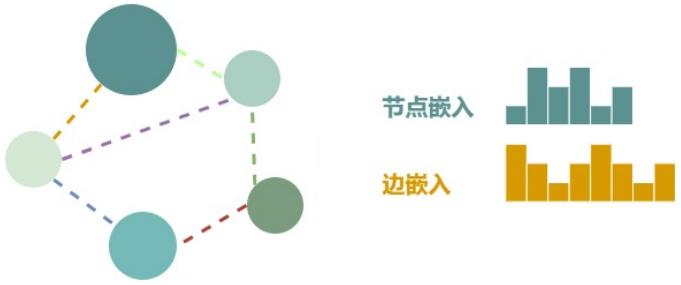
联邦图神经网络

联邦图链接异质性

动机与挑战

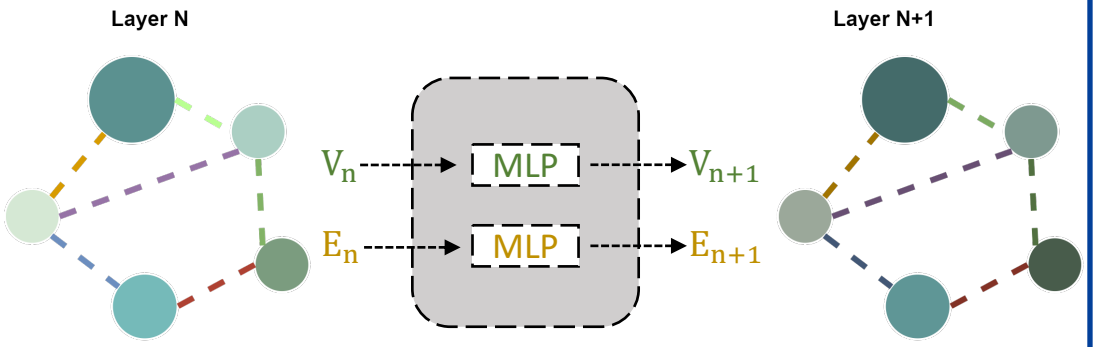
图神经网络

Graph Neural Network（GNN）



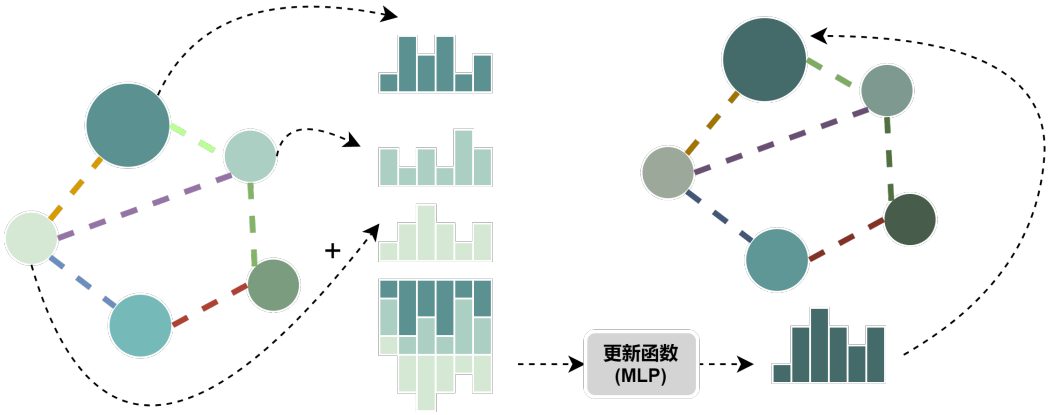
特征提取

- **节点特征**
- **边特征**



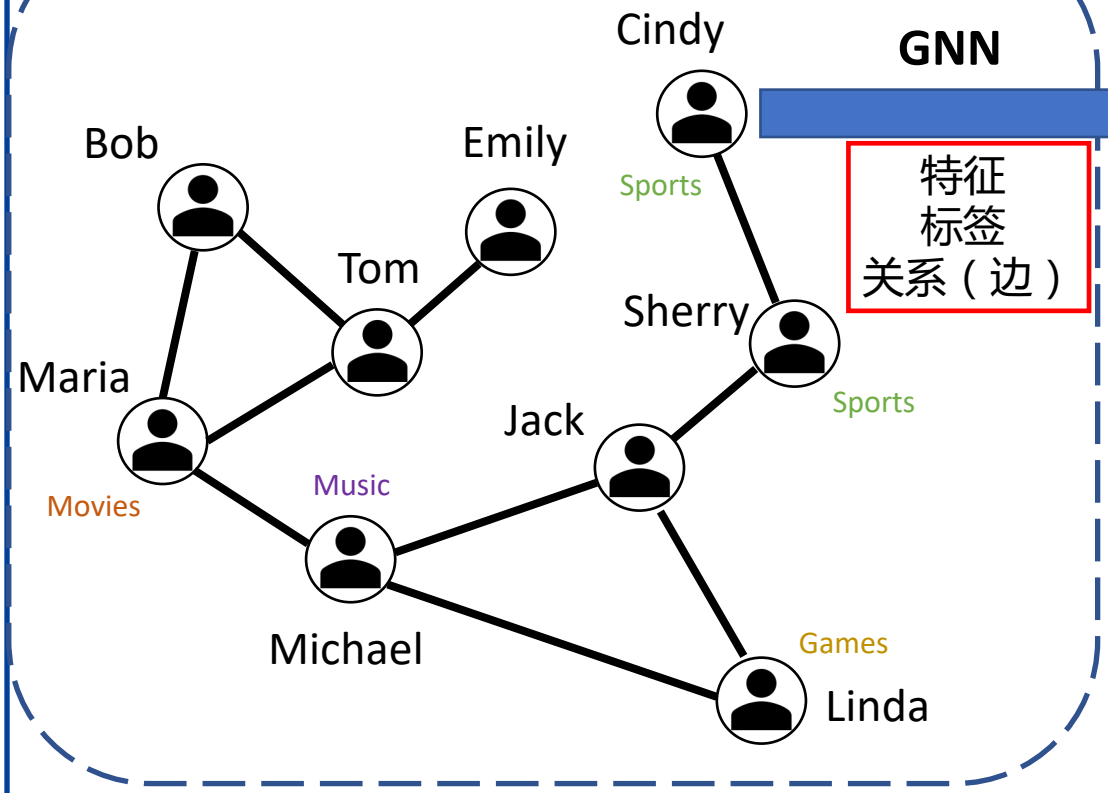
消息传递

- **获取邻居节点信息**
- **聚合邻居节点信息**
- **更新中心节点信息**



图神经网络应用

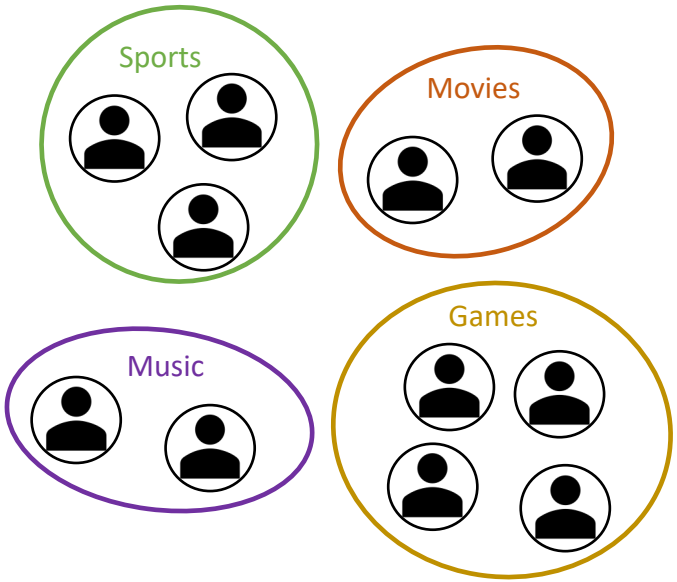
开发某 App 的公司收集到的数据



GNN

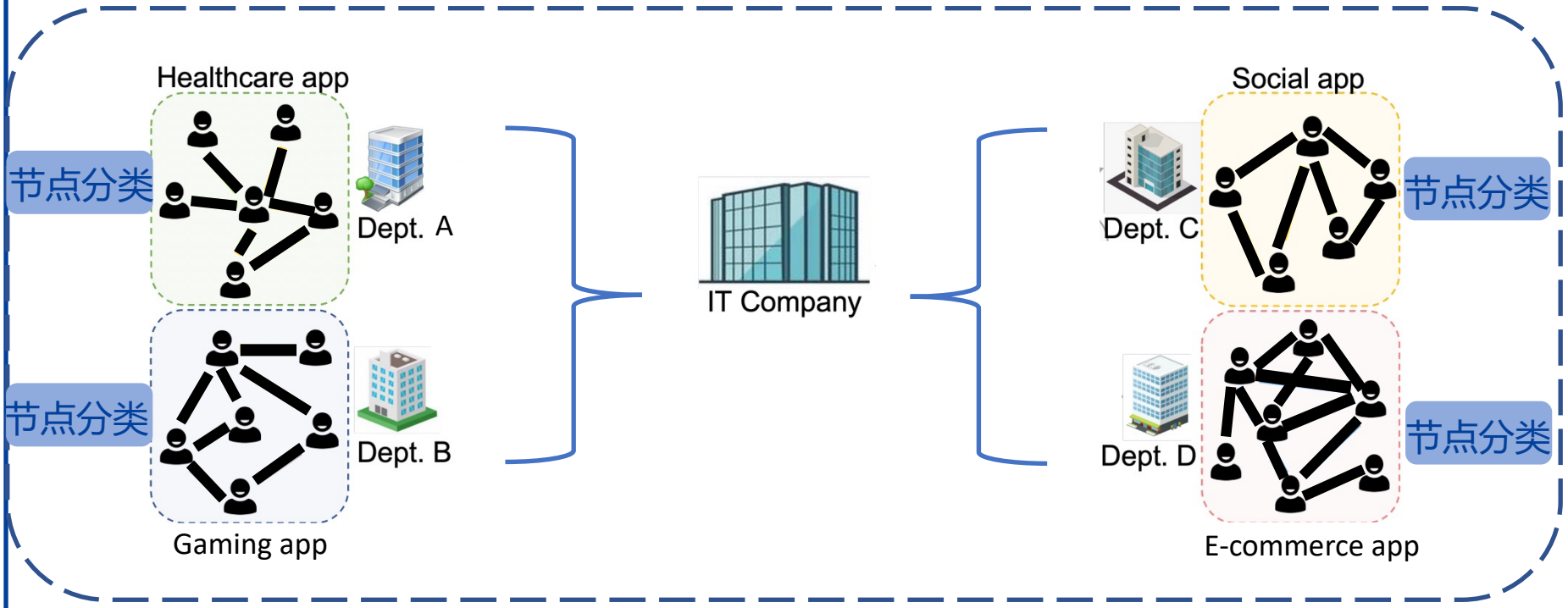
特征
标签
关系（边）

节点分类



社交网络图
（单关系图，边为好友关系）

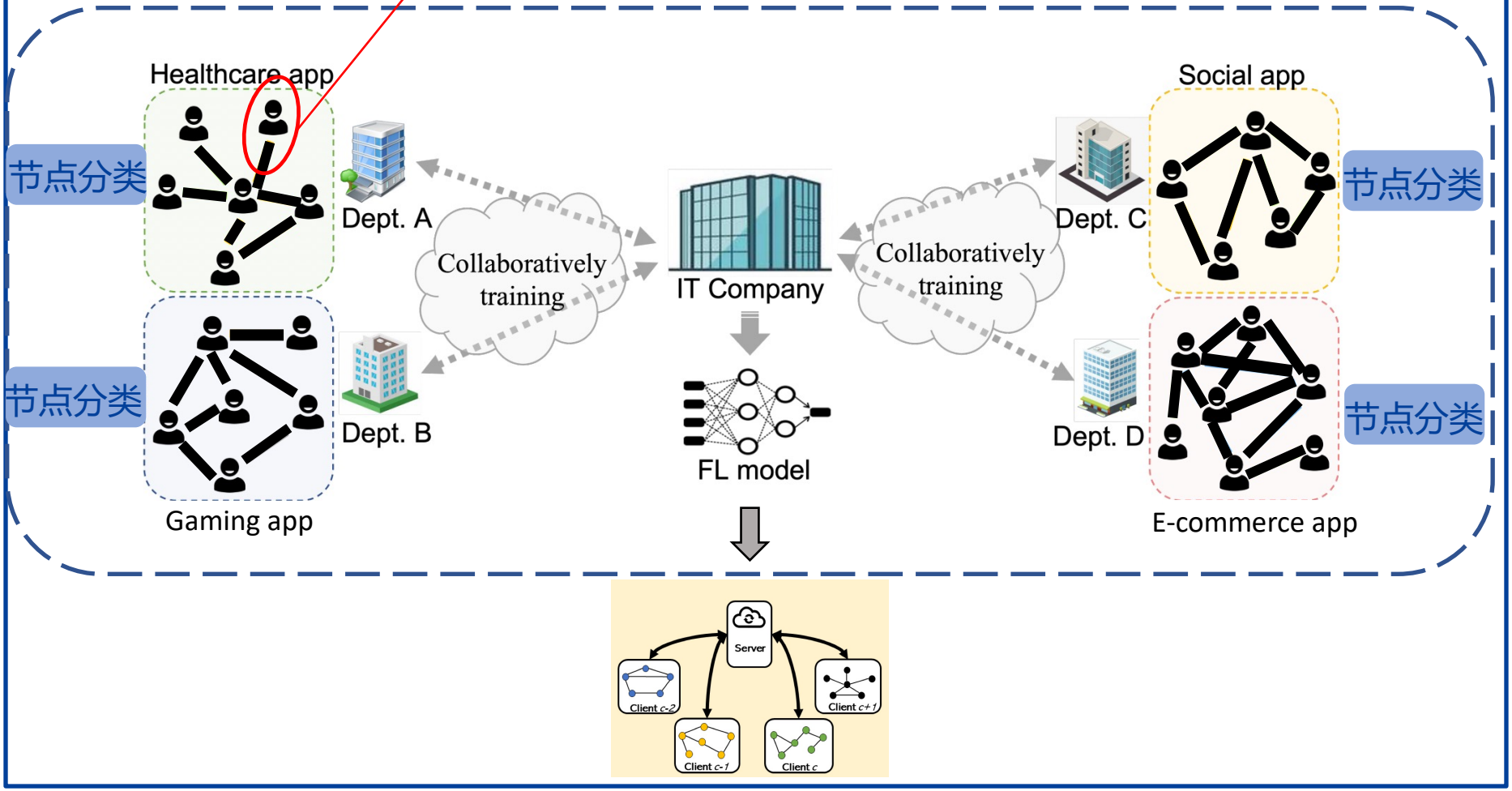
图神经网络应用



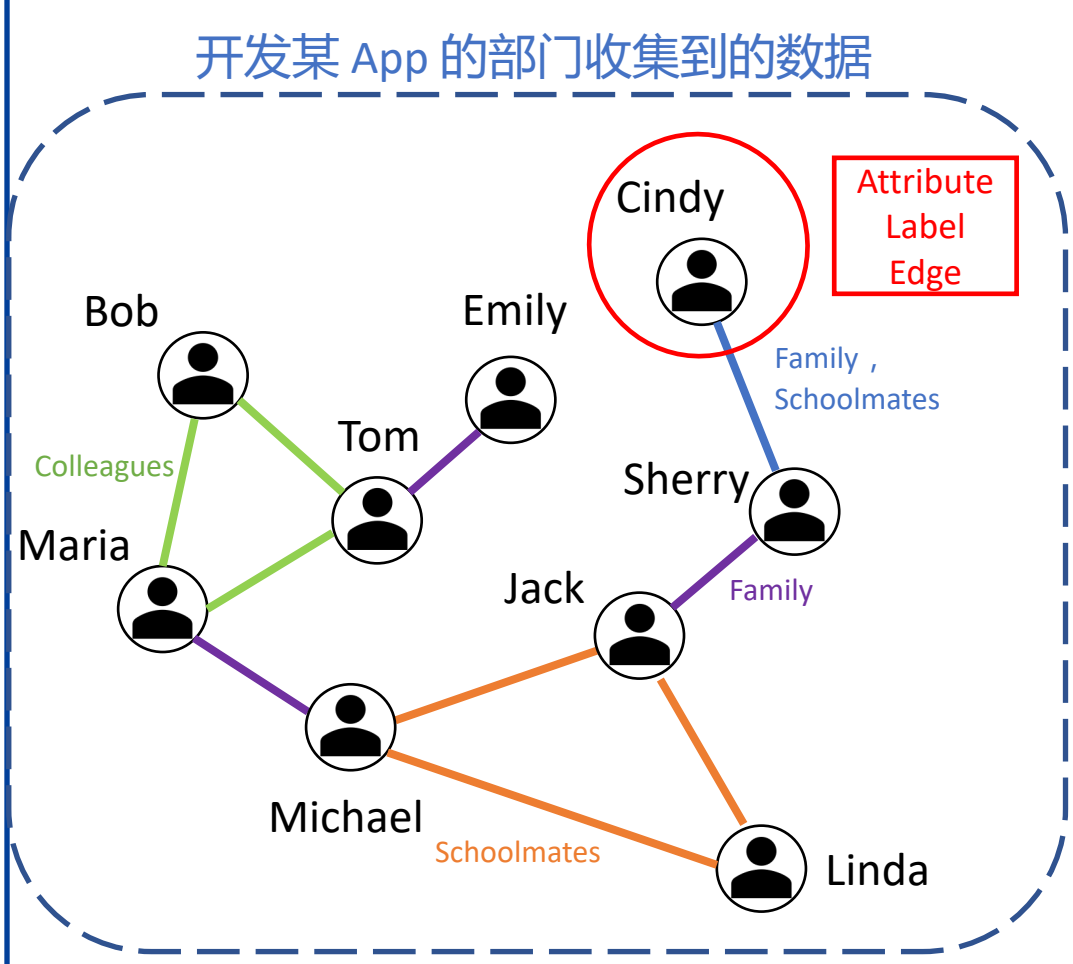
如何提升每个部门开发 App 的效果？

联邦图神经网络

节点：标签、特征、是否有相连边（及其显式属性）



边同质图的隐式链接



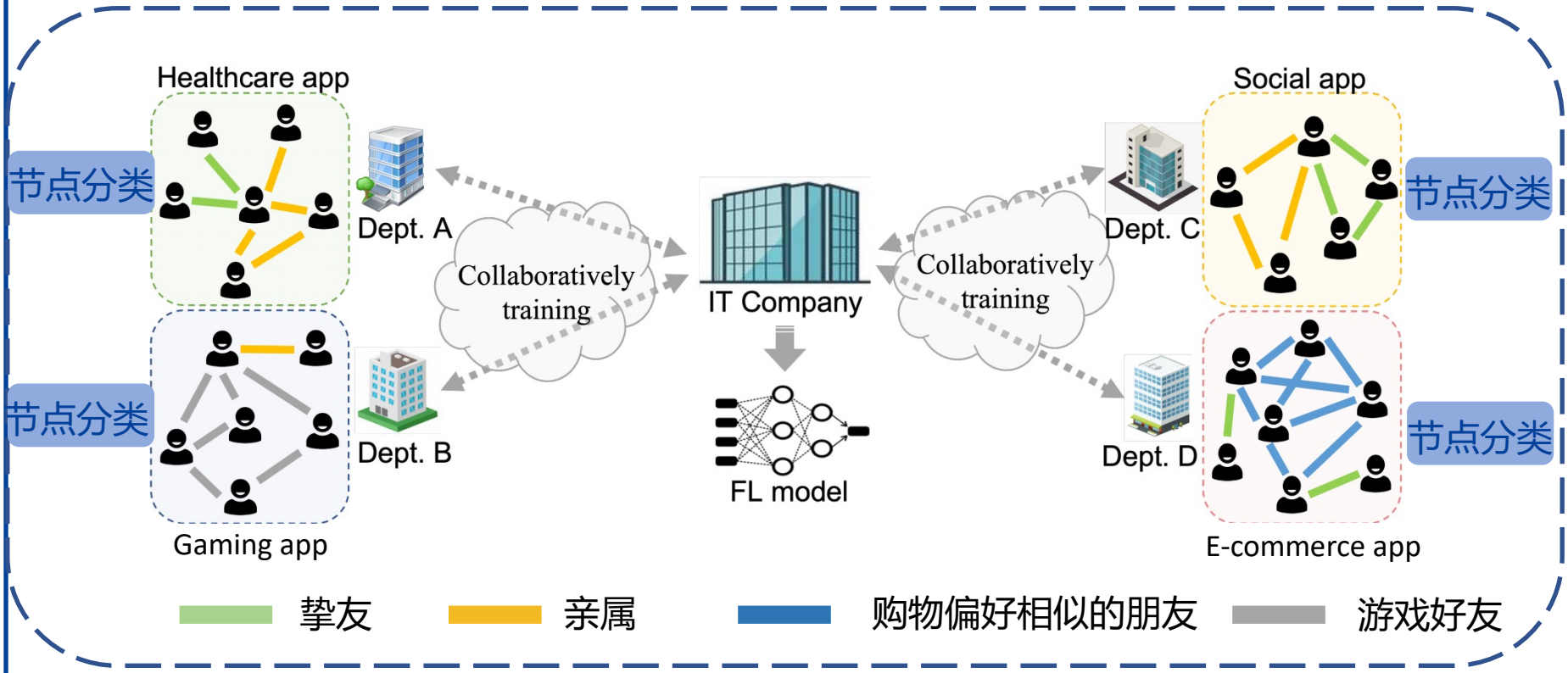
GNN

节点分类

节点标签、特征，是否有相连边

节点标签、特征，边的隐藏属性

基于隐式链接的联邦图

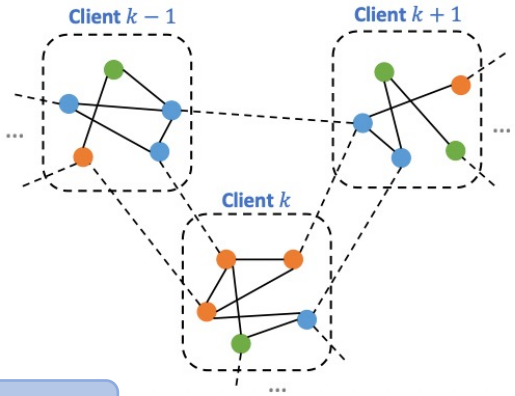

Healthcare app
节点分类
Dept. A
Collaboratively training
IT Company
FL model
Collaboratively training
Dept. C
Social app
节点分类
节点分类
Dept. B
Dept. D
节点分类
Gaming app
E-commerce app
挚友
亲属
购物偏好相似的朋友
游戏好友

联邦图的链接异质性

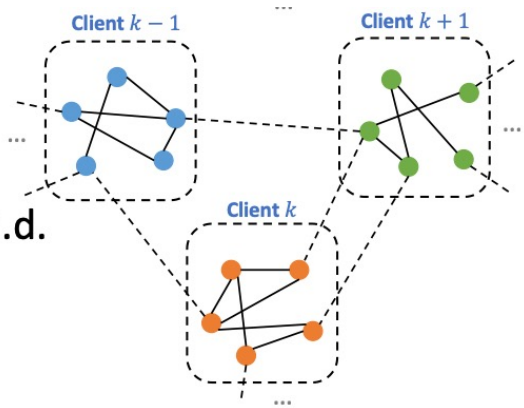
FL 中的一个重要挑战——**non-IID**

节点

i.i.d.

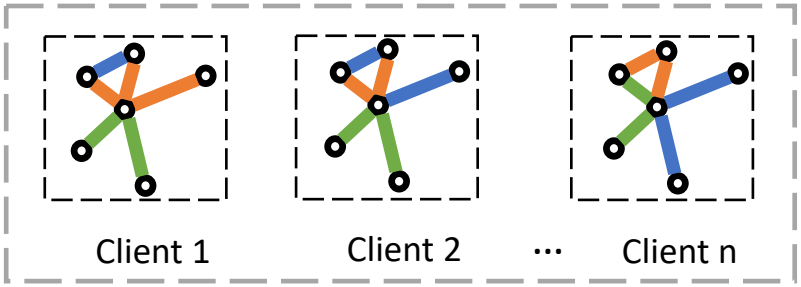


non-i.i.d.

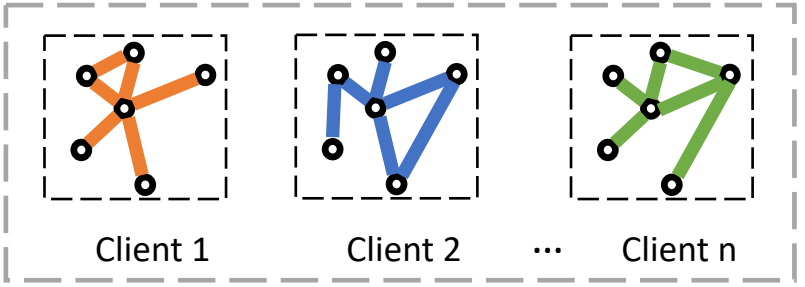


链接

i.i.d.



non-i.i.d.



FL 场景下的图的隐式链接类型的异质性尚未得到研究！

研究的必要性

在全局、 FL 场景下训练模型时，是否应该**针对不同的隐式链接类型**进行不同的处理？（即考虑**区别对待**不同隐式链接是否对模型有帮助）

- 提取 4 种隐式关系作为预定义的隐式链接类型，
- 构造 4 个单一子图（只包含一种隐式链接类型），
- 1 个混合子图（包含 4 种隐式链接类型）。

Link-type	*reference*	*share authors*	*share keywords*	*same year*	mixed

研究的必要性

针对不同隐式链接类型的处理**能否影响**图学习**模型性能**（准确率）？

全局

Link-type	reference	share authors	share keywords	same year	mixed
GCN	0.4066	0.3209	0.4159	0.1862	0.3403
mGCN	0.4025	0.3202	0.4155	0.1828	**0.4017**

多通道

每个通道对应不同链接类型进行消息传递

联邦

Link-type	reference	share authors	share keywords	same year	mixed
Fed-GCN	0.3280				N/A
Fed-mGCN	**0.4125**				N/A

简化设置：每个客户端存有具有单一隐式链接类型的子图

研究挑战

设计应对隐式链接类型异质性问题的 FL 框架时应考虑：

- 如何检测出潜在的隐式链接类型
- 如何在 FL 场景中区别对待不同隐式链接类型
- 客户端上传到服务器进行聚合时是否应针对不同隐私链接类型做不同的处理

客户端

1. 每个客户端存有包含**不同隐式链接类型**的图数据
2. 每个客户端存有包含**相同隐式链接类型**但**分布不同**（比例）的图数据

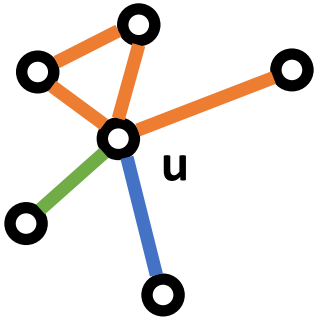
→ 表示隐式链接类型异质性的不同程度

提纲

框架设计

- m-GCN
- Fed-mGCN
- FEDLIT

多通道图卷积神经网络（m-GCN）



共享特征投影层

（θ^s）

输入原始节点 u 的特征x_u，

输出 node embedding：

$\theta^s = [\vartheta_0^s, \vartheta^s]$

$\mathrm{h}_u^s = \sigma(\vartheta_{0,u}^s + \sum_{u \in V} \vartheta_u^s x_u)$

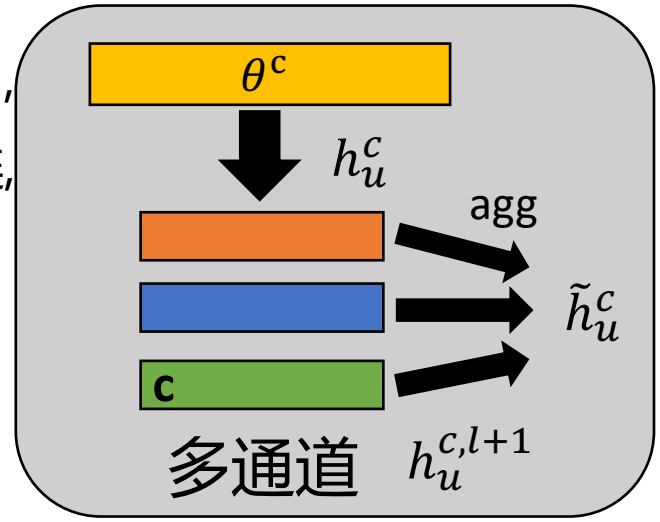
h_u^s

每个通道负责一个链接类型 c，

假设节点 u 与类型 c 的边相连，

将 h_u^s 作为输入，输出 h_u^c

在对应通道中进行 L 层图卷积



分类器

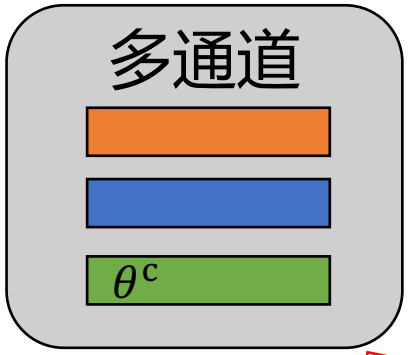
（θ^t）

下游任务

（节点分类...）

联邦多通道图卷积神经网络（Fed-mGCN）

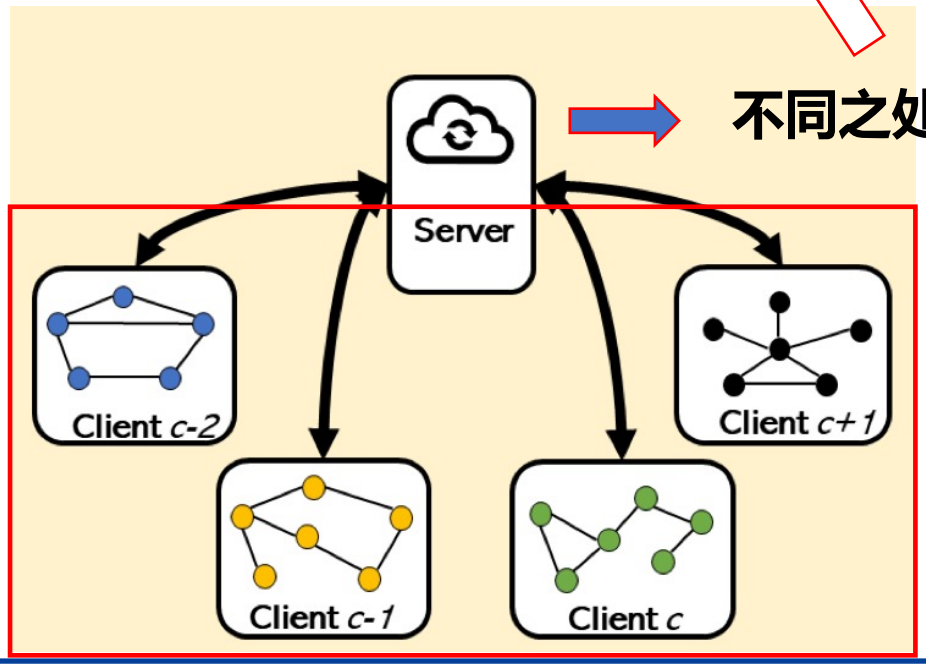
$$\theta^{c,(r+1)} = \sum_{n=1}^{|\{n\}_{\exists c \in \mathbb{C}_n}|} \frac{|V_n^c|}{|V^c|} \theta_n^{c,(r)}$$



$$\theta^{s,(r+1)} = \sum_{n=1}^{|N|} \frac{|V_n|}{|V|} \theta_n^{s,(r)} \qquad \theta^{t,(r+1)} = \sum_{n=1}^{|N|} \frac{|V_n|}{|V|} \theta_n^{t,(r)}$$

共享特征投影层（θ^s）　　分类器（θ^t）

FedAvg



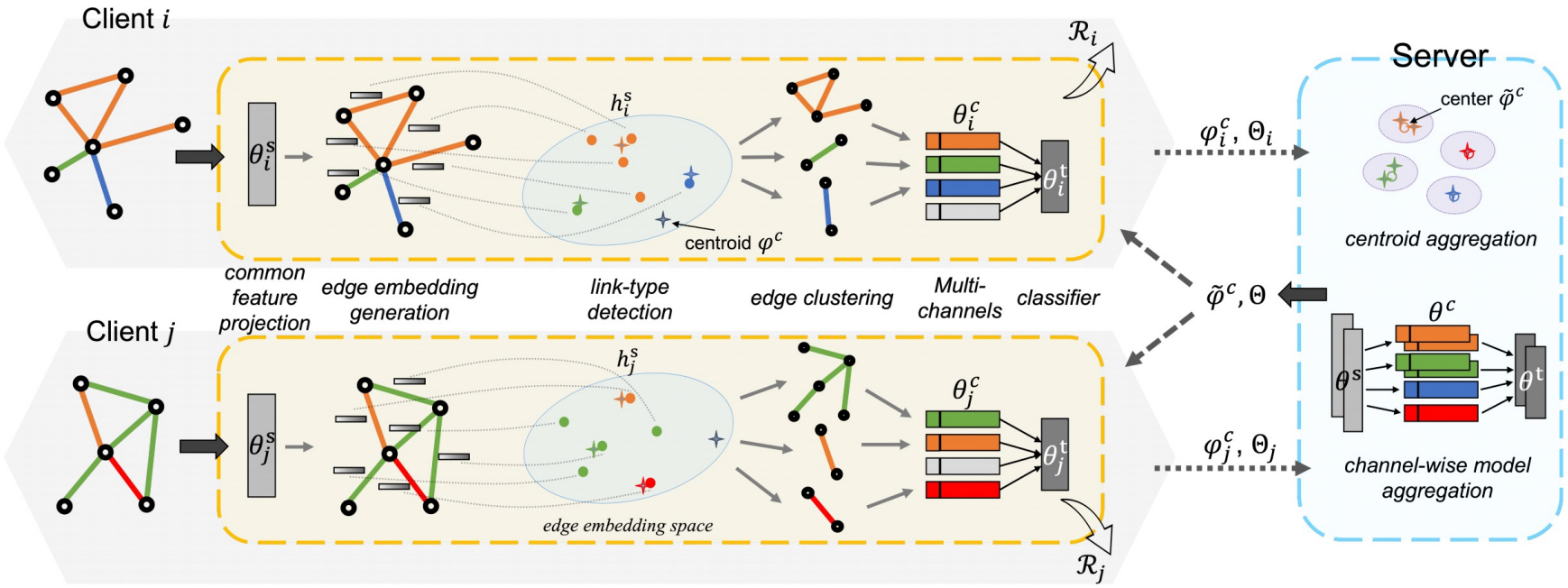
不同之处：服务器中的模型聚合方法

与普通 FL-GCN **相同**：客户端训练本地模型，再传到服务器进行通信。

本文设计的方法：FedLit

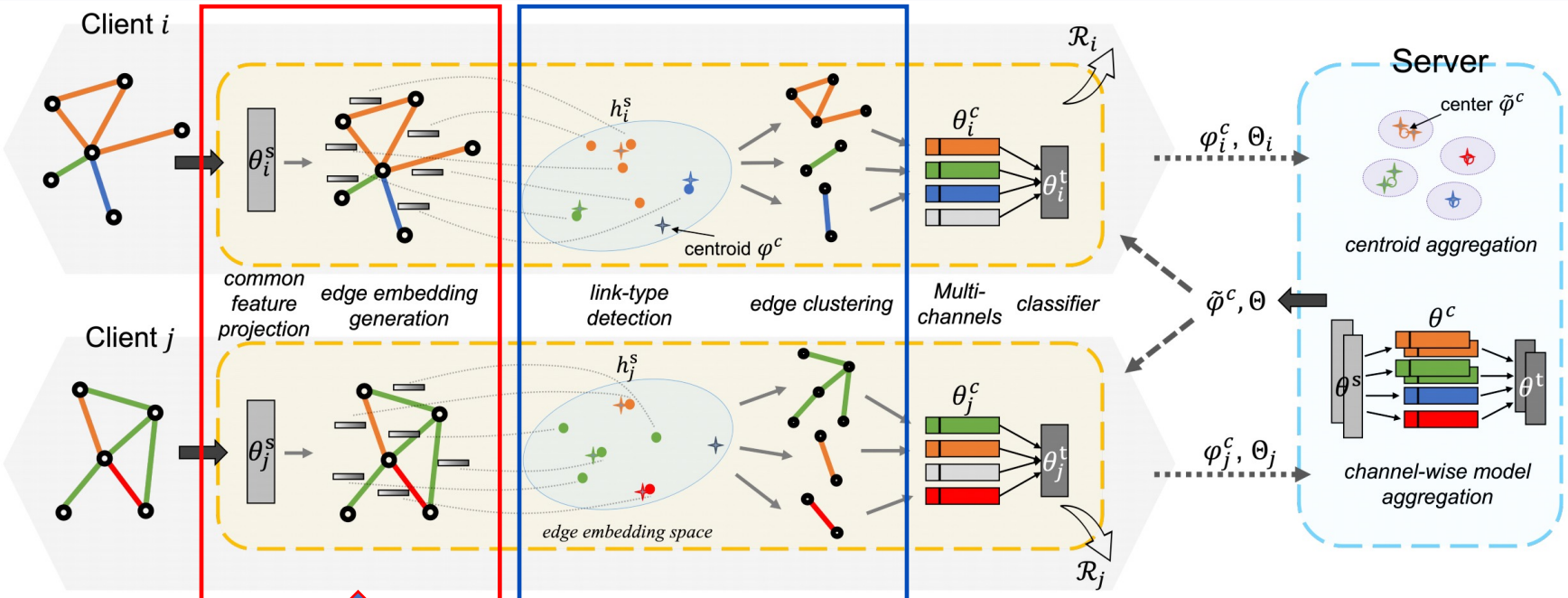
动态的隐式链接类型感知的聚类联邦学习框架

（ a dynamic latent link-type-aware clustered FL framework ）



动态检测边的隐式链接类型，并聚类相应的边

动态检测+聚类



$$\theta^s = [\vartheta_0^s, \vartheta^s]$$

$$\mathrm{h}_u^s = \sigma(\vartheta_{0,u}^s + \textstyle\sum_{u \in V} \vartheta_u^s x_u)$$

$$z_{uv} = \mathrm{h}_u^s \oplus \mathrm{h}_v^s$$

concatenate

预定义隐式链接类型数量为 k，

如果边 e_{uv}的嵌入 z_{uv} 离质心 φ^c 最近，

它将被分配给链接类型为 c 的 簇。

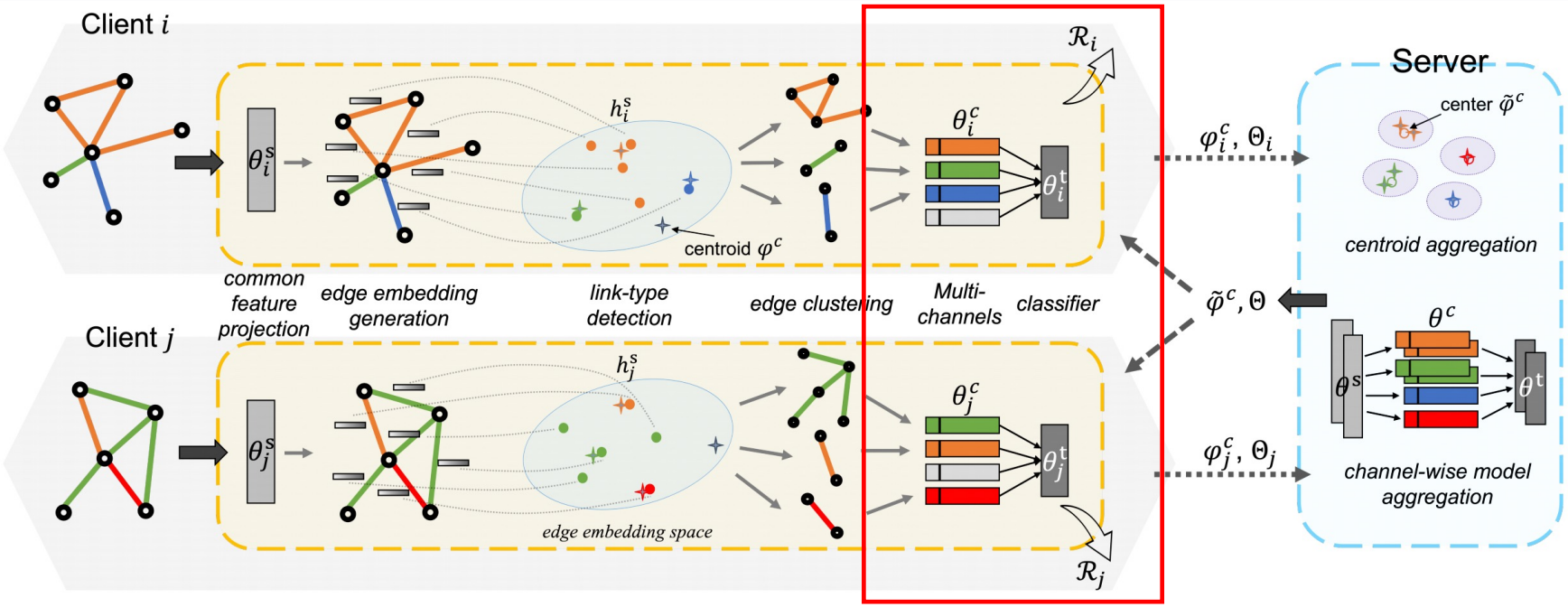
检测

使用 K-Means 聚类，结合 EM(Expectation Maximization)，

最小化簇内距离并更新质心 φ^c。

聚类

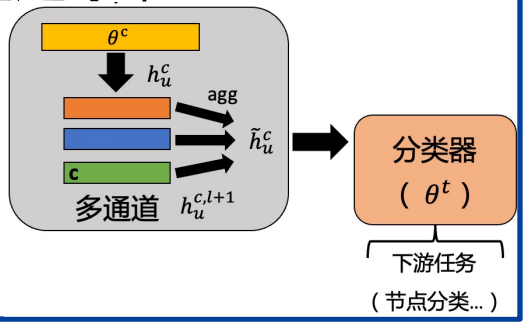
联邦动态聚类



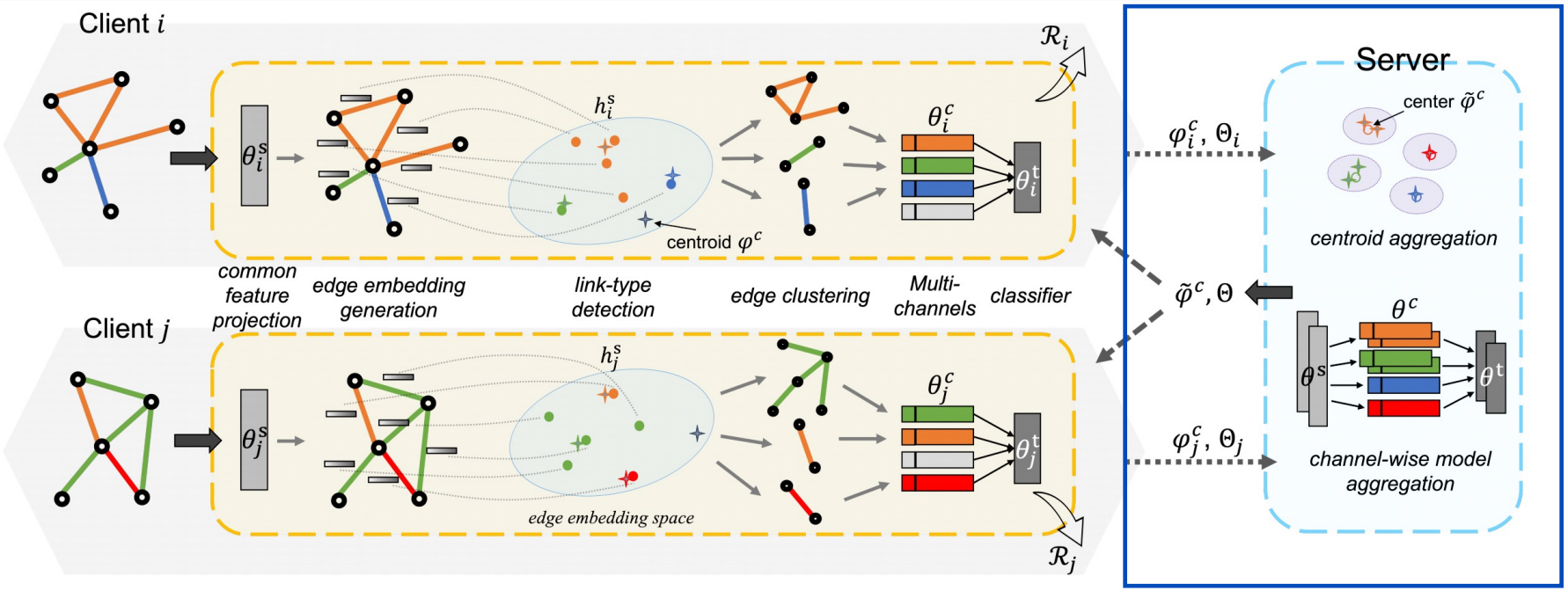
每个簇的质心 φ^c 都与本地模型中的一个通道相**对应**，

隐式链接类型为 c 的边会被传输到与质心 φ^c **相关联的通道中**，

其余操作与 m-GCN 中相同 。



联邦动态聚类



在FL通信过程中，每个客户端传输本地模型参数和更新后的 k 个质心集合到服务器。

服务器端

将收到的N×K个质心分成k个组 $\emptyset = \{\emptyset_1, \emptyset_2, \ldots, \emptyset_k\}$ 。

目标：中央每组应对应一种链接类型，与本地通道对齐，且组内距离最小化（组内的所有质心间距离最近）。

限制条件：每组内所有质心应来自不同客户端 。

如何做到？

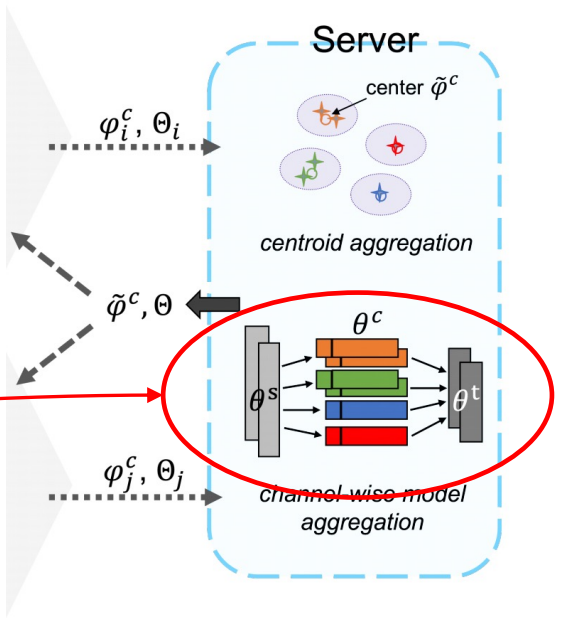
联邦动态聚类

第 1 轮

服务器从每个客户端中随机选一个质心形成一个组 ø，

共进行 k 次得到 $\{ø_1, ø_2, \ldots, ø_k\}$，

并计算各组的中心 $\tilde{\varphi}$（平均值）



接下来的每轮

聚合更新

所有质心分组完成

Similarity

$\{\varphi\}^{NK}$

$\{\tilde{\varphi}\}^k$

select

最相似的一对：φ_n - $\tilde{\varphi}_c$

assign

$ø_c$

去除所有$\{\varphi_n\} - \tilde{\varphi}_c$对

保证每个客户端只能给一个质心到某一组

客户端

提纲

实验设计

实验设置

实验结果

实验设置

出版物数据集： DBLP-DM 、 PUBMED–DIABETES

电子健康记录数据集： NELL 、 MIMIC3

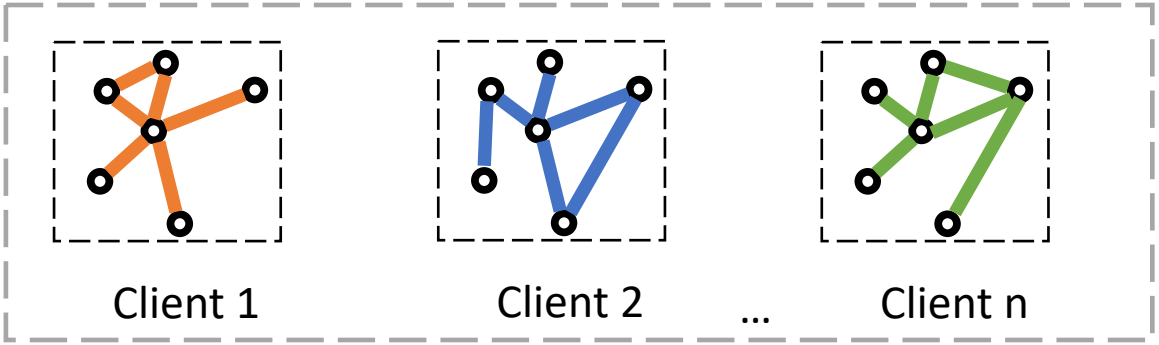
Dataset	#Node	#Feature	#Class	#Total Edge	#Oracle Edge i)	ii)	iii)	iv)
DBLP-DM	46,582	200	12	7,097,924	206,219	1,285,315	2,818,120	2,788,270
PUBMED-DIABETES	13,778	200	3	588,529	20,035	118,441	74,971	375,082
NELL	41,671	2,792	5	39,250,315	91,229	6,499,135	24,529,421	8,130,530
MIMIC3	58,495	6,671	6	30,603,469	23,068	27,244,566	2,413,231	922,604

将数据分割为三部分，以模拟不同程度的隐式链接类型异质性：

Distinct

模拟 FL 中客户端存有不同隐式链接类型的图的情况，

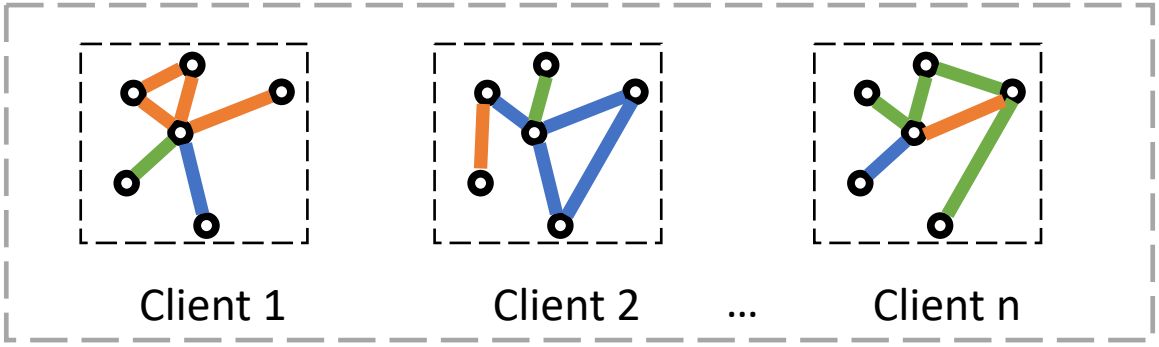
并通过让客户端仅持有一种隐式链接类型来**极端化**这种情况 。



实验设置

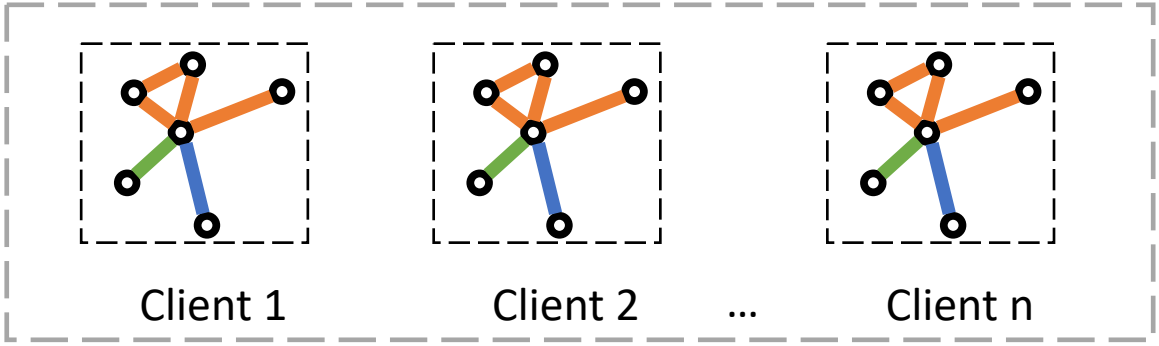
Dominant

模拟 FL 中更合理的情况：客户端存有所有隐式链接类型的图，但存在随机选择的主导隐式链接类型（其余类型的边占较小部分）。



Balanced

模拟 FL 中不太可能出现的情况：客户端上的图具有相同的边分布（最小的隐式链接类型异质性，即完全同质图）。



全局实验结果

Dataset	DBLP-DM			PUBMED-DIABETES			NELL			MIMIC3		
GCN	0.3378 (±0.0022)			0.8468 (±0.0024)			0.3758 (±0.0210)			0.3603 (±0.0021)		
cGCN [$k = \pi$]	0.3712 (±0.0042)			0.8781 (±0.0055)			0.4983 (±0.0198)			0.3655 (±0.0039)		
cGCN [$k > \pi$]	0.3884 (±0.0048)			**0.8795** (±0.0028)			0.5024 (±0.0069)			0.3715 (±0.0030)		
mGCN	**0.4148** (±0.0026)			0.8778 (±0.0023)			**0.5162** (±0.0129)			**0.3973** (±0.0028)		
	distinct	dominant	balanced	distinct	dominant	balanced	distinct	dominant	balanced	distinct	dominant	balanced
Fed-GCN	0.3033 (±0.0028)	0.3088 (±0.0025)	0.3226 (±0.0036)	0.7776 (±0.0046)	0.8008 (±0.0043)	0.8284 (±0.0029)	0.2345 (±0.0184)	0.2260 (±0.0016)	0.3491 (±0.0107)	0.3399 (±0.0055)	0.3388 (±0.0026)	0.3584 (±0.0032)
Fed-mGCN	**0.3930** (±0.0016)	**0.3779** (±0.0050)	**0.4046** (±0.0015)	0.8758 (±0.0025)	0.8657 (±0.0058)	0.8745 (±0.0022)	0.4447 (±0.0178)	0.4050 (±0.0055)	**0.5178** (±0.0065)	**0.3830** (±0.0015)	0.3474 (±0.0020)	0.3762 (±0.0011)
FEDLIT [$k = \pi$]	0.3238 (±0.0049)	0.3371 (±0.0066)	0.3400 (±0.0083)	**0.8776** (±0.0044)	0.8779 (±0.0028)	0.8771 (±0.0070)	0.4158 (±0.0071)	0.3970 (±0.0067)	0.4750 (±0.0152)	0.3552 (±0.0017)	0.3541 (±0.0037)	0.3750 (±0.0048)
FEDLIT [$k > \pi$]	0.3533 (±0.0040)	0.3701 (±0.0080)	0.3669 (±0.0062)	0.8759 (±0.0092)	**0.8820** (±0.0047)	**0.8811** (±0.0023)	**0.4715** (±0.0111)	**0.4243** (±0.0253)	0.5091 (±0.0060)	0.3685 (±0.0018)	**0.3593** (±0.0038)	**0.3765** (±0.0046)

mGCN：在已知预定义的隐式链接类型的图上训练；

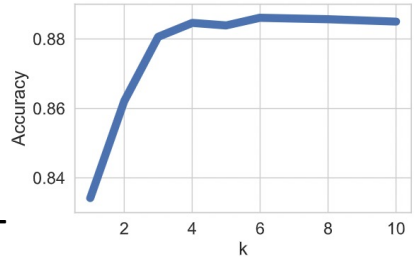
cGCN：将FedLit的聚类机制引入mGCN，无需获取其预定义的隐式链接类型。

Fed-mGCN：在已知预定义的隐式链接类型的图上进行联邦训练；

FEDLIT：无需获取预定义的隐式链接类型。

π：预定义的隐式链接类型数量

k ：设置的通道数量



FEDLIT

提纲

总结与讨论

本文工作的优势与局限性

思考

总结与讨论

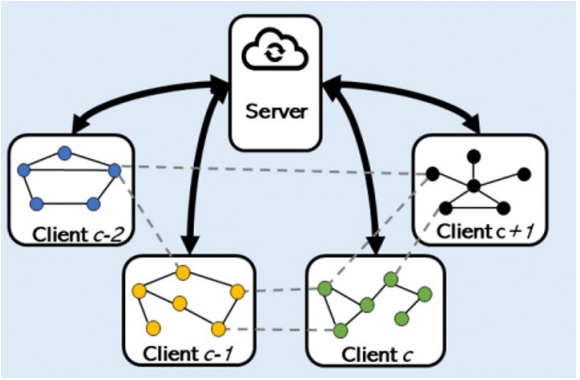
本文的工作解决了不同客户端之间隐式链接类型的异质性问题，

但仍有局限：

① 无法生成空簇

② 效率：Gu Z, Zhang K, Bai G, et al. Dynamic Activation of Clients and Parameters for Federated Learning over Heterogeneous Graphs[C]. ICDE, 2023.

③ 隐私



思考：

- 本文的场景中客户端之间没有相连边，“缺失边”的链接类型是否需要考虑？
- 单关系图中的隐式链接类型，多关系图呢？

東南大學
SOUTHEAST UNIVERSITY

欢迎老师和同学们指正！