HealSci 和兴医疗

北京大学计算机学院
School of Computer Science

Federated Learning on Graph Neural Networks

吴玉理

2023/12/1 模型组分享会

目录

- 背景介绍

- 前置知识
  - 联邦学习（FL, Federated Learning）
  - 图神经网络（GNN, Graph Neural Network）

- 论文分享（FL+GNN)
  - (2021 NIPS) Subgraph Federated Learning with Missing Neighbor Generation
  - (2022 Nature Communications) A federated graph neural network framework for privacy-preserving personalization
  - (2023 ICML) Personalized Subgraph Federated Learning

目录

- 背景介绍

- 前置知识
  - 联邦学习（FL, Federated Learning）
  - 图神经网络（GNN, Graph Neural Network）

- 论文分享（FL+GNN)
  - (2021 NIPS) Subgraph Federated Learning with Missing Neighbor Generation
  - (2022 Nature Communications) A federated graph neural network framework for privacy-preserving personalization
  - (2023 ICML) Personalized Subgraph Federated Learning

背景介绍

2022年4月9日，《中共中央、国务院关于构建更加完善的要素市场化配置体制机制的意见》中数据作为一种新型生产要素首次正式出现在官方文件，并提出要加快培育数据要素市场，其中一大方面即为**加强数据资源整合和安全保护**。数据安全行业也在**数据融合与隐私保护的双重驱动**下迅速发展。

数据融合

随着信息经济发展，数据已和其他要素一起融入经济价值创造过程，对生产力发展有广泛影响。

而伴随我国医疗健康产业迅速发展，医疗健康大数据成为新的热点。大数据的应用通过与医疗机构、高校和政府联合开展产学研合作，实现对健康医疗大数据价值的深度挖掘，开展重大专科疾病课题的研究、推动基层诊疗、智慧养老等，将成为造福民生、改善人民生活的重要部分。**数据的核心价值在于共享和应用，多方数据的融合应用具有巨大意义。**

数据安全

隐私保护

近三年来，欧盟GDPR、美国加利福尼亚州CCPA和我国的《数据安全法》（草案）、《个人信息保护法》（草案）等代表性法律法规出台，严格要求在数据使用过程中做好隐私保护。

而由于医疗数据具有的特殊性、敏感性、变现价值极高，其早已成为隐私泄露的重灾区，这也在很大程度上影响了医疗机构之间共享数据。**数据应用者迫切需要找到可靠的方法，合法合规地实现数据的共享流通。**

附草案说明

中华人民共和国
数据安全法

法律出版社

附草案说明

中华人民共和国
个人信息保护法

法律出版社

GB

中华人民共和国国家标准

信息安全技术
健康医疗数据安全指南

Information security technology—Guide for health data security

国家市场监督管理总局
国家标准化管理委员会 发布

T/GDWJ

团体标准

T/GDWJ 013—2022

广东省健康医疗数据安全分类分级管理技术规范

Technical specification for categorization and classification of health data security

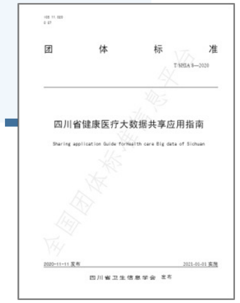
广东省卫生经济学会 发布

团体标准

广东省健康医疗数据脱敏技术规范

Technical specification for desensitization of medical data

广州市标准化促进会
广东省健康医疗大数据标准工作组 发布



团体标准

四川省健康医疗大数据共享应用指南

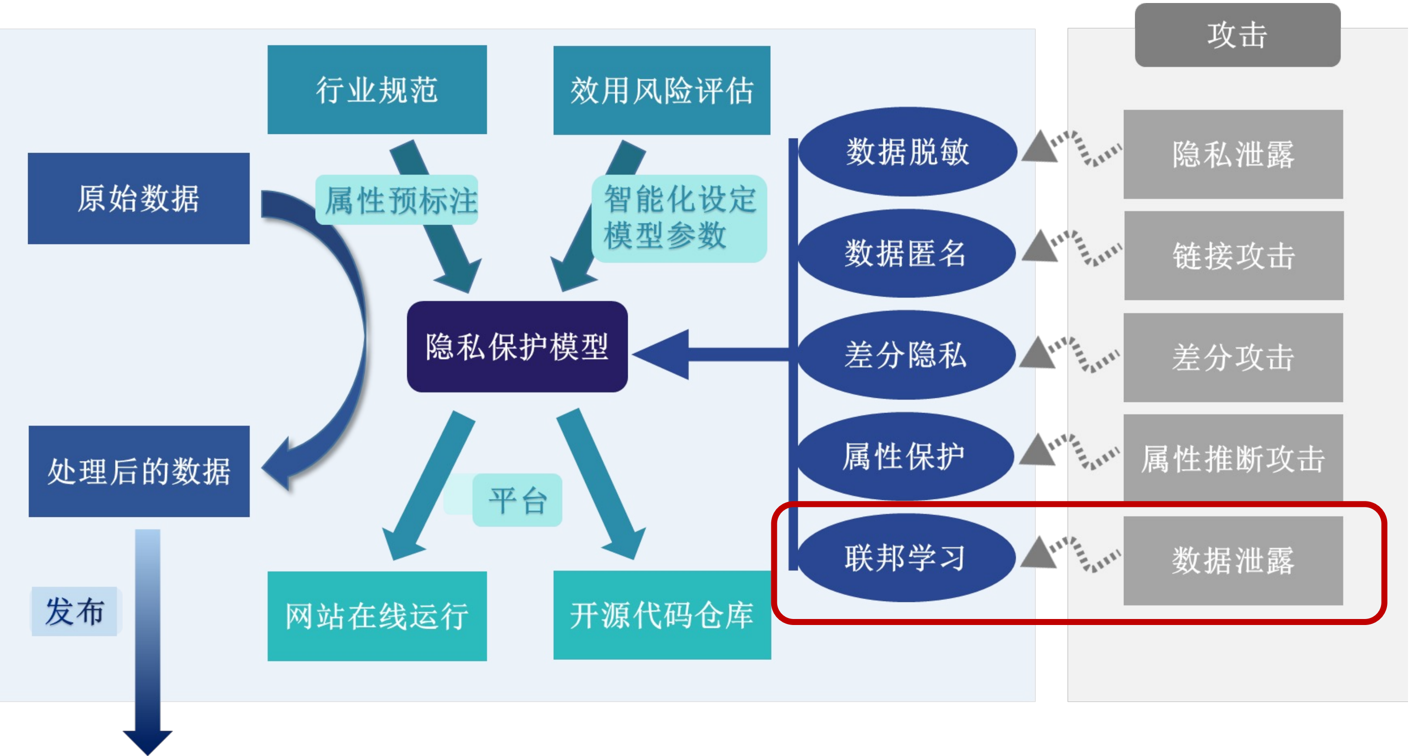
Sharing application Guide forHealth care Big data of Sichuan

四川省卫生信息学会 发布

DataGuard开源安全数据发布平台

http://101.132.17.93/

- 随着信息化在医疗行业不断深入，大数据成为医疗行业发展的驱动力，但不断涌现的隐私泄露问题，促使相关部门出台医疗行业数据规范条例。为满足医疗行业合法合规发布和使用数据的需求，本项目开发了“DataGuard”开源安全数据发布平台。
- 数据所有者遵循简单操作即可实现对医疗数据的匿名化和隐私化处理，合法合规发布数据，方便数据挖掘者获取数据，并进行后续的数据融合、分析、挖掘等环节。本平台搭载了经典的数据匿名算法和新兴的基于人工智能技术的隐私保护模型，可以抵御攻击者的链接攻击、差分攻击、属性推断攻击等，严密、高效地保护用户隐私。用户可以借助网站在线运行模型，或者下载开源代码本地运行算法，得到隐私保护处理后的数据，以及详尽的数据效用和风险评估报告。本平台实现了自动化、智能化、规范化的数据隐私保护，促进了医疗大数据的资源整合与价值迁移。



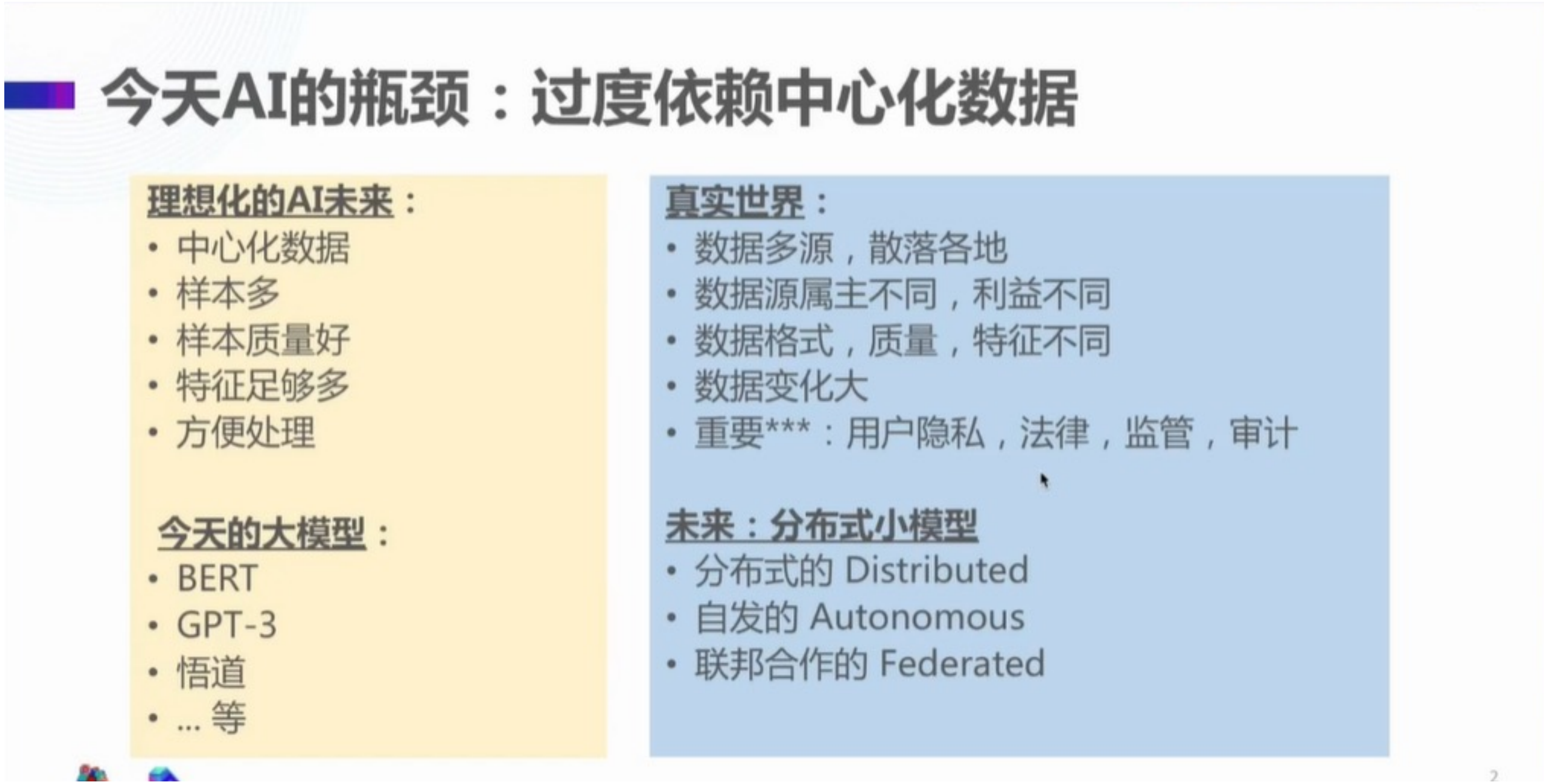
目录

- 背景介绍

- 前置知识
  - 联邦学习（FL, Federated Learning）
  - 图神经网络（GNN, Graph Neural Network）

- 论文分享（FL+GNN)
  - (2021 NIPS) Subgraph Federated Learning with Missing Neighbor Generation
  - (2022 Nature Communications) A federated graph neural network framework for privacy-preserving personalization
  - (2023 ICML) Personalized Subgraph Federated Learning

联邦学习（FL, Federated Learning）



今天AI的瓶颈：过度依赖中心化数据

理想化的AI未来：
- 中心化数据
- 样本多
- 样本质量好
- 特征足够多
- 方便处理

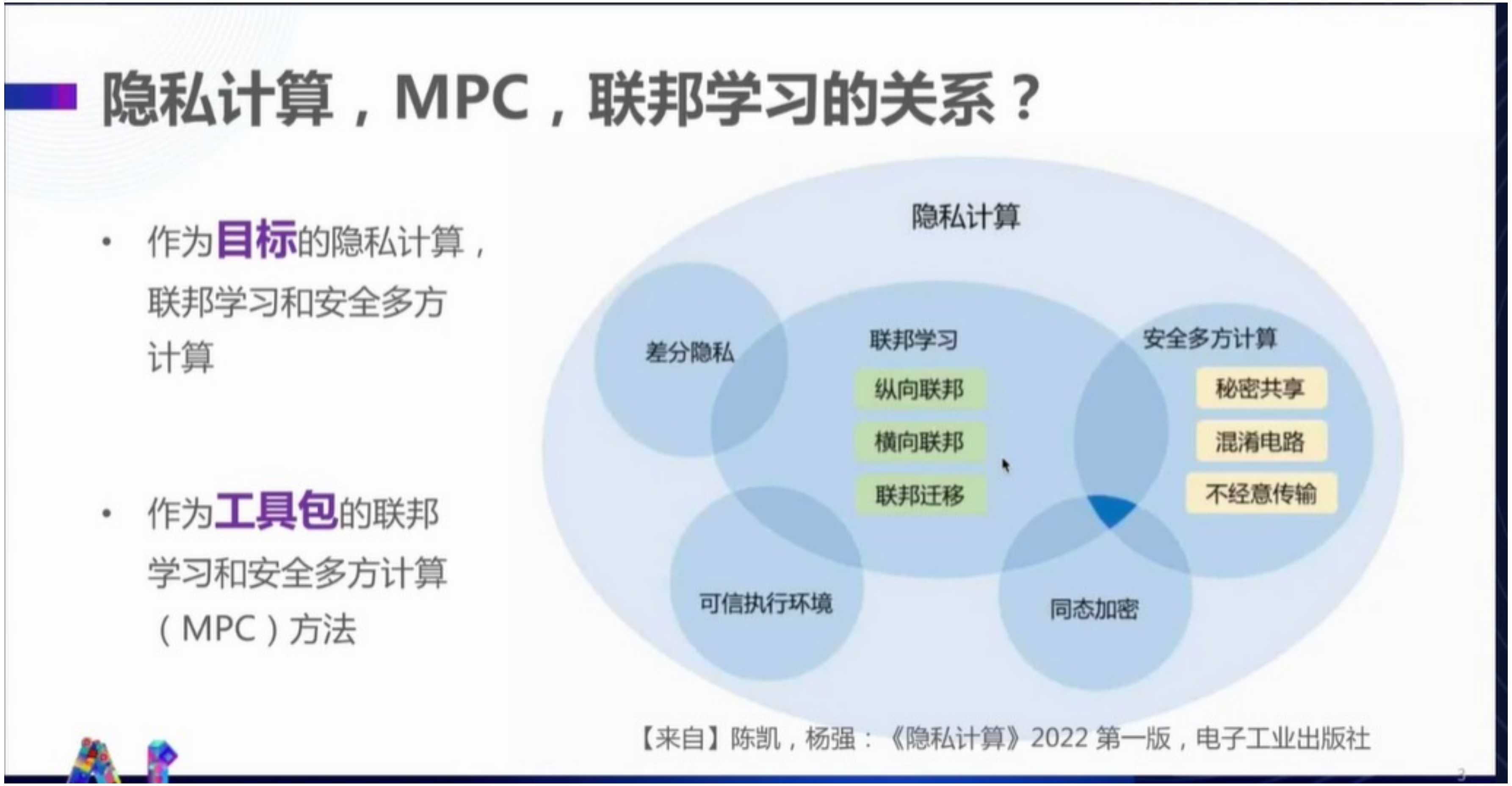
今天的大模型：
- BERT
- GPT-3
- 悟道
- ... 等

真实世界：
- 数据多源，散落各地
- 数据源属主不同，利益不同
- 数据格式，质量，特征不同
- 数据变化大
- 重要***：用户隐私，法律，监管，审计

未来：分布式小模型
- 分布式的 Distributed
- 自发的 Autonomous
- 联邦合作的 Federated

来源：【2022年机器之心 AI 科技年会】杨强教授：可信联邦学习 https://www.bilibili.com/video/BV1334y147HP/

联邦学习（FL, Federated Learning）



隐私计算，MPC，联邦学习的关系？

- 作为**目标**的隐私计算，联邦学习和安全多方计算
- 作为**工具包**的联邦学习和安全多方计算（MPC）方法

隐私计算

差分隐私

联邦学习

纵向联邦

横向联邦

联邦迁移

安全多方计算

秘密共享

混淆电路

不经意传输

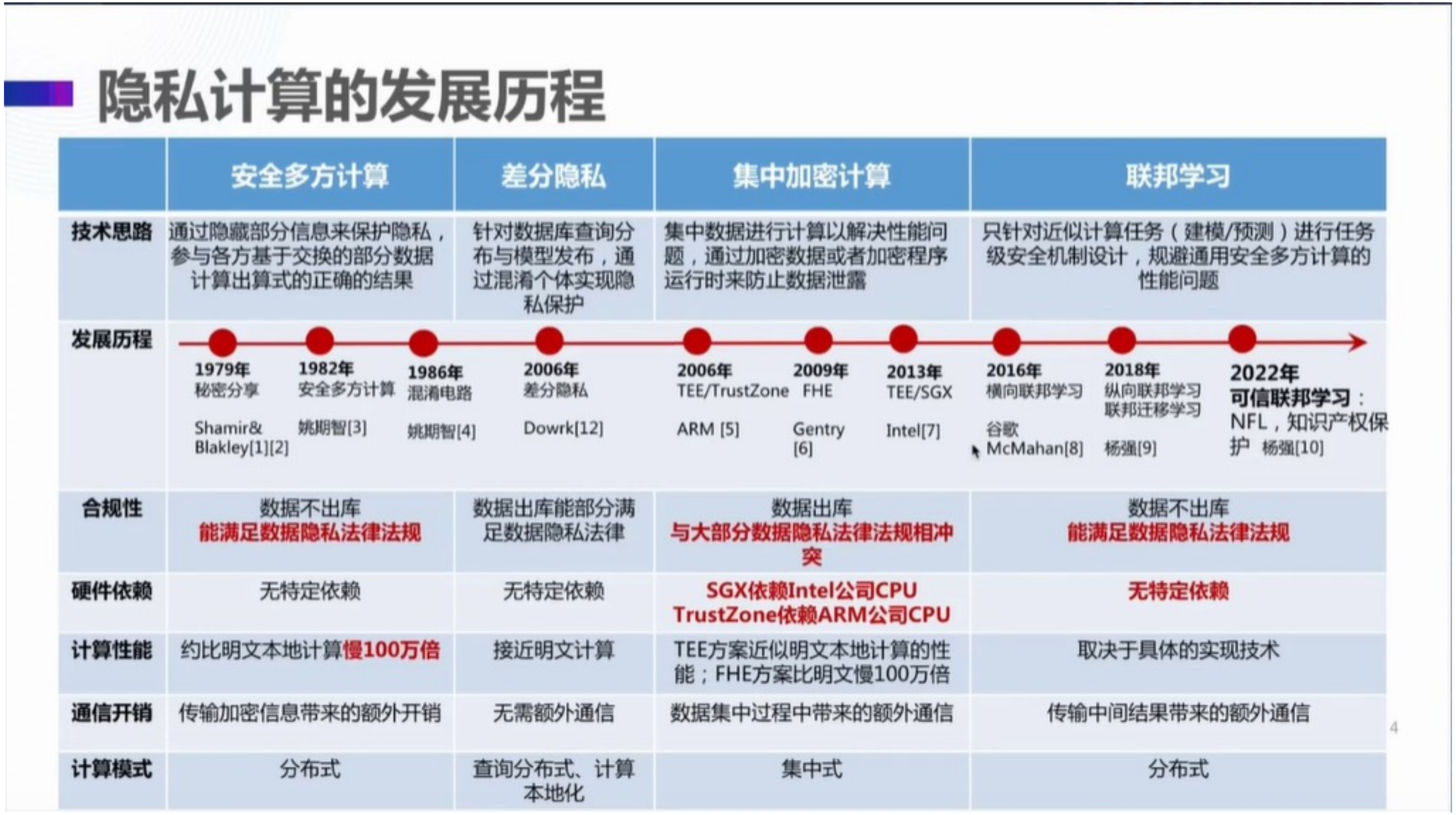
可信执行环境

同态加密

【来自】陈凯，杨强：《隐私计算》2022 第一版，电子工业出版社

来源：【2022年机器之心 AI 科技年会】杨强教授：可信联邦学习 https://www.bilibili.com/video/BV1334y147HP/

联邦学习（FL, Federated Learning）

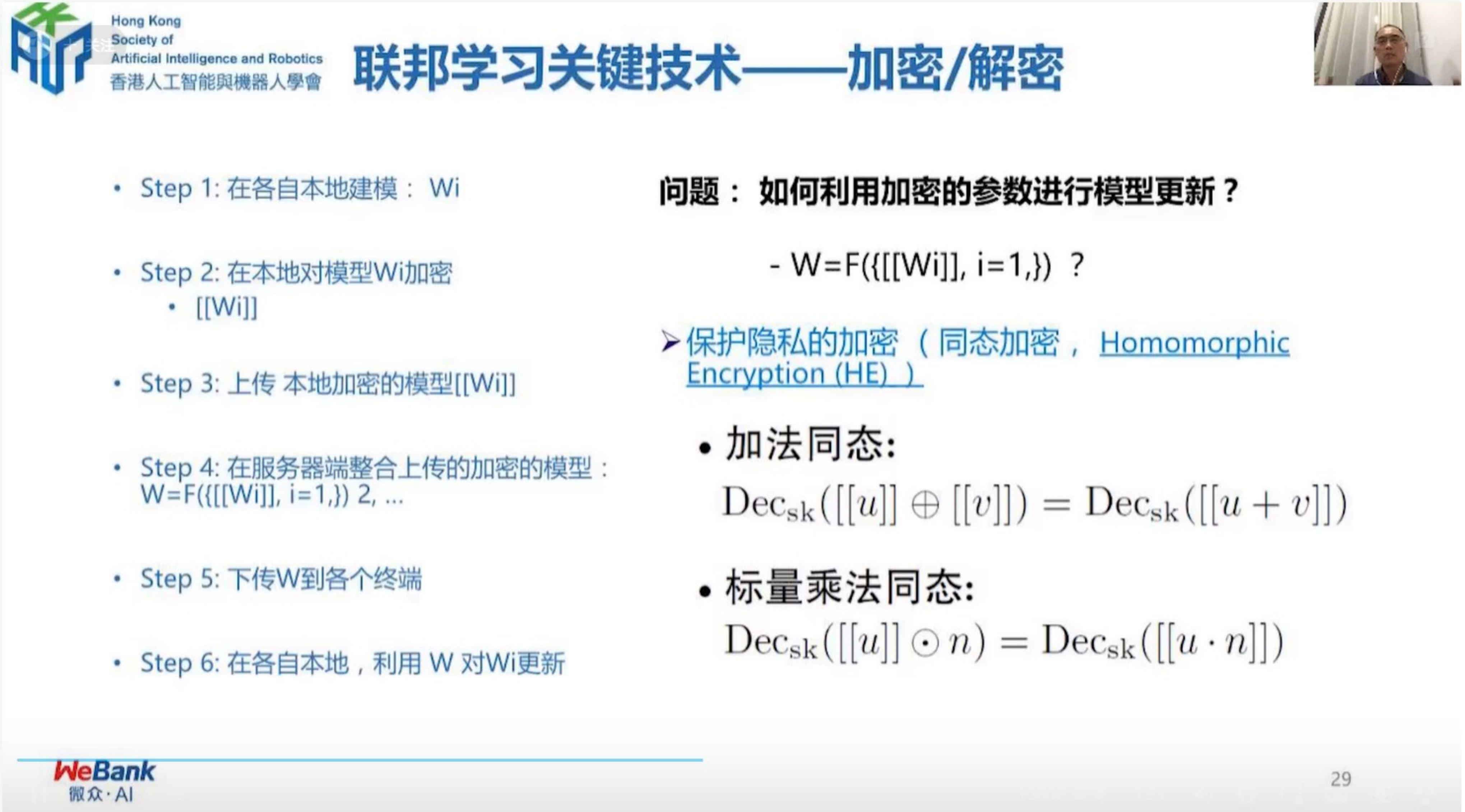


隐私计算的发展历程

	安全多方计算	差分隐私	集中加密计算	联邦学习
技术思路	通过隐藏部分信息来保护隐私，参与各方基于交换的部分数据计算出算式的正确的结果	针对数据库查询分布与模型发布，通过混淆个体实现隐私保护	集中数据进行计算以解决性能问题，通过加密数据或者加密程序运行时来防止数据泄露	只针对近似计算任务（建模/预测）进行任务级安全机制设计，规避通用安全多方计算的性能问题
发展历程	**1979年** 秘密分享 Shamir& Blakley[1][2] **1982年** 安全多方计算 姚期智[3] **1986年** 混淆电路 姚期智[4]	**2006年** 差分隐私 Dowrk[12]	**2006年** TEE/TrustZone ARM [5] **2009年** FHE Gentry [6] **2013年** TEE/SGX Intel[7]	**2016年** 横向联邦学习 谷歌 McMahan[8] **2018年** 纵向联邦学习 联邦迁移学习 杨强[9] **2022年** **可信联邦学习**：NFL，知识产权保护 杨强[10]
合规性	数据不出库 **能满足数据隐私法律法规**	数据出库能部分满足数据隐私法律	数据出库 **与大部分数据隐私法律法规相冲突**	数据不出库 **能满足数据隐私法律法规**
硬件依赖	无特定依赖	无特定依赖	**SGX依赖Intel公司CPU** **TrustZone依赖ARM公司CPU**	**无特定依赖**
计算性能	约比明文本地计算**慢100万倍**	接近明文计算	TEE方案近似明文本地计算的性能；FHE方案比明文慢100万倍	取决于具体的实现技术
通信开销	传输加密信息带来的额外开销	无需额外通信	数据集中过程中带来的额外通信	传输中间结果带来的额外通信
计算模式	分布式	查询分布式、计算本地化	集中式	分布式

来源：【2022年机器之心 AI 科技年会】杨强教授：可信联邦学习 https://www.bilibili.com/video/BV1334y147HP/

联邦学习（FL, Federated Learning）



Hong Kong Society of Artificial Intelligence and Robotics
香港人工智能與機器人學會

联邦学习关键技术——加密/解密

- Step 1: 在各自本地建模： Wi
- Step 2: 在本地对模型Wi加密
  - [[Wi]]
- Step 3: 上传 本地加密的模型[[Wi]]
- Step 4: 在服务器端整合上传的加密的模型：W=F({[[Wi]], i=1,}) 2, ...
- Step 5: 下传W到各个终端
- Step 6: 在各自本地，利用 W 对Wi更新

问题： 如何利用加密的参数进行模型更新？

- W=F({[[Wi]], i=1,}) ？

➢保护隐私的加密 （同态加密， Homomorphic Encryption (HE) ）

- 加法同态:

$$\mathrm{Dec}_{\mathrm{sk}}([[u]] \oplus [[v]]) = \mathrm{Dec}_{\mathrm{sk}}([[u+v]])$$

- 标量乘法同态:

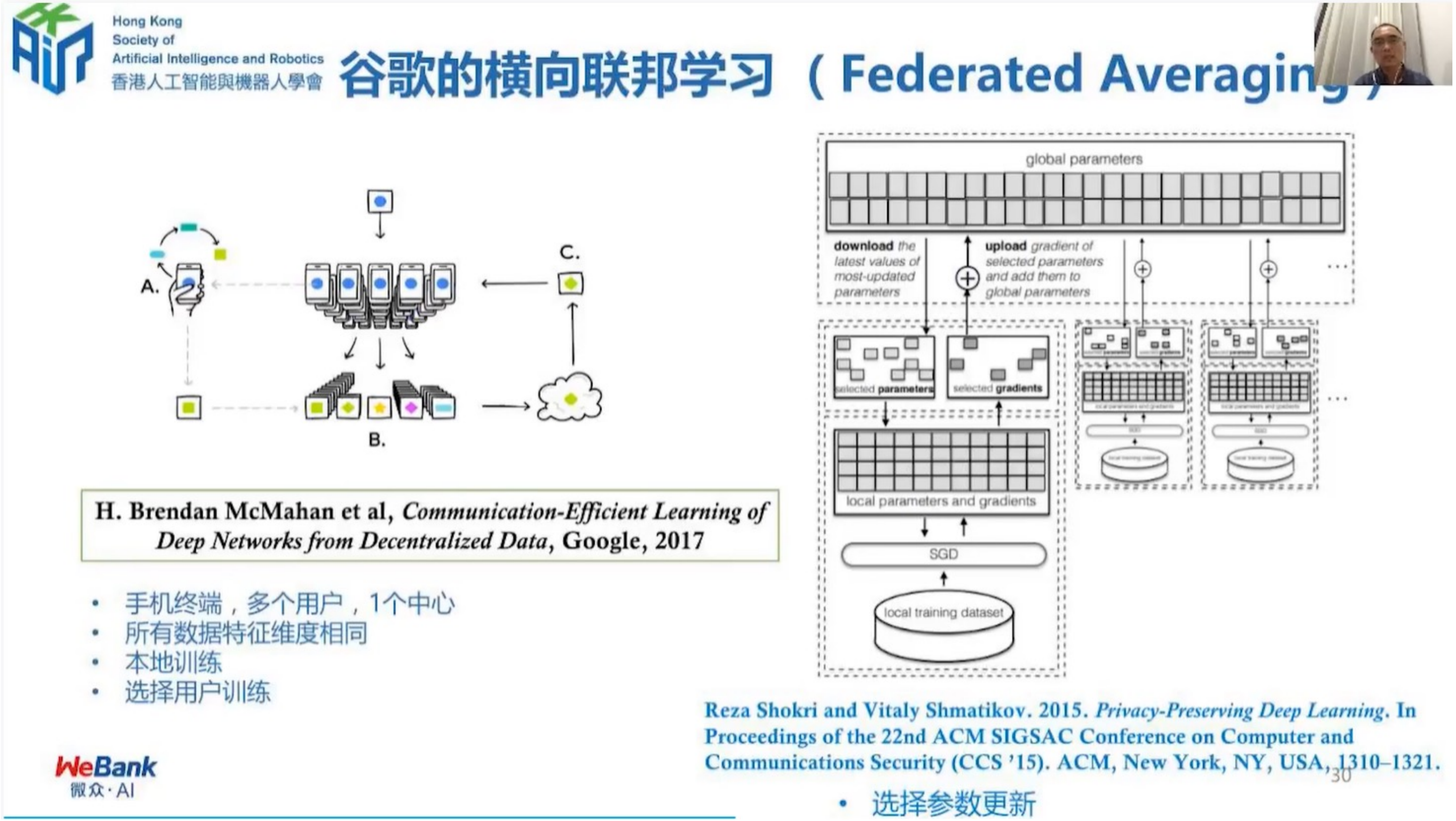
$$\mathrm{Dec}_{\mathrm{sk}}([[u]] \odot n) = \mathrm{Dec}_{\mathrm{sk}}([[u \cdot n]])$$

WeBank 微众·AI

29

来源：微众银行首席人工智能官杨强教授在HKSAIR《AI金融》课程：带你认识联邦学习与四大应用场景 https://www.bilibili.com/video/BV11T4y1E7Fi/

联邦学习（FL, Federated Learning)


Hong Kong
Society of
Artificial Intelligence and Robotics
香港人工智能與機器人學會
A.
B.
C.
global parameters
download the latest values of most-updated parameters
upload gradient of selected parameters and add them to global parameters
selected parameters
selected gradients
local parameters and gradients
SGD
local training dataset
WeBank
微众·AI
30

谷歌的横向联邦学习（Federated Averaging）

H. Brendan McMahan et al, *Communication-Efficient Learning of Deep Networks from Decentralized Data*, Google, 2017

- 手机终端，多个用户，1个中心
- 所有数据特征维度相同
- 本地训练
- 选择用户训练

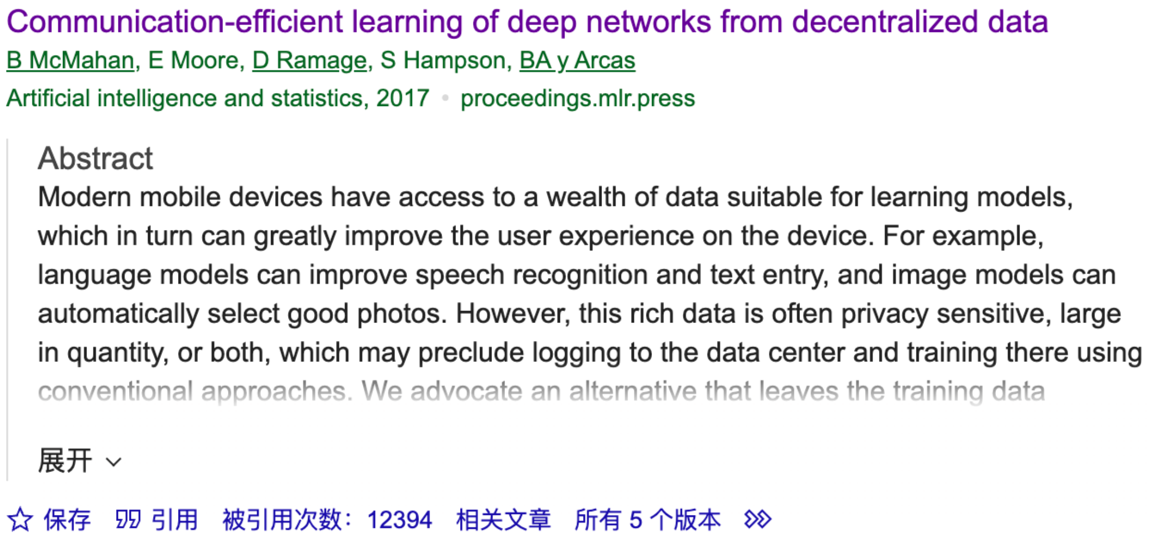
Reza Shokri and Vitaly Shmatikov. 2015. *Privacy-Preserving Deep Learning*. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15). ACM, New York, NY, USA, 1310–1321.

- 选择参数更新

来源：微众银行首席人工智能官杨强教授在HKSAIR《AI金融》课程：带你认识联邦学习与四大应用场景 https://www.bilibili.com/video/BV11T4y1E7Fi/

联邦学习（FL, Federated Learning）

FedAVG



Communication-efficient learning of deep networks from decentralized data

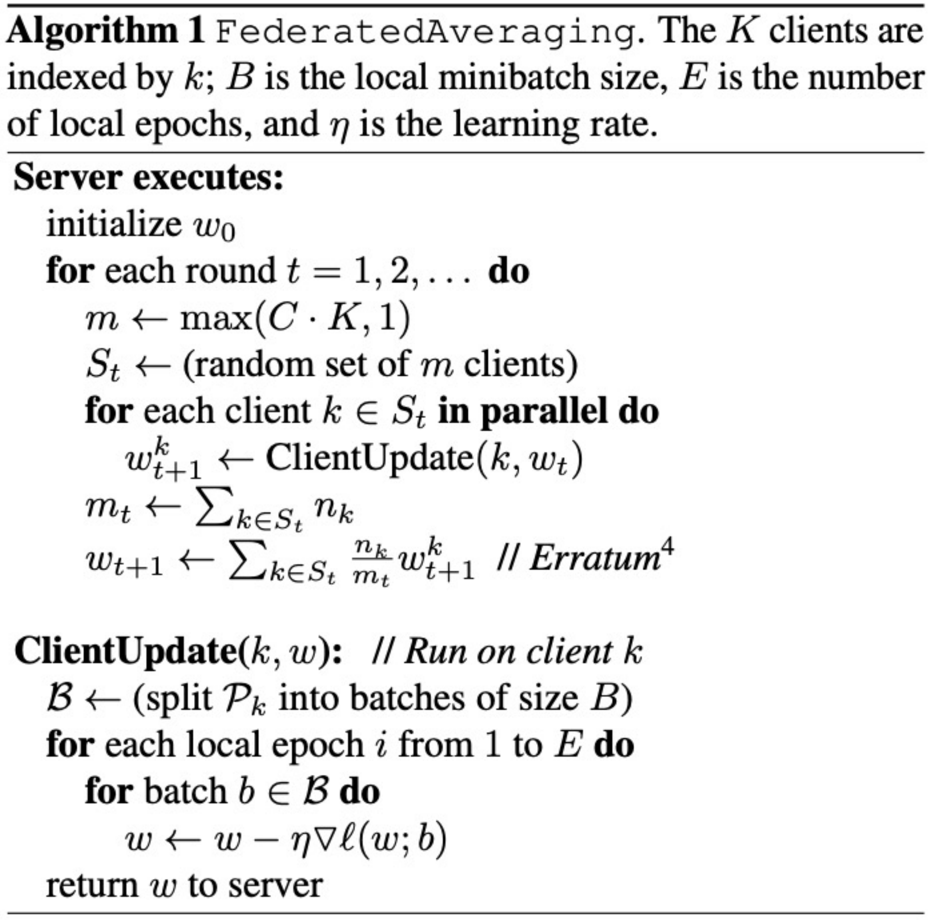
B McMahan, E Moore, D Ramage, S Hampson, BA y Arcas

Artificial intelligence and statistics, 2017 • proceedings.mlr.press

Abstract

Modern mobile devices have access to a wealth of data suitable for learning models, which in turn can greatly improve the user experience on the device. For example, language models can improve speech recognition and text entry, and image models can automatically select good photos. However, this rich data is often privacy sensitive, large in quantity, or both, which may preclude logging to the data center and training there using conventional approaches. We advocate an alternative that leaves the training data

展开 ⌄
☆ 保存
99 引用
被引用次数：12394
相关文章
所有 5 个版本
≫

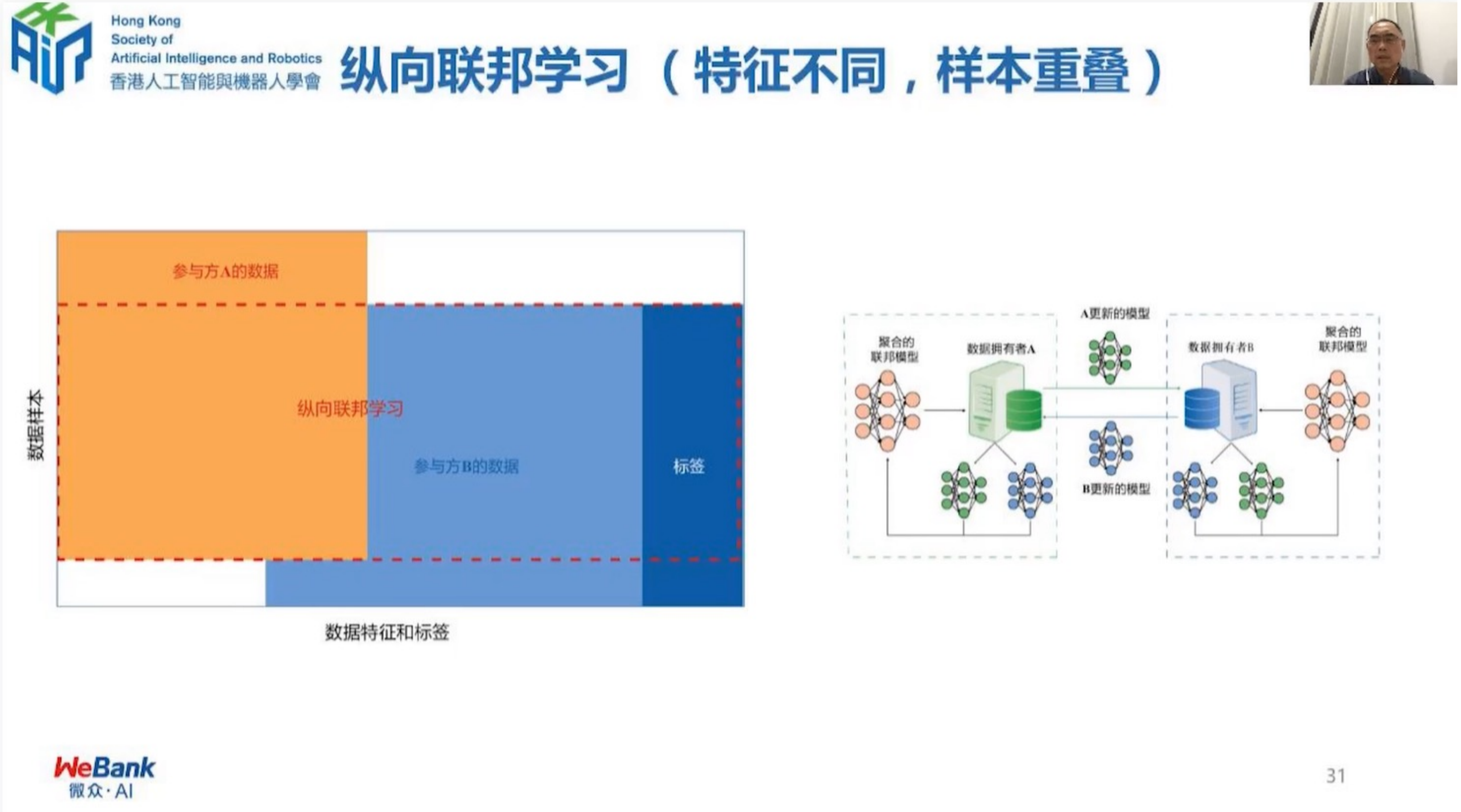


Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

```
Server executes:
  initialize $w_0$
  for each round $t = 1, 2, \dots$ do
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of $m$ clients)
    for each client $k \in S_t$ in parallel do
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $m_t \leftarrow \sum_{k \in S_t} n_k$
    $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$  // Erratum[4]

ClientUpdate($k, w$):  // Run on client $k$
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size $B$)
  for each local epoch $i$ from 1 to $E$ do
    for batch $b \in \mathcal{B}$ do
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return $w$ to server
```

联邦学习（FL, Federated Learning）

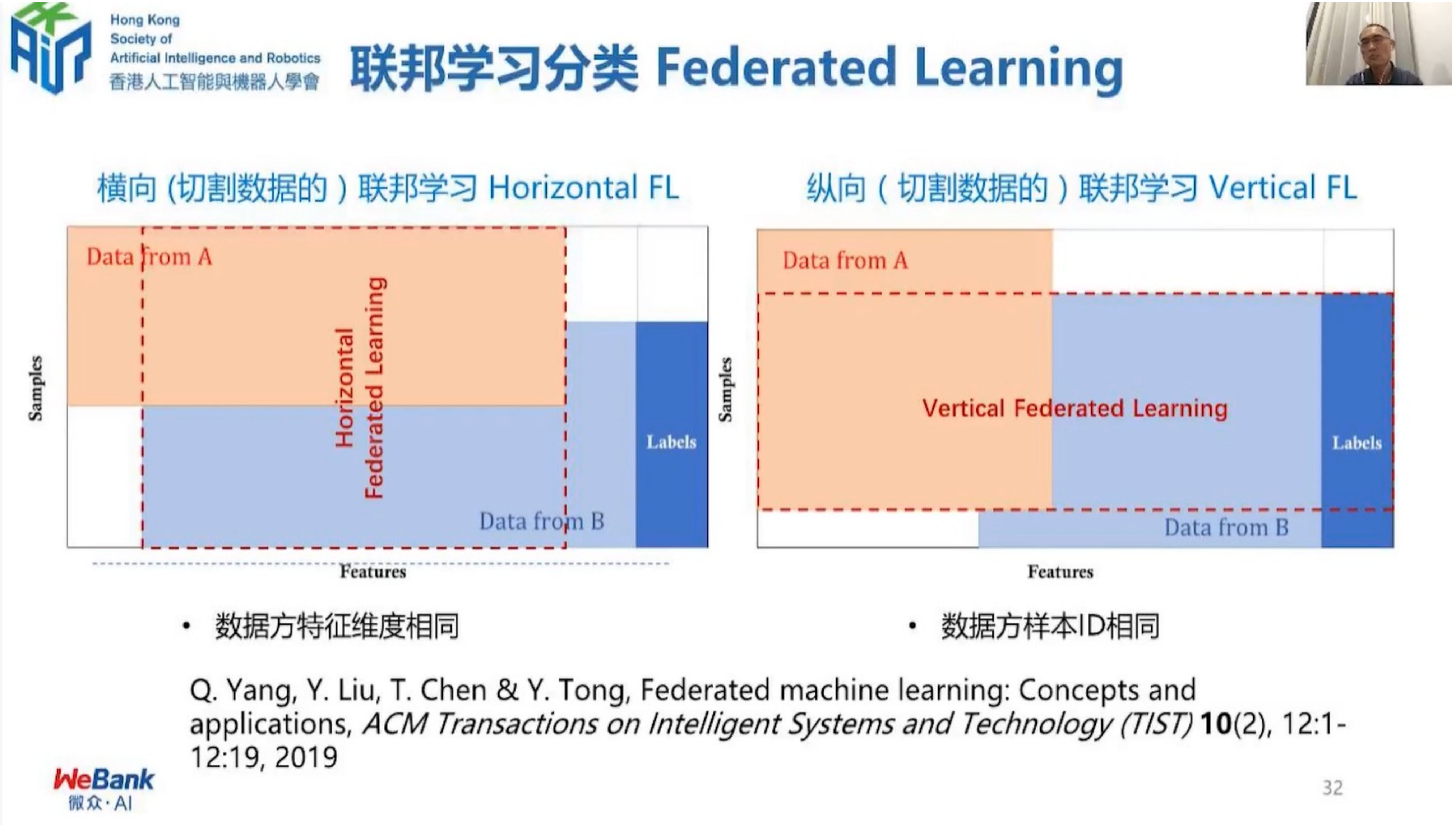


Hong Kong Society of Artificial Intelligence and Robotics 香港人工智能與機器人學會

纵向联邦学习（特征不同，样本重叠）

来源：微众银行首席人工智能官杨强教授在HKSAIR《AI金融》课程：带你认识联邦学习与四大应用场景 https://www.bilibili.com/video/BV11T4y1E7Fi/

联邦学习（FL, Federated Learning）



联邦学习分类 Federated Learning

横向 (切割数据的) 联邦学习 Horizontal FL

纵向 (切割数据的) 联邦学习 Vertical FL

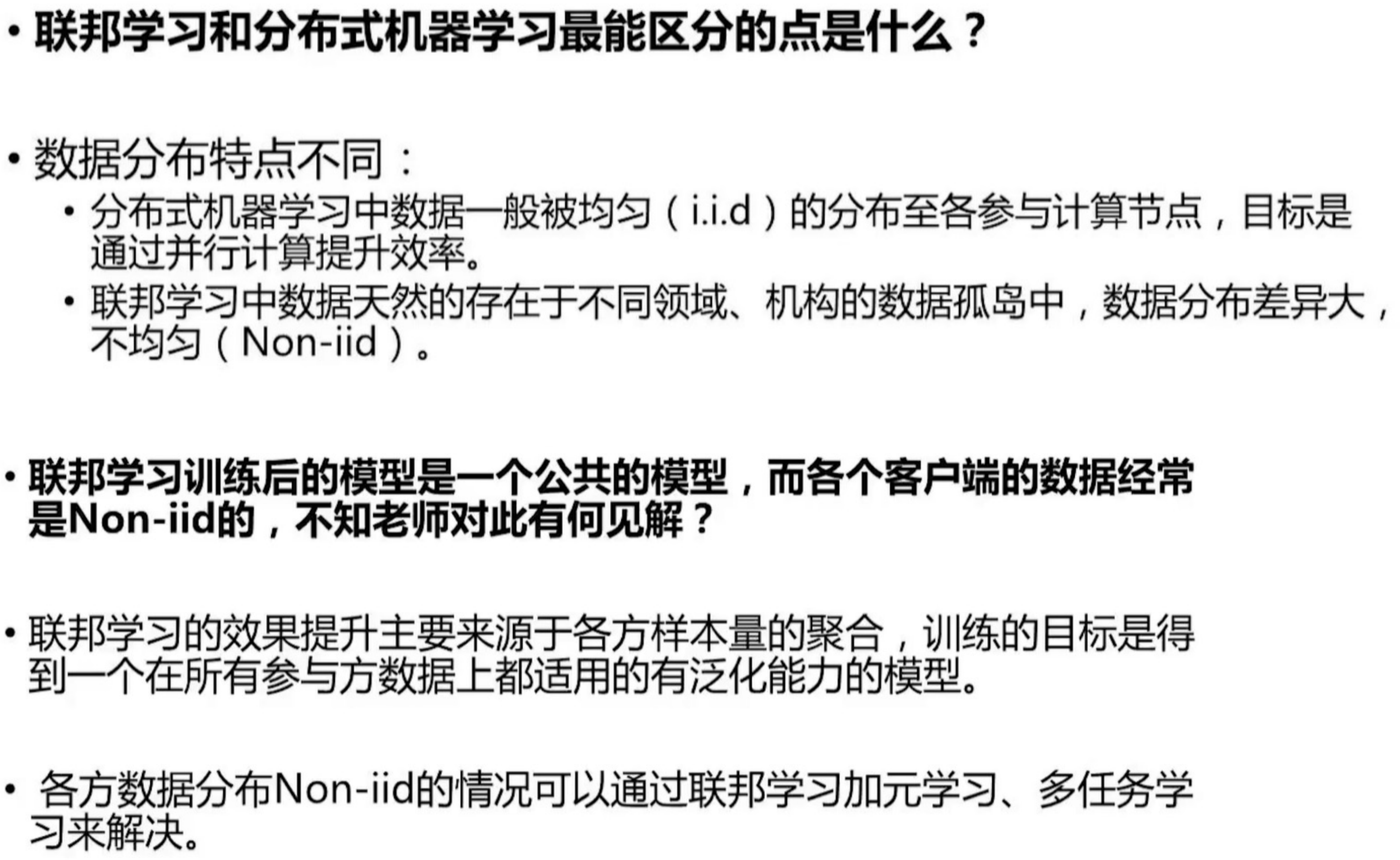
- 数据方特征维度相同

- 数据方样本ID相同

Q. Yang, Y. Liu, T. Chen & Y. Tong, Federated machine learning: Concepts and applications, *ACM Transactions on Intelligent Systems and Technology (TIST)* **10**(2), 12:1-12:19, 2019

来源：微众银行首席人工智能官杨强教授在HKSAIR《AI金融》课程：带你认识联邦学习与四大应用场景 https://www.bilibili.com/video/BV11T4y1E7Fi/

联邦学习（FL, Federated Learning）



- **联邦学习和分布式机器学习最能区分的点是什么？**

- 数据分布特点不同：
  - 分布式机器学习中数据一般被均匀（i.i.d）的分布至各参与计算节点，目标是通过并行计算提升效率。
  - 联邦学习中数据天然的存在于不同领域、机构的数据孤岛中，数据分布差异大，不均匀（Non-iid）。

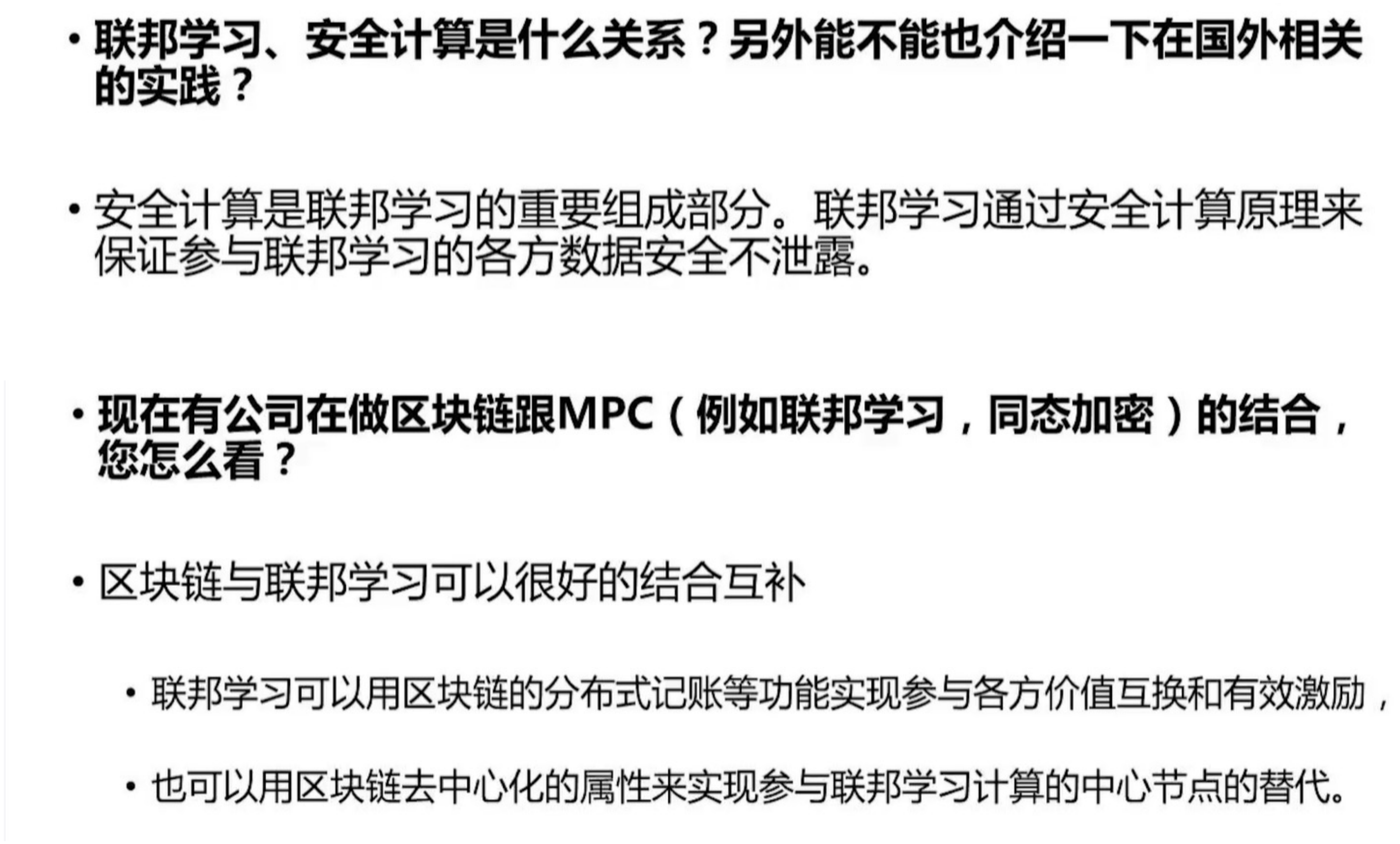
- **联邦学习训练后的模型是一个公共的模型，而各个客户端的数据经常是Non-iid的，不知老师对此有何见解？**

- 联邦学习的效果提升主要来源于各方样本量的聚合，训练的目标是得到一个在所有参与方数据上都适用的有泛化能力的模型。

- 各方数据分布Non-iid的情况可以通过联邦学习加元学习、多任务学习来解决。

来源：微众银行首席人工智能官杨强教授在HKSAIR《AI金融》课程：带你认识联邦学习与四大应用场景 https://www.bilibili.com/video/BV11T4y1E7Fi/

联邦学习（FL, Federated Learning）



- **联邦学习、安全计算是什么关系？另外能不能也介绍一下在国外相关的实践？**

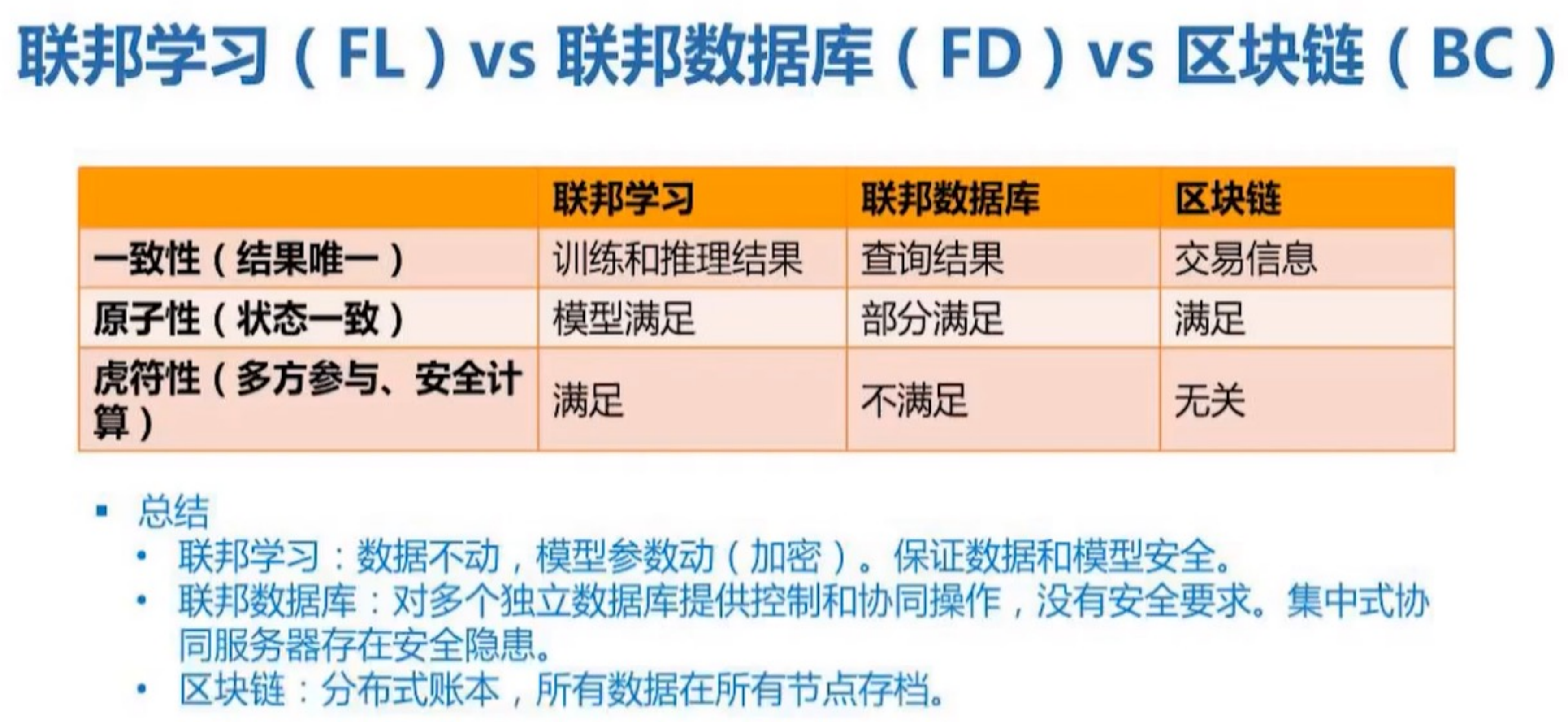
- 安全计算是联邦学习的重要组成部分。联邦学习通过安全计算原理来保证参与联邦学习的各方数据安全不泄露。

- **现在有公司在做区块链跟MPC（例如联邦学习，同态加密）的结合，您怎么看？**

- 区块链与联邦学习可以很好的结合互补
  - 联邦学习可以用区块链的分布式记账等功能实现参与各方价值互换和有效激励，
  - 也可以用区块链去中心化的属性来实现参与联邦学习计算的中心节点的替代。

来源：微众银行首席人工智能官杨强教授在HKSAIR《AI金融》课程：带你认识联邦学习与四大应用场景 https://www.bilibili.com/video/BV11T4y1E7Fi/

联邦学习（FL, Federated Learning）



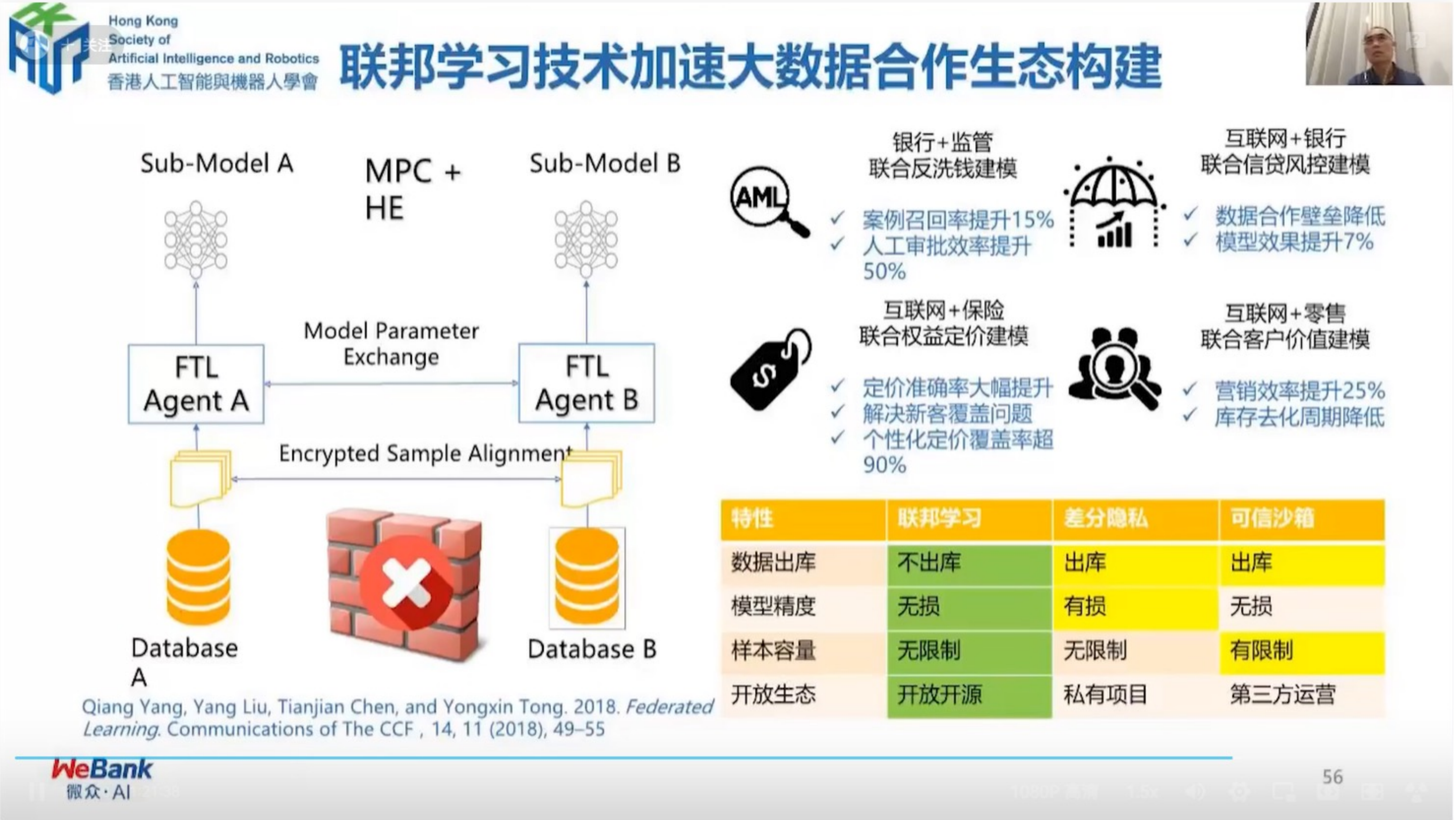
联邦学习（FL）vs 联邦数据库（FD）vs 区块链（BC）

	联邦学习	联邦数据库	区块链
一致性（结果唯一）	训练和推理结果	查询结果	交易信息
原子性（状态一致）	模型满足	部分满足	满足
虎符性（多方参与、安全计算）	满足	不满足	无关

- 总结
  - 联邦学习：数据不动，模型参数动（加密）。保证数据和模型安全。
  - 联邦数据库：对多个独立数据库提供控制和协同操作，没有安全要求。集中式协同服务器存在安全隐患。
  - 区块链：分布式账本，所有数据在所有节点存档。

来源：微众银行首席人工智能官杨强教授在HKSAIR《AI金融》课程：带你认识联邦学习与四大应用场景 https://www.bilibili.com/video/BV11T4y1E7Fi/

联邦学习（FL, Federated Learning）



Hong Kong Society of Artificial Intelligence and Robotics
香港人工智能與機器人學會

联邦学习技术加速大数据合作生态构建

Sub-Model A

MPC + HE

Sub-Model B

Model Parameter Exchange

FTL Agent A

FTL Agent B

Encrypted Sample Alignment

Database A

Database B

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2018. *Federated Learning*. Communications of The CCF , 14, 11 (2018), 49–55

AML

银行+监管
联合反洗钱建模

- ✓ 案例召回率提升15%
- ✓ 人工审批效率提升50%

互联网+银行
联合信贷风控建模

- ✓ 数据合作壁垒降低
- ✓ 模型效果提升7%

互联网+保险
联合权益定价建模

- ✓ 定价准确率大幅提升
- ✓ 解决新客覆盖问题
- ✓ 个性化定价覆盖率超90%

互联网+零售
联合客户价值建模

- ✓ 营销效率提升25%
- ✓ 库存去化周期降低

特性	联邦学习	差分隐私	可信沙箱
数据出库	不出库	出库	出库
模型精度	无损	有损	无损
样本容量	无限制	无限制	有限制
开放生态	开放开源	私有项目	第三方运营

WeBank
微众·AI

56

来源：微众银行首席人工智能官杨强教授在HKSAIR《AI金融》课程：带你认识联邦学习与四大应用场景 https://www.bilibili.com/video/BV11T4y1E7Fi/

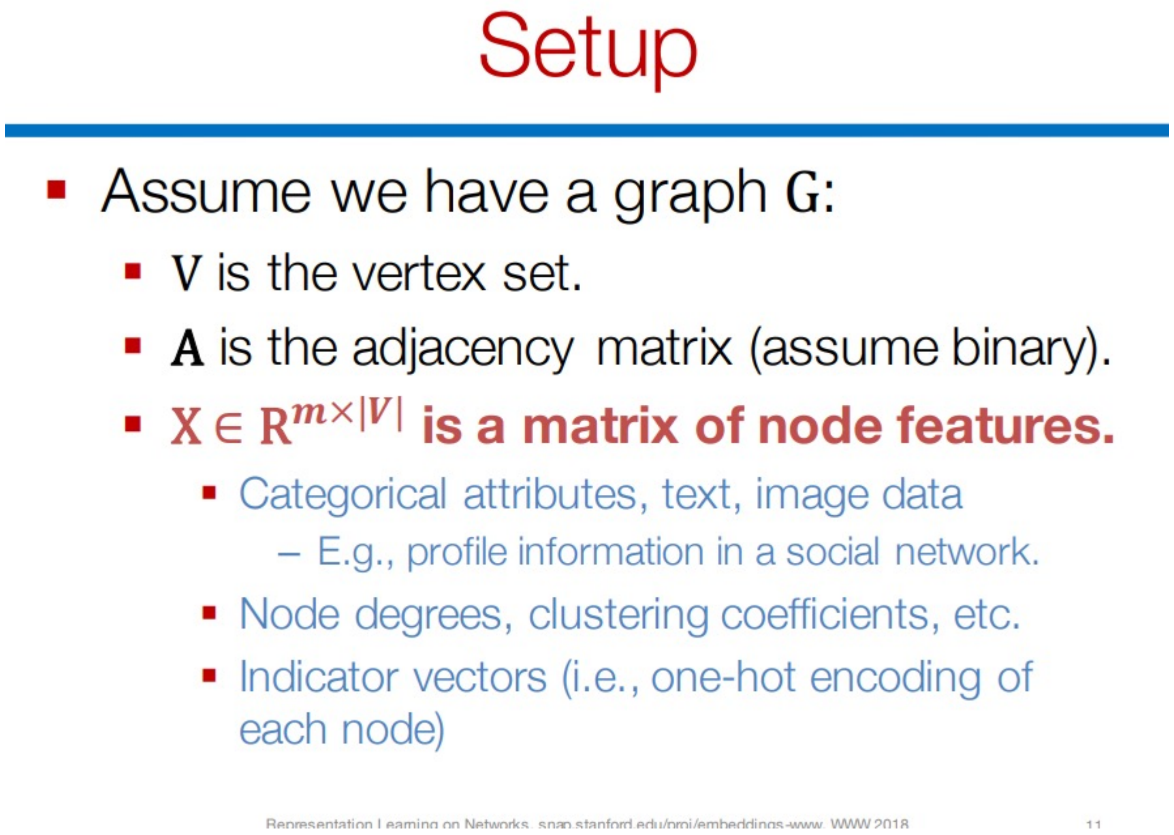
目录

- 背景介绍

- 前置知识
  - 联邦学习（FL, Federated Learning）
  - 图神经网络（GNN, Graph Neural Network）

- 论文分享（FL+GNN)
  - (2021 NIPS) Subgraph Federated Learning with Missing Neighbor Generation
  - (2022 Nature Communications) A federated graph neural network framework for privacy-preserving personalization
  - (2023 ICML) Personalized Subgraph Federated Learning

图神经网络（GNN, Graph Neural Network）



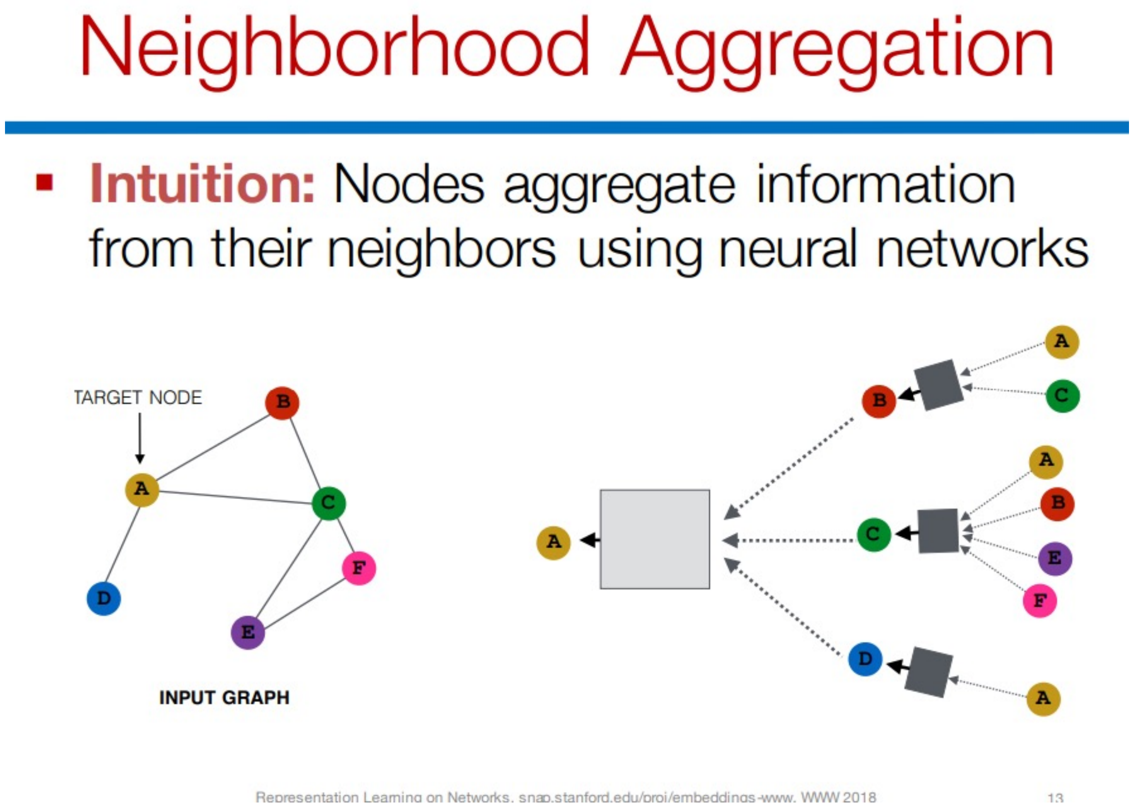
Setup

- Assume we have a graph $\mathbf{G}$:
  - $\mathbf{V}$ is the vertex set.
  - $\mathbf{A}$ is the adjacency matrix (assume binary).
  - $\mathbf{X} \in \mathbf{R}^{m \times |V|}$ **is a matrix of node features.**
    - Categorical attributes, text, image data
      - E.g., profile information in a social network.
    - Node degrees, clustering coefficients, etc.
    - Indicator vectors (i.e., one-hot encoding of each node)

Representation Learning on Networks, snap.stanford.edu/proj/embeddings-www, WWW 2018 11

Source: WWW-18 Tutorial Representation Learning on Networks https://snap.stanford.edu/proj/embeddings-www/

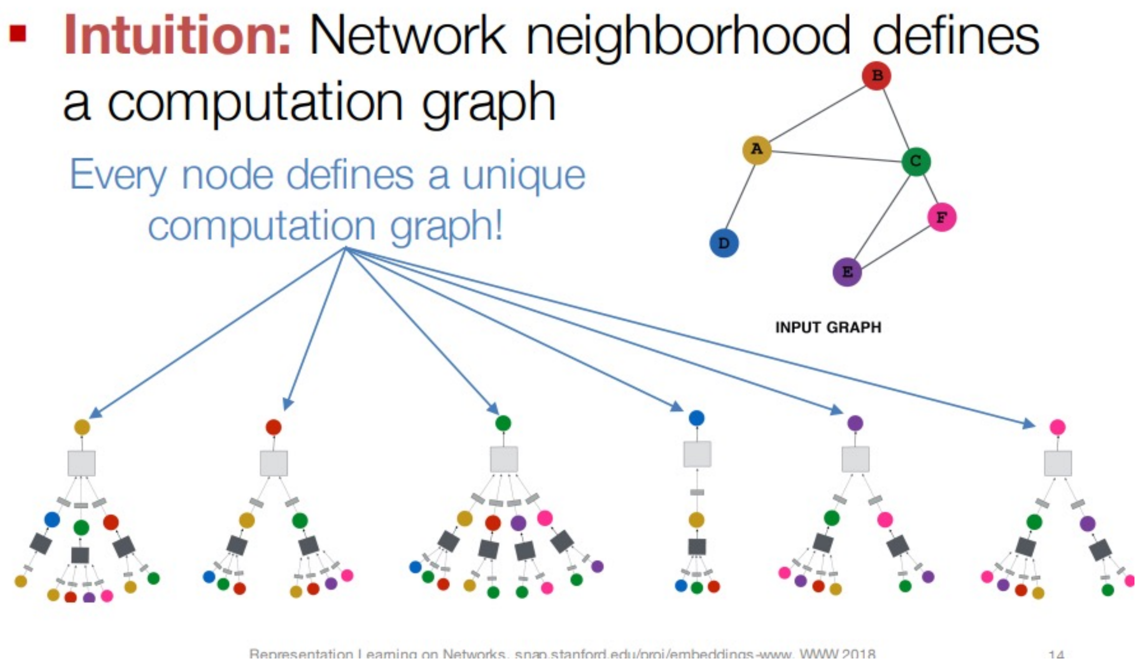
图神经网络（GNN, Graph Neural Network）



Neighborhood Aggregation

- **Intuition:** Nodes aggregate information from their neighbors using neural networks

Representation Learning on Networks, snap.stanford.edu/proj/embeddings-www, WWW 2018 13



- **Intuition:** Network neighborhood defines a computation graph

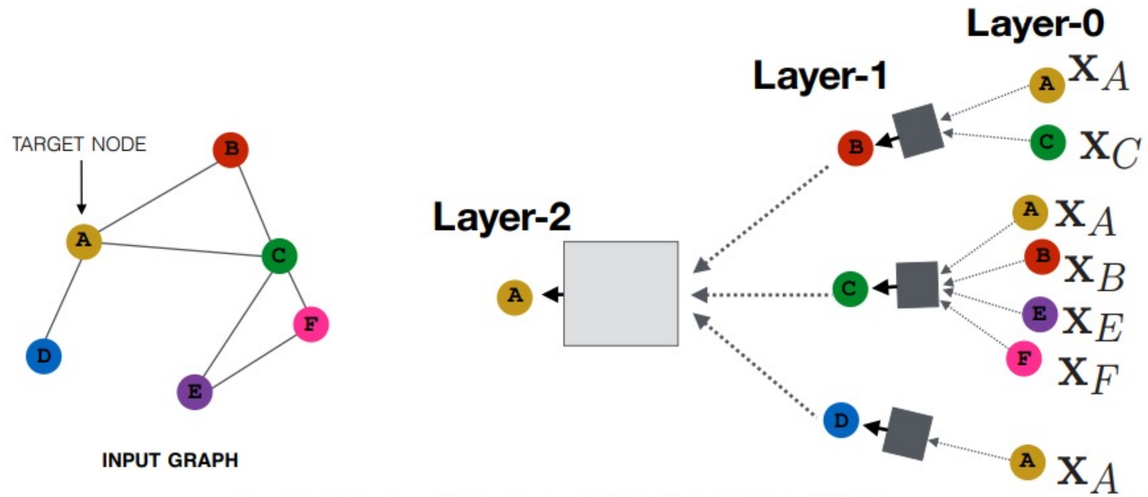
Every node defines a unique computation graph!

Representation Learning on Networks, snap.stanford.edu/proj/embeddings-www, WWW 2018 14

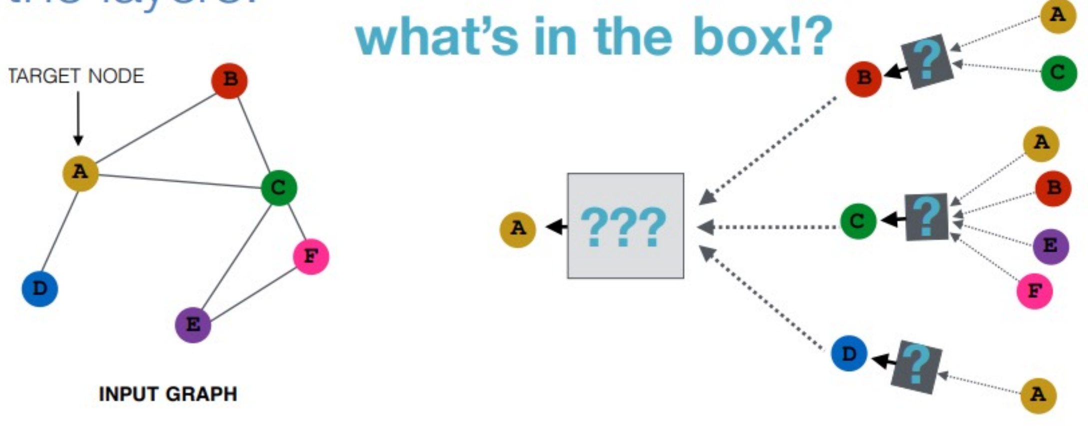
Source: WWW-18 Tutorial Representation Learning on Networks https://snap.stanford.edu/proj/embeddings-www/

图神经网络（GNN, Graph Neural Network）

- Nodes have embeddings at each layer.
- Model can be arbitrary depth.
- “layer-0” embedding of node $\mathbf{u}$ is its input feature, i.e. $\mathbf{x}_{\mathbf{u}}$.



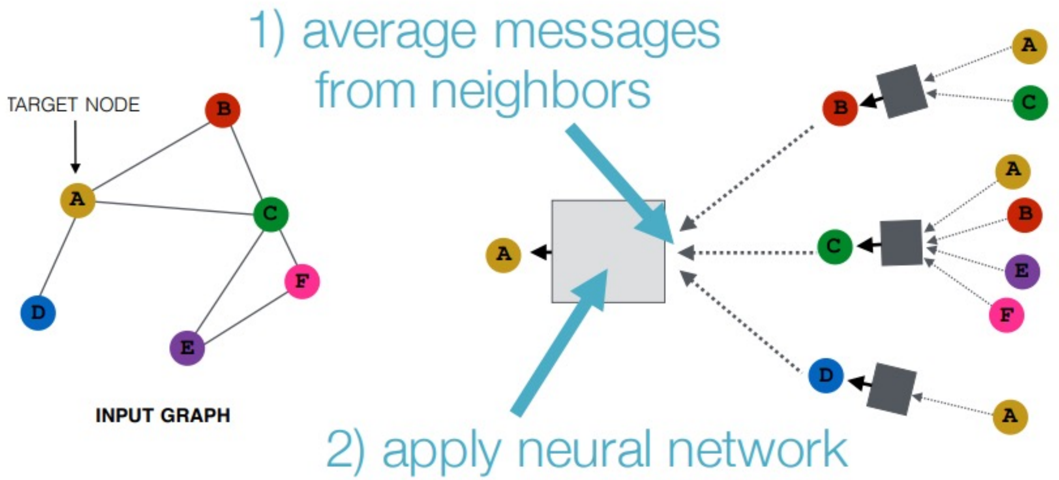
- Key distinctions are in how different approaches aggregate information across the layers.



Source: WWW-18 Tutorial Representation Learning on Networks https://snap.stanford.edu/proj/embeddings-www/

图神经网络（GNN, Graph Neural Network）

- **Basic approach:** Average neighbor information and apply a neural network.



- **Basic approach:** Average neighbor messages and apply a neural network.

Initial "layer 0" embeddings are equal to node features

$$\boxed{\mathbf{h}_v^0 = \mathbf{x}_v}$$

previous layer embedding of *v*

$$\boxed{\mathbf{h}_v^k} = \boxed{\sigma}\left(\mathbf{W}_k \boxed{\sum_{u\in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|}} + \mathbf{B}_k \boxed{\mathbf{h}_v^{k-1}}\right), \quad \forall k > 0$$

kth layer embedding of *v*

non-linearity (e.g., ReLU or tanh)

average of neighbor's previous layer embeddings

- GraphSAGE:

concatenate self embedding and neighbor embedding

$$\mathbf{h}_v^k = \sigma\left(\left[\mathbf{W}_k \cdot \boxed{\text{AGG}\left(\{\mathbf{h}_u^{k-1}, \forall u \in N(v)\}\right)}, \mathbf{B}_k\mathbf{h}_v^{k-1}\right]\right)$$

generalized aggregation

Source: WWW-18 Tutorial Representation Learning on Networks https://snap.stanford.edu/proj/embeddings-www/

目录

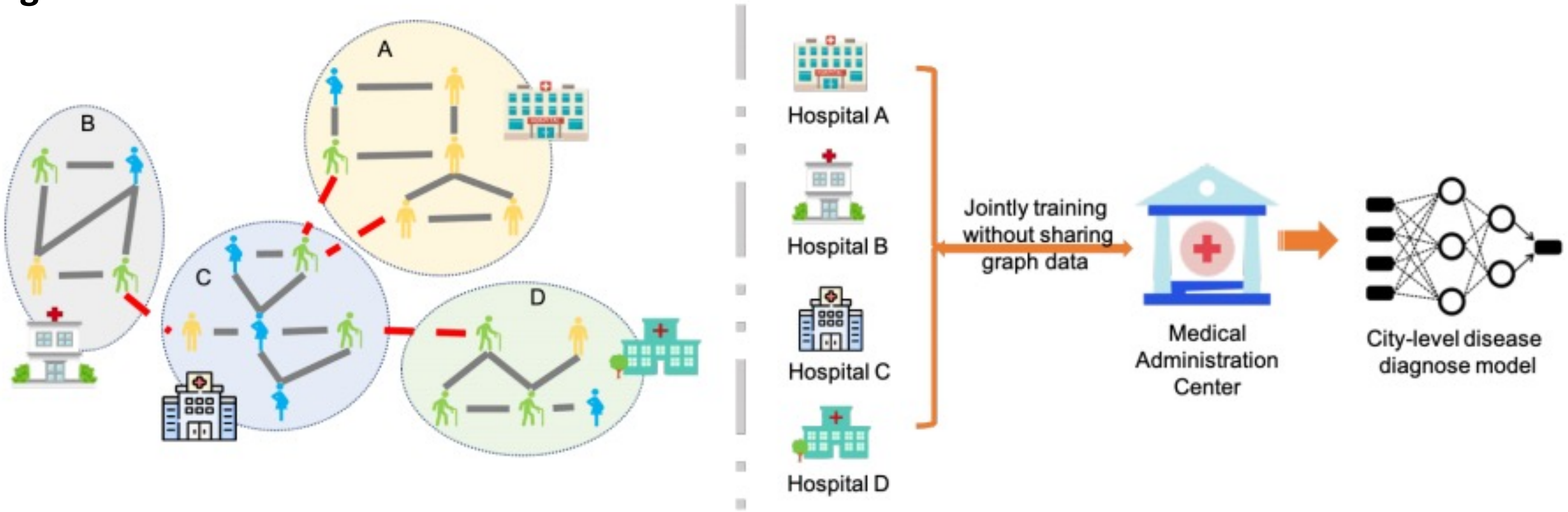
- 背景介绍

- 前置知识
  - 联邦学习（FL, Federated Learning）
  - 图神经网络（GNN, Graph Neural Network）

- 论文分享（FL+GNN)
  - (2021 NIPS) Subgraph Federated Learning with Missing Neighbor Generation
  - (2022 Nature Communications) A federated graph neural network framework for privacy-preserving personalization
  - (2023 ICML) Personalized Subgraph Federated Learning

Subgraph Federated Learning with Missing Neighbor Generation (2021 NIPS)

Motivating Scenario


A
B
C
D
Hospital A
Hospital B
Hospital C
Hospital D
Jointly training without sharing graph data
Medical Administration Center
City-level disease diagnose model

- Scenario: residents of a city may go to different hospitals for various reasons ⟶ their healthcare data are stored only within the hospitals they visit
- When any healthcare problem is to be studied in the whole city, a single powerful graph mining model is needed to conduct effective inference over the entire global patient network.
- However, it is rather difficult to let all hospitals share their patient networks with others to train the graph mining model due to conflicts of interests.

Subgraph Federated Learning with Missing Neighbor Generation (2021 NIPS)

- **Goal:** train a ***powerful and generalizable*** graph mining model over ***multiple distributed subgraphs*** without actual data sharing

- **Challenge 1: How to jointly learn from multiple local subgraphs?**
  - How to capture the global data distribution and avoid prone to overfitting
  - How to integrate multiple graph mining models into a universally applicable one
- Solution 1: FedSage: Training GraphSage with FedAvg.

- **Challenge 2: How to deal with missing links across local subgraphs?**
  - data samples in graphs are connected and correlated
  - data samples in each subgraph can potentially have connections to those in other subgraphs.
- Solution 2: FedSage+: Generating missing neighbors along FedSage.
  - add a missing neighbor generator
    - generate potential missing links within the subgraph ⟶ generate missing neighbors across distributed subgraphs

Subgraph Federated Learning with Missing Neighbor Generation (2021 NIPS)

FedSage+

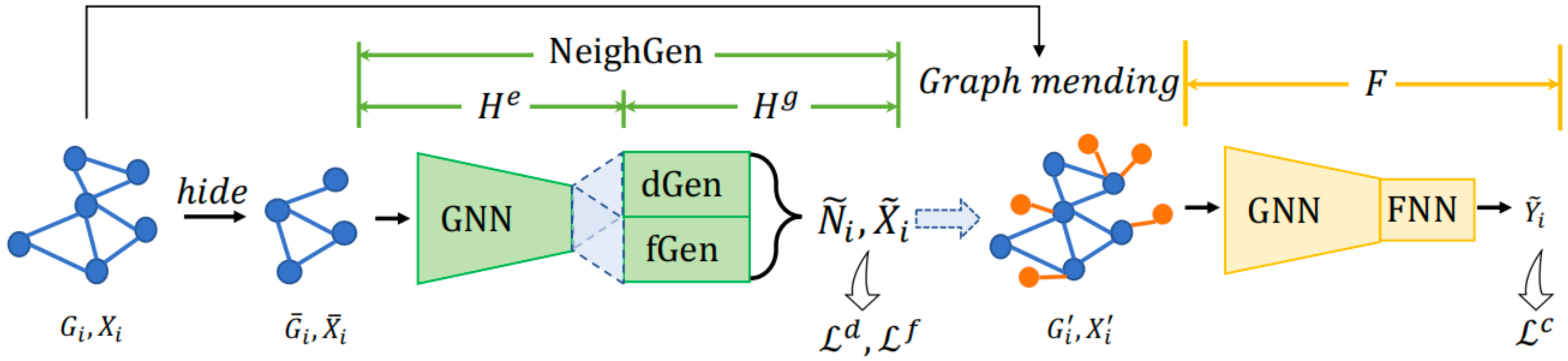


Figure 2: Joint training of missing neighbor generation and node classification.

- Missing Neighbor Generator (NeighGen)
  - Neural architecture of NeighGen.
  - Graph mending simulation.
  - Neighbor Generation.

$$\widetilde{n}_v = \sigma((\theta^d)^T \cdot n_v), \text{ and } \widetilde{x}_v = R\left(\sigma\left((\theta^f)^T \cdot (z_v + \mathbf{N}(0,1))\right), \widetilde{n}_v\right).$$

$$\mathcal{L}^n = \lambda^d \mathcal{L}^d + \lambda^f \mathcal{L}^f = \lambda^d \frac{1}{|\bar{V}_i|} \sum_{v \in \bar{V}_i} L_1^S(\widetilde{n}_v - n_v) + \lambda^f \frac{1}{|\bar{V}_i|} \sum_{v \in \bar{V}_i} \sum_{p \in [\widetilde{n}_v]} \min_{u \in \mathcal{N}_{G_i}(v) \cap V_i^h} (||\widetilde{x}_v^p - x_u||_2^2),$$

Subgraph Federated Learning with Missing Neighbor Generation (2021 NIPS)

FedSage+

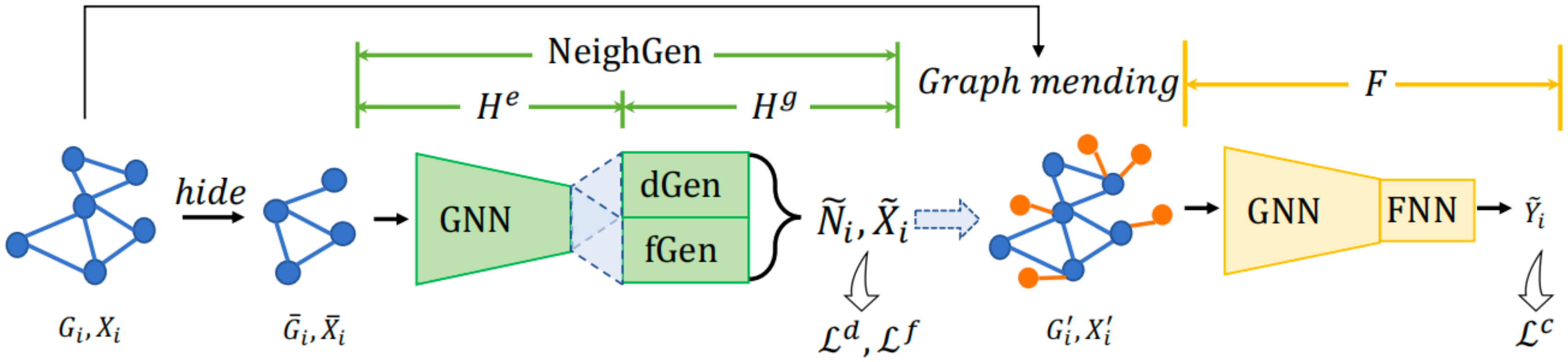


Figure 2: Joint training of missing neighbor generation and node classification.

- Federated Learning of GraphSage and NeighGen
  - Problem: cooperation through directly averaging weights of NeighGen across the system can negatively affect its performance
  - Goal: generating diverse missing neighbors in each subgraph
  - Solution: add a cross-subgraph feature reconstruction loss into fGen

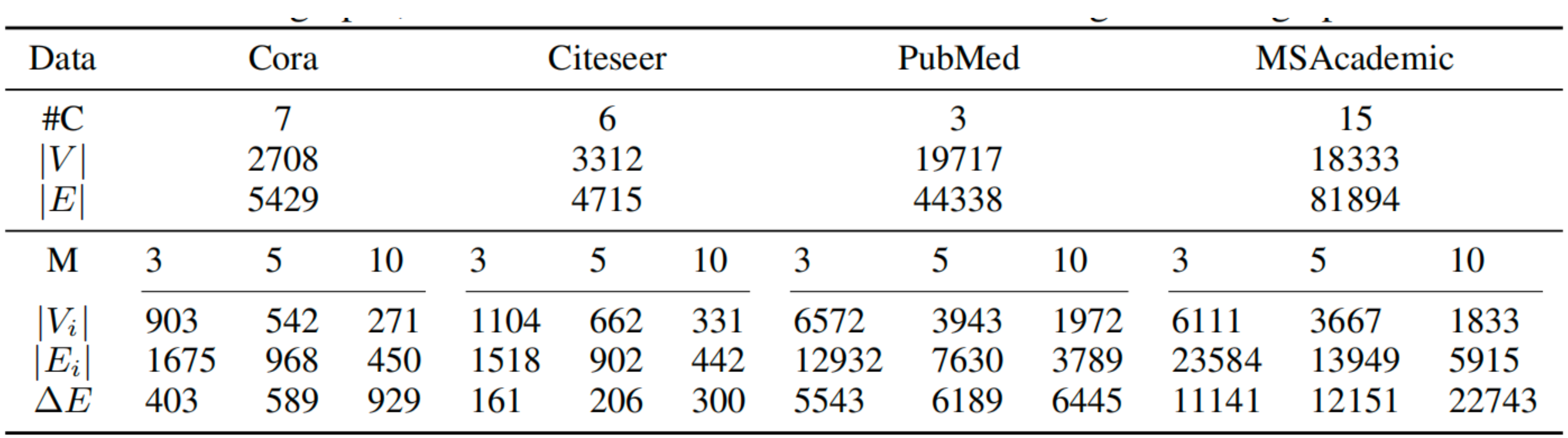
picked as the closest node from G_j other than G_i to simulate the neighbor of $v \in \bar{V}_i$ missed into G_j

$$\mathcal{L}_i^f = \frac{1}{|\bar{V}_i|}\sum_{v\in\bar{V}_i}\sum_{p\in[\tilde{n}_v]}\left(\min_{u\in\mathcal{N}_{G_i}(v)\cap V_i^h}(||\tilde{x}_v^p - x_u||_2^2) + \boxed{\alpha\sum_{j\in[M]/i}\min_{u\in V_j}(||H_i^g(z_v)^p - x_u||_2^2)}\right)$$

Subgraph Federated Learning with Missing Neighbor Generation (2021 NIPS)

Experiments

graph impairing ratio *h*, we set *h*% ∈ [3.4%, 27.8%]

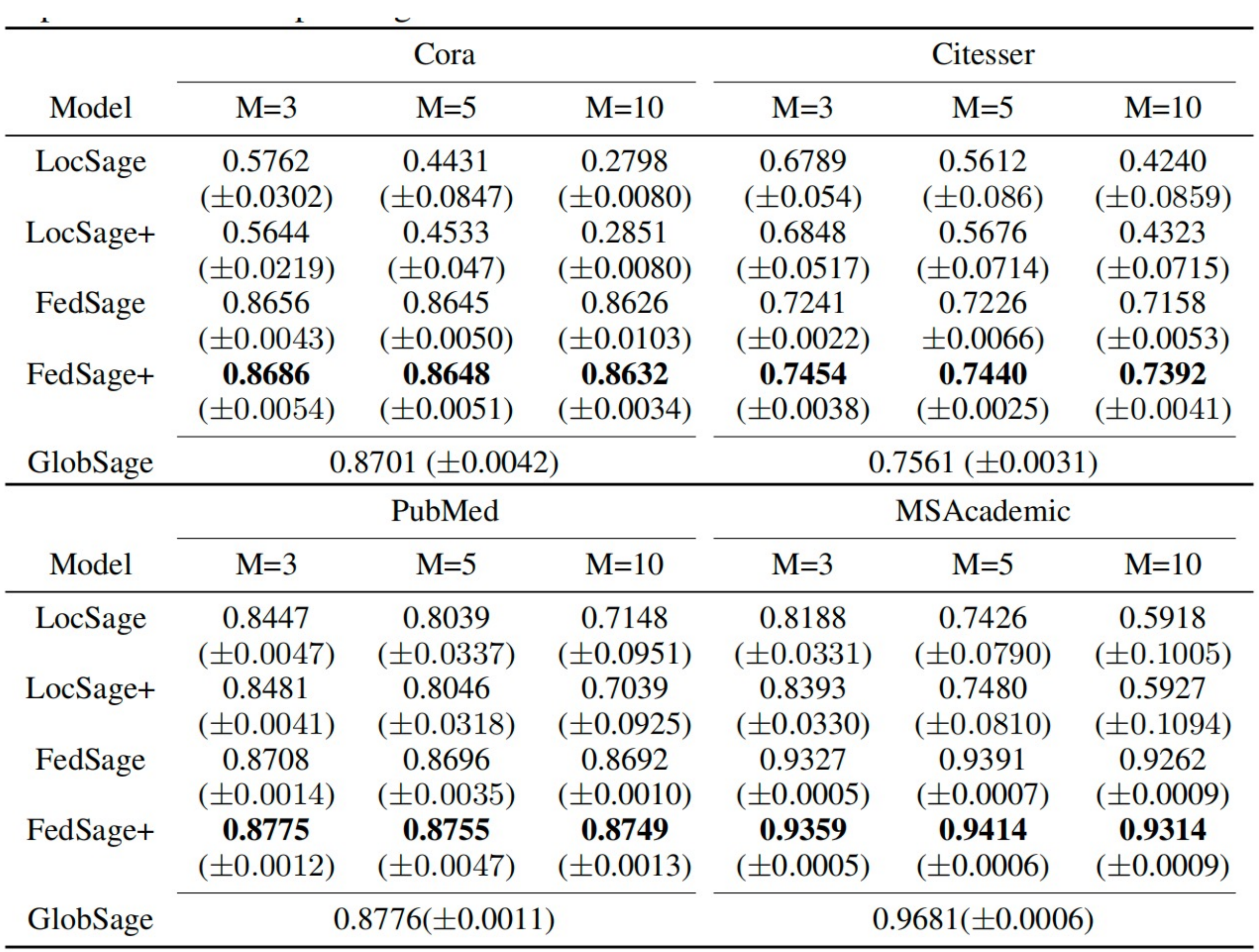


Data	Cora			Citeseer			PubMed			MSAcademic		
#C	7			6			3			15		
$\|V\|$	2708			3312			19717			18333		
$\|E\|$	5429			4715			44338			81894		
M	3	5	10	3	5	10	3	5	10	3	5	10
$\|V_i\|$	903	542	271	1104	662	331	6572	3943	1972	6111	3667	1833
$\|E_i\|$	1675	968	450	1518	902	442	12932	7630	3789	23584	13949	5915
ΔE	403	589	929	161	206	300	5543	6189	6445	11141	12151	22743

To synthesize the distributed subgraph system, we find hierarchical graph clusters on each dataset with the Louvain algorithm and use the clustering results with 3, 5, and 10 clusters of similar sizes to obtain subgraphs for data owners.

Subgraph Federated Learning with Missing Neighbor Generation (2021 NIPS)

Experiments



	Cora			Citesser		
Model	M=3	M=5	M=10	M=3	M=5	M=10
LocSage	0.5762 (±0.0302)	0.4431 (±0.0847)	0.2798 (±0.0080)	0.6789 (±0.054)	0.5612 (±0.086)	0.4240 (±0.0859)
LocSage+	0.5644 (±0.0219)	0.4533 (±0.047)	0.2851 (±0.0080)	0.6848 (±0.0517)	0.5676 (±0.0714)	0.4323 (±0.0715)
FedSage	0.8656 (±0.0043)	0.8645 (±0.0050)	0.8626 (±0.0103)	0.7241 (±0.0022)	0.7226 ±0.0066)	0.7158 (±0.0053)
FedSage+	**0.8686** (±0.0054)	**0.8648** (±0.0051)	**0.8632** (±0.0034)	**0.7454** (±0.0038)	**0.7440** (±0.0025)	**0.7392** (±0.0041)
GlobSage	0.8701 (±0.0042)			0.7561 (±0.0031)		

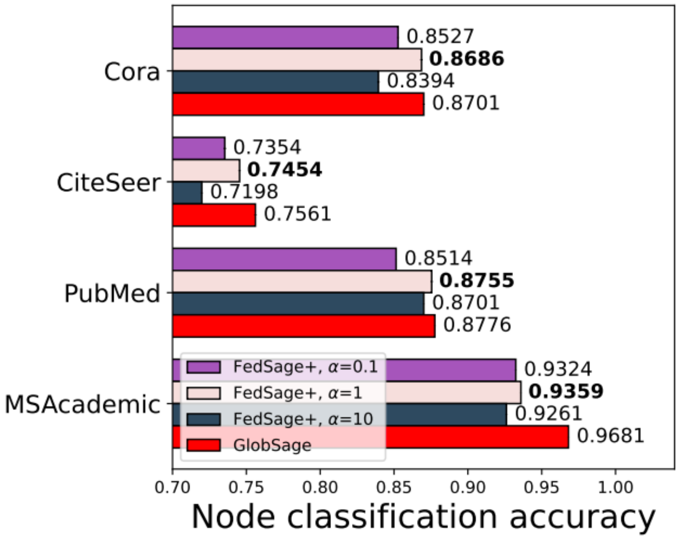
	PubMed			MSAcademic		
Model	M=3	M=5	M=10	M=3	M=5	M=10
LocSage	0.8447 (±0.0047)	0.8039 (±0.0337)	0.7148 (±0.0951)	0.8188 (±0.0331)	0.7426 (±0.0790)	0.5918 (±0.1005)
LocSage+	0.8481 (±0.0041)	0.8046 (±0.0318)	0.7039 (±0.0925)	0.8393 (±0.0330)	0.7480 (±0.0810)	0.5927 (±0.1094)
FedSage	0.8708 (±0.0014)	0.8696 (±0.0035)	0.8692 (±0.0010)	0.9327 (±0.0005)	0.9391 (±0.0007)	0.9262 (±0.0009)
FedSage+	**0.8775** (±0.0012)	**0.8755** (±0.0047)	**0.8749** (±0.0013)	**0.9359** (±0.0005)	**0.9414** (±0.0006)	**0.9314** (±0.0009)
GlobSage	0.8776(±0.0011)			0.9681(±0.0006)		

- **GlobSage**: the GraphSage model trained on the original global graph without missing links (as an upper bound for FL framework with GraphSage model alone)
- **LocSage**: one GraphSage model trained solely on each subgraph
- **LocSage+**: the GraphSage plus NeighGen model jointly trained solely on each subgraph

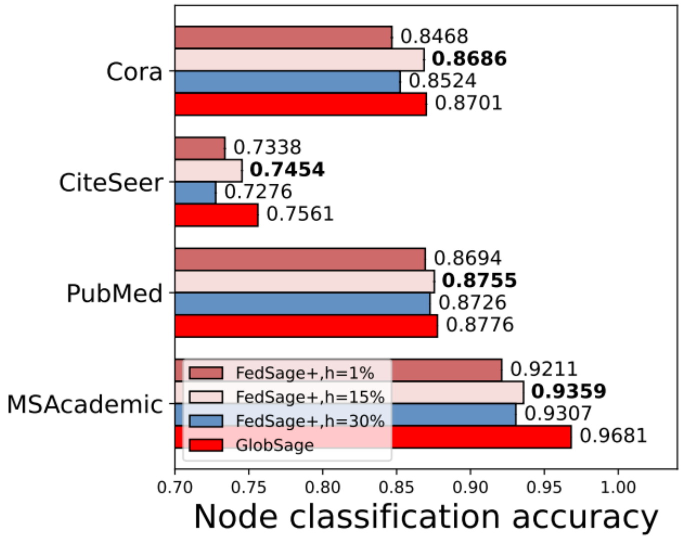
Subgraph Federated Learning with Missing Neighbor Generation (2021 NIPS)

Experiments

Hyper-parameter studies

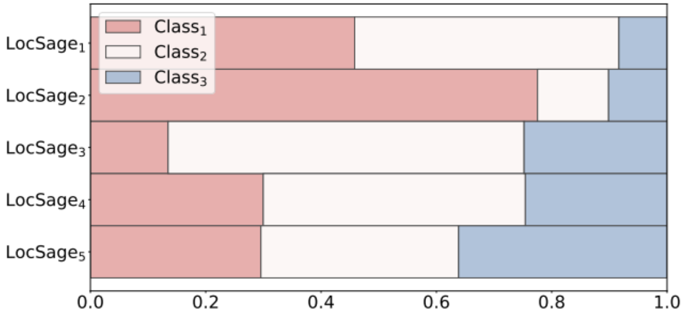


(a) Hyper-parameter study for α with $h = 15\%$.

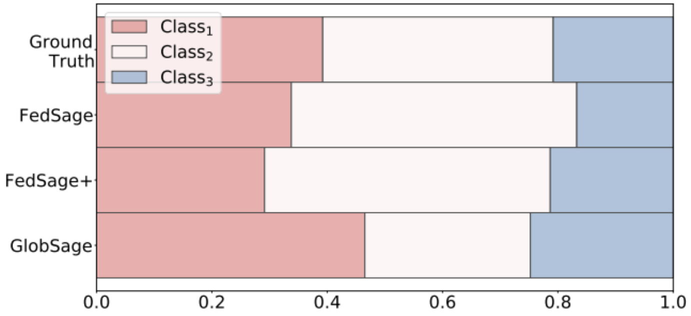


(b) Hyper-parameter study for h with $\alpha = 1$.

Case studies



(a) Local model predictions



(b) Global ground-truth vs. model predictions

Subgraph Federated Learning with Missing Neighbor Generation (2021 NIPS)

Conclusion

- This work aims at obtaining a ***generalized*** node classification model in a ***distributed subgraph system*** without direct data sharing.
- To tackle the realistic yet unexplored issue of ***missing cross-subgraph links,*** we design a novel ***missing neighbor generator*** NeighGen with the corresponding local and federated training processes.
- Experimental results evidence the distinguished elevation brought by our FedSage and FedSage+ frameworks , which is consistent with our theoretical implications.

目录

- 背景介绍

- 前置知识
  - 联邦学习（FL, Federated Learning）
  - 图神经网络（GNN, Graph Neural Network）

- 论文分享（FL+GNN)
  - (2021 NIPS) Subgraph Federated Learning with Missing Neighbor Generation
  - (2022 Nature Communications) A federated graph neural network framework for privacy-preserving personalization
  - (2023 ICML) Personalized Subgraph Federated Learning

A federated graph neural network framework for privacy-preserving personalization (2022 Nature Communications)

Background

- **Demand**
  - **Personalization** is a critical direction in the development of the Web
    - It can ease the burden of information overload by providing different users with different services based on their preferences and characteristics to better satisfy their personal needs.
  - E.g. Personalized healthcare services
    - can help people's health management and provide effective therapy plans based on an individual's mental and physical conditions.

- **Problem**
  - user data is usually highly privacy-sensitive and its centralized storage and exploitation can lead to users' ***privacy concerns and the risk of data leakage***
  - under the pressure of some strict ***data protection regulations*** such as General Data Protection Regulation (GDPR), online platforms may not be able to centrally store user data to learn GNN models for personalization in the future.

A federated graph neural network framework for privacy-preserving personalization (2022 Nature Communications)

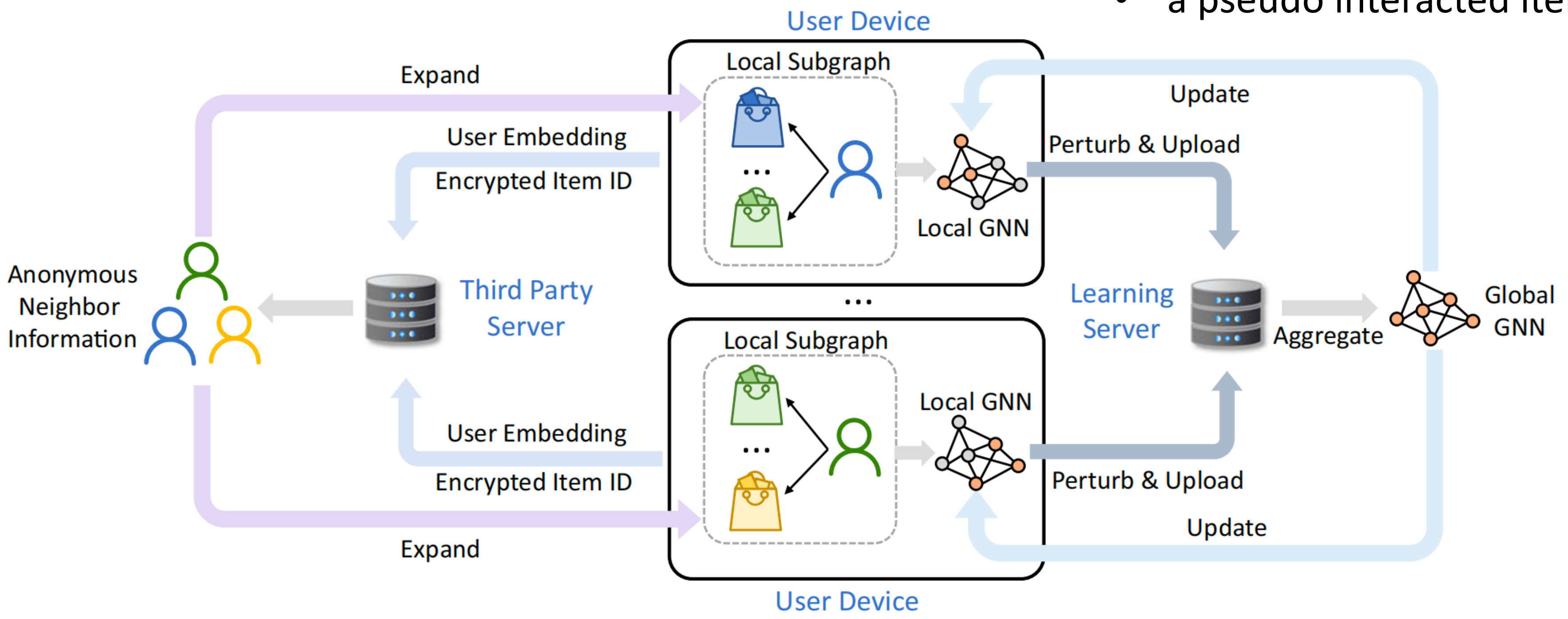
Challenge

- the local GNN model trained on local user data may convey private information
  - it is challenging to ***protect user privacy when synthesizing the global GNN model*** from the local ones
- the local user data may only contain ***first-order interactions*** between user and items, *higher-order interaction information is not available* since user data cannot be directly exchanged and linked among different clients due to privacy restrictions
  - Prior work on subgraph-level federated learning assumes that each client has a large subgraph and there is no sufficient interaction across different subgraphs decentralized on different clients.
  - However, in personalization scenarios *the decentralized subgraphs can be very small*, and the interactions across different subgraphs can be critical for understanding userinterest

A federated graph neural network framework for privacy-preserving personalization (2022 Nature Communications)

- mining high-order user-item interaction information
- local differential privacy (LDP)
- a pseudo interacted item sampling method

Overall framework of FedPerGNN



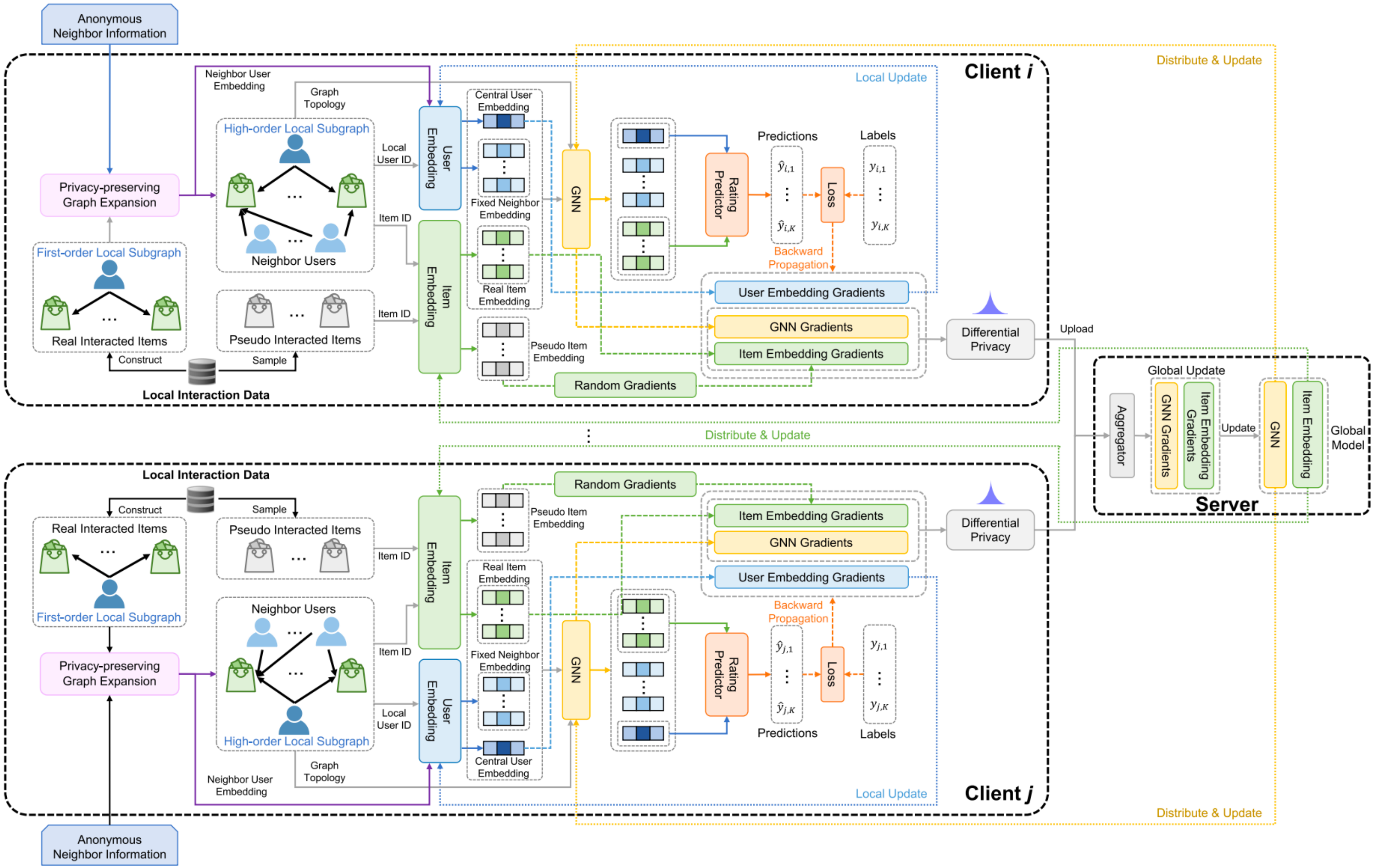
Privacy-Preserving Graph Expansion

Privacy-Preserving Model Update

- third-party server: conduct the ***privacy-preserving graph expansion protocol*** to incorporate high-order graph information into local model learning under privacy protection.
  - The devices upload the user embedding and encrypted item IDs to this server for finding user neighbors, and the embeddings of anonymous neighbor users are distributed to user devices for expanding local subgraphs.

A federated graph neural network framework for privacy-preserving personalization (2022 Nature Communications)

Detailed framework of FedPerGNN



A federated graph neural network framework for privacy-preserving personalization (2022 Nature Communications)

privacy-preserving user-item graph expansion protocol

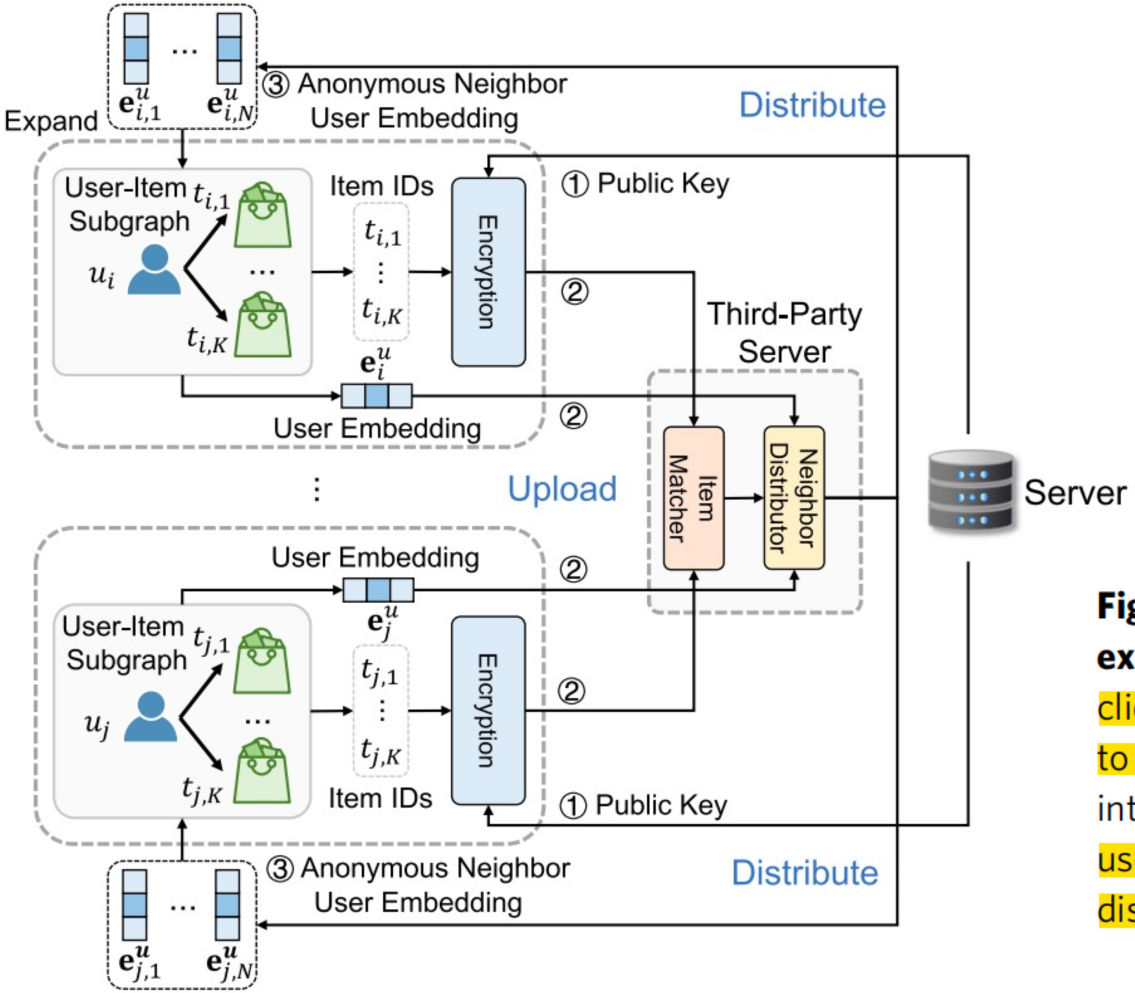


Fig. 8 The framework of the privacy-preserving user-item graph expansion protocol. The server first generates and sends a public key to clients for encrypting local item IDs, and the clients upload the ciphertexts to a third-party server for matching the same items. The users with co-interacted items are regarded as neighbors, and the anonymous neighbor user embeddings with their corresponding connected encrypted items are distributed to the clients for expanding local subgraphs.

A federated graph neural network framework for privacy-preserving personalization (2022 Nature Communications)

Experiments

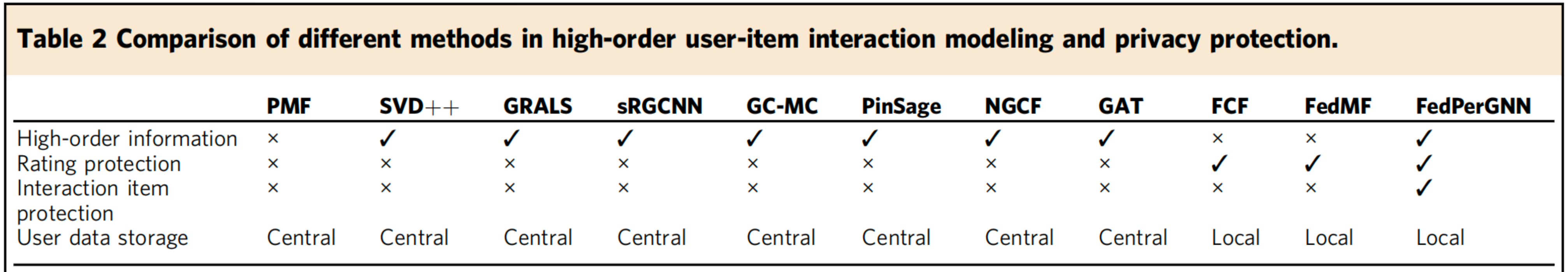
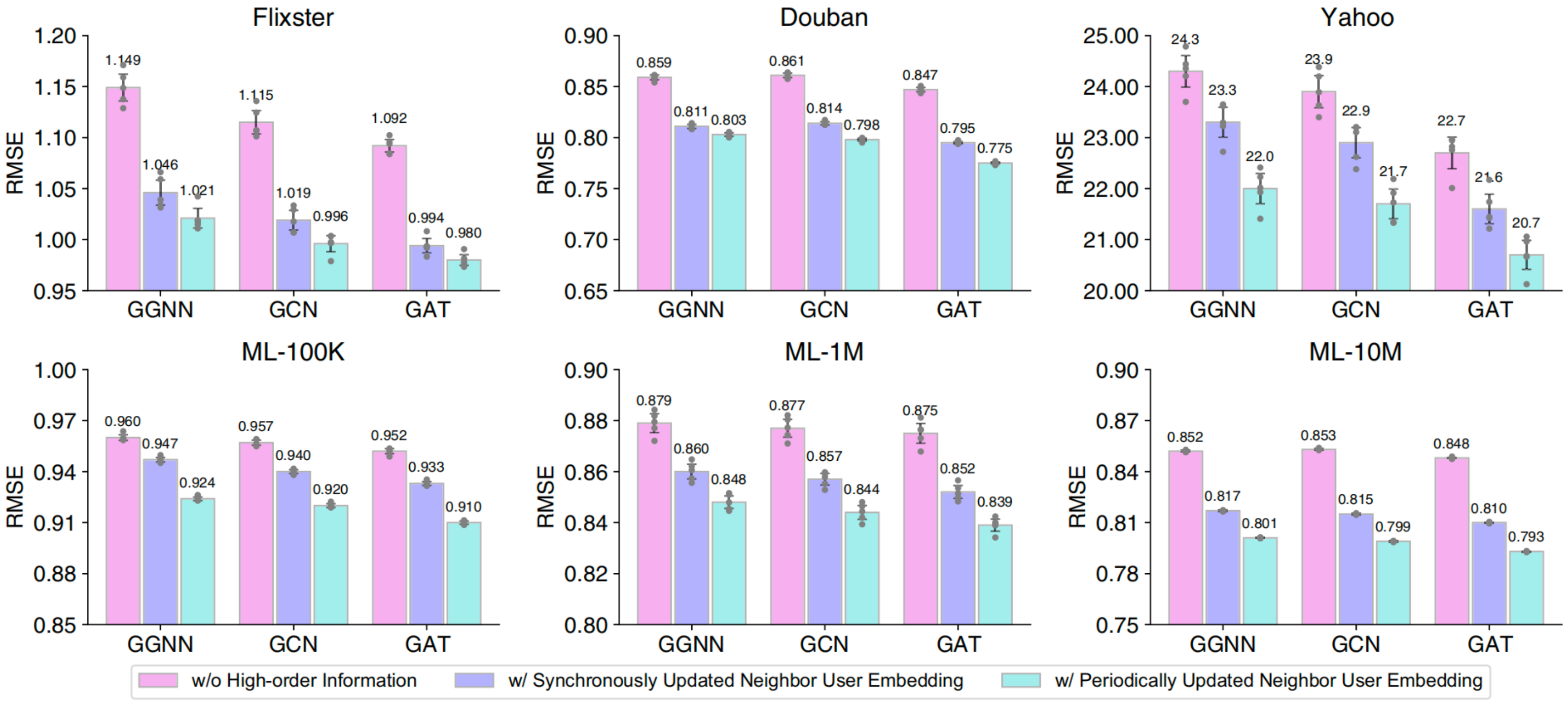


Table 2 Comparison of different methods in high-order user-item interaction modeling and privacy protection.

	PMF	SVD++	GRALS	sRGCNN	GC-MC	PinSage	NGCF	GAT	FCF	FedMF	FedPerGNN
High-order information	×	✓	✓	✓	✓	✓	✓	✓	×	×	✓
Rating protection	×	×	×	×	×	×	×	×	✓	✓	✓
Interaction item protection	×	×	×	×	×	×	×	×	×	×	✓
User data storage	Central	Central	Central	Central	Central	Central	Central	Central	Local	Local	Local

ormation but cannot protect user
ncorporate high-order information



目录

- 背景介绍

- 前置知识
  - 联邦学习（FL, Federated Learning）
  - 图神经网络（GNN, Graph Neural Network）

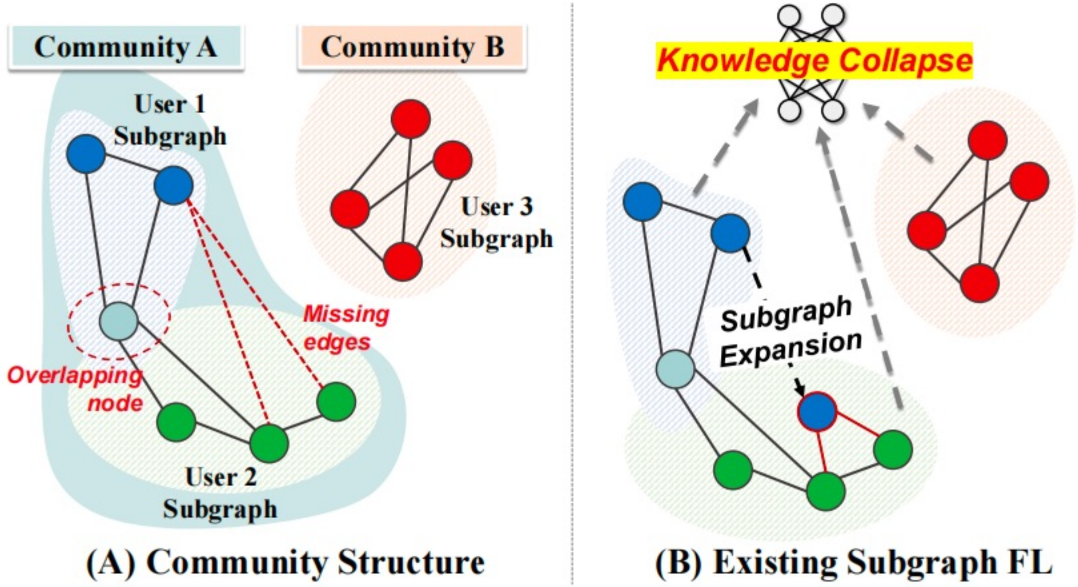
- 论文分享（FL+GNN)
  - (2021 NIPS) Subgraph Federated Learning with Missing Neighbor Generation
  - (2022 Nature Communications) A federated graph neural network framework for privacy-preserving personalization
  - (2023 ICML) Personalized Subgraph Federated Learning

Personalized Subgraph Federated Learning (2023 ICML)

Challenge

- How to deal with potentially ***missing edges*** between subgraphs that are not captured by individual data owners, but may carry important information?
  - expanding the local subgraph from other subgraphs
    - by exactly augmenting the relevant nodes from the other subgraphs at the other clients
    - by estimating the nodes using the node information in the other subgraphs
  - However, such sharing of node information may compromise data privacy and can incur high communication costs.

- ***heterogeneity*** among subgraphs
  - E.g. User 1 and 3 subgraphs in Communities A and B, respectively, in Figure 1 (A) – are sometimes completely disjoint, having opposite properties.

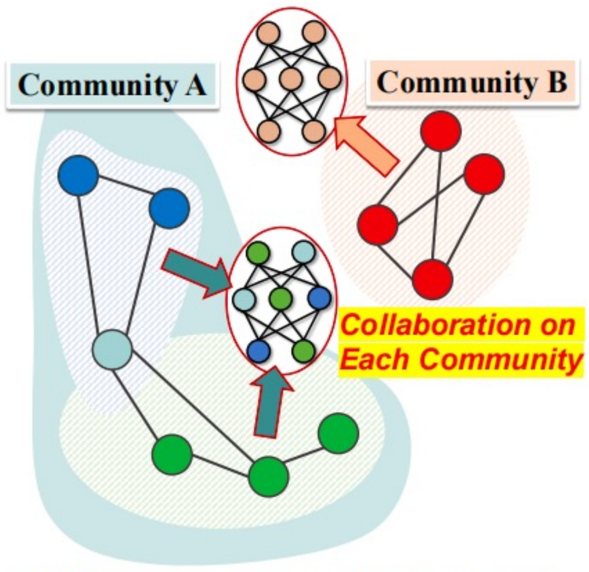


(A) Community Structure

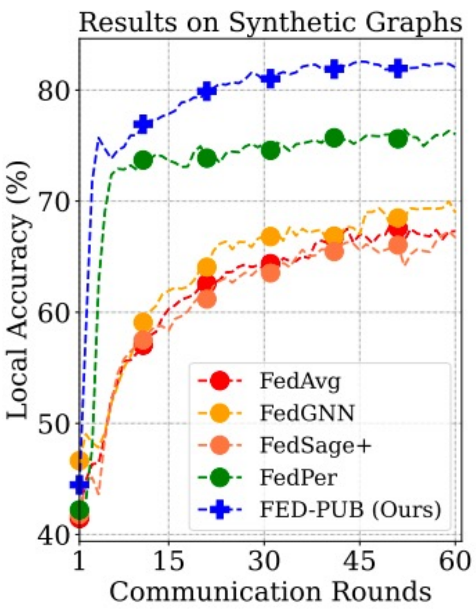
(B) Existing Subgraph FL

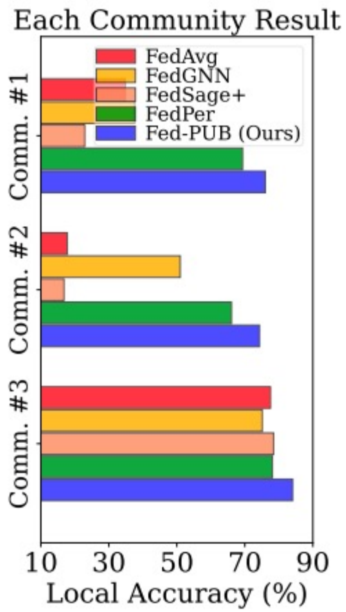
Personalized Subgraph Federated Learning (2023 ICML)

- Goal: jointly improve the interrelated local models trained on the ***interconnected local subgraphs***, for instance, subgraphs belonging to the same community, by sharing weights among them.



(C) Our Personalized Subgraph FL





$$\min_{\{\boldsymbol{\theta}_i,\boldsymbol{\mu}_i\}_{i=1}^{K}} \sum_{G_i \subseteq \mathcal{G}} \mathcal{L}(G_i;\boldsymbol{\theta}_i,\boldsymbol{\mu}_i),\ \boldsymbol{\theta}_i \leftarrow \boldsymbol{\mu}_i \odot \left(\sum_{j=1}^{K} \alpha_{ij}\boldsymbol{\theta}_j\right)$$

$$\text{with } \alpha_{ik} \gg \alpha_{il} \text{ for } G_k \subseteq C \text{ and } G_l \not\subseteq C, \qquad (2)$$

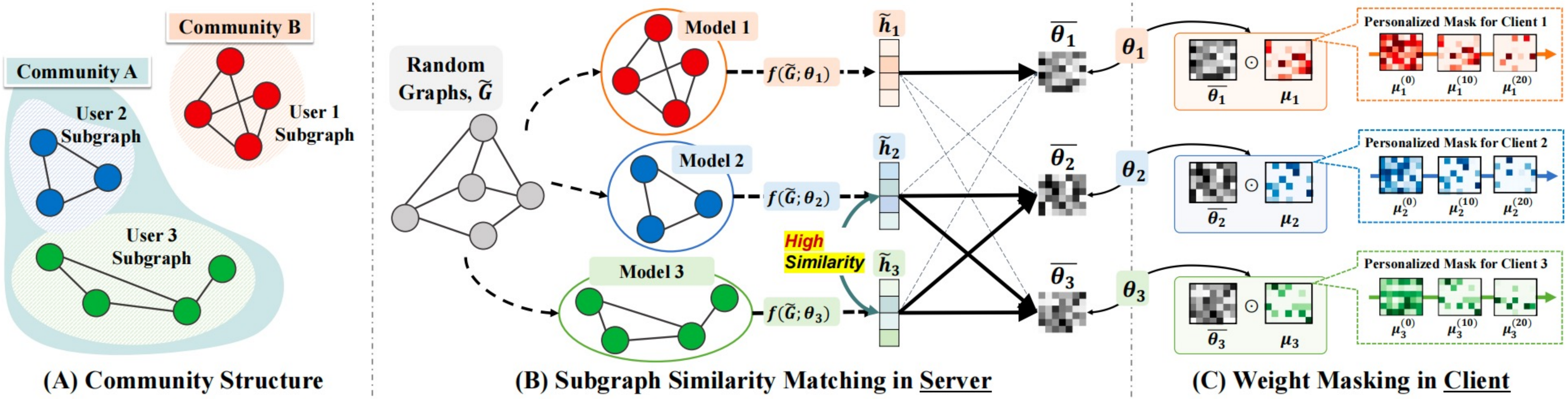
Knowledge collapse results
where local models belonging to two small communities (Communities 1 and 2)

Method Design:

- use functional embeddings of GNNs on random graphs to obtain ***similarity*** scores between two local GNNs
- use them to perform weighted averaging of the model parameters at the server
- learn and apply personalized sparse masks (the similarity scores of the parameters) on the local GNN at each client to obtain only the subnetwork

Personalized Subgraph Federated Learning (2023 ICML)

Method



(A) Community Structure (B) Subgraph Similarity Matching in Server (C) Weight Masking in Client

Figure 2: **(A) Two communities**, where Community A and B consist of two and one subgraphs, respectively. **(B) Similarity Matching**: we first forward randomly generated graphs to models $f(\tilde{G};\boldsymbol{\theta}_i)$, and obtain functional embeddings $\tilde{\boldsymbol{h}}_i$, which are then used to estimate subgraph similarities. Then, the similarities are used in weight aggregation, resulting in personalized model weights $\bar{\boldsymbol{\theta}}_i$. **(C) Weight Masking:** transmitted weights from the server to clients $\bar{\boldsymbol{\theta}}_i$ are masked and shifted by local masks $\boldsymbol{\mu}_i$ for localization to local subgraphs.

Personalized Weight Aggregation

$$\bar{\boldsymbol{\theta}}_i \leftarrow \sum_{j=1}^{K} \alpha_{ij} \cdot \boldsymbol{\theta}_j, \quad \alpha_{ij} = \frac{\exp(\tau \cdot S(i,j))}{\sum_k \exp(\tau \cdot S(i,k))},$$

Personalized Parameter Masking

$$\min_{(\boldsymbol{\theta}_i,\boldsymbol{\mu}_i)} \mathcal{L}(G_i;\boldsymbol{\theta}_i,\boldsymbol{\mu}_i) + \lambda_1\|\boldsymbol{\mu}_i\|_1 + \lambda_2\|\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}}_i\|_2^2.$$

Personalized Subgraph Federated Learning (2023 ICML)

Experiments

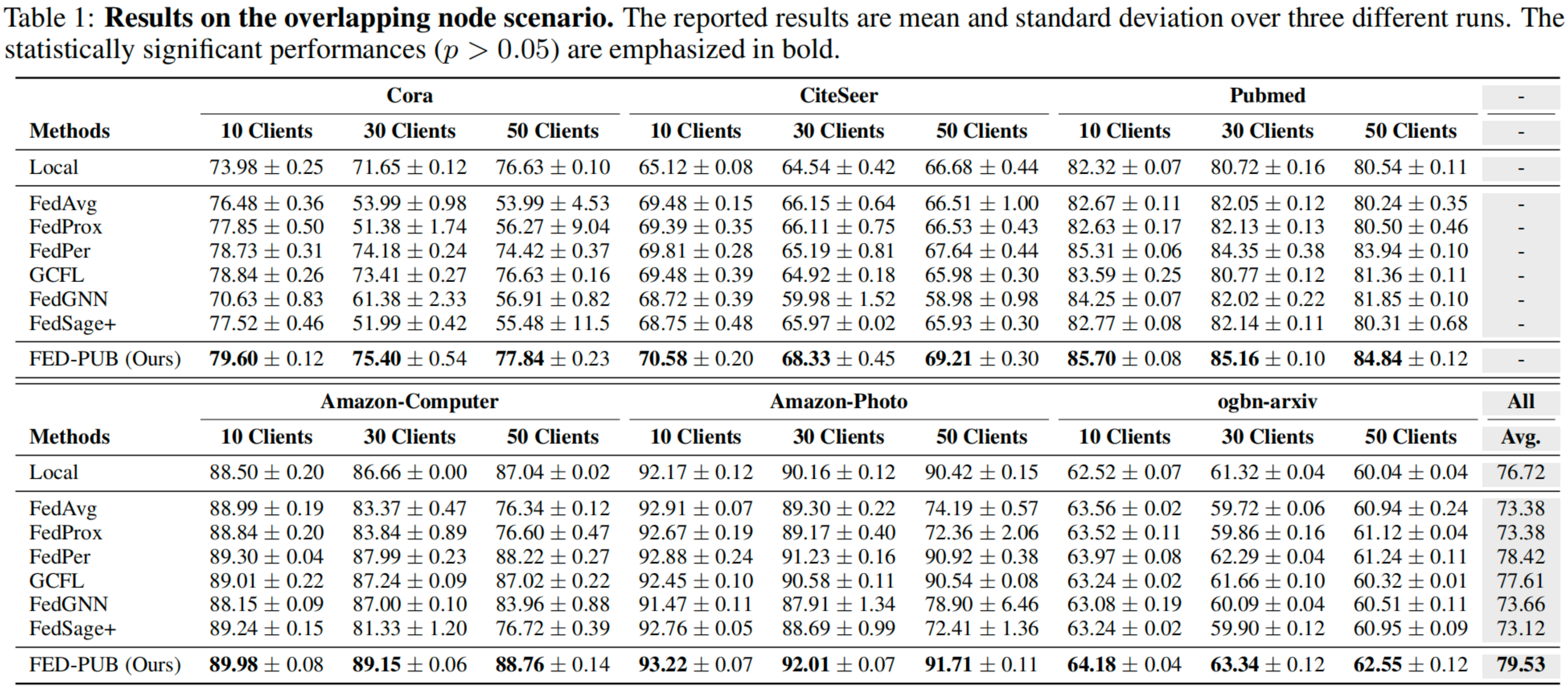


Table 1: **Results on the overlapping node scenario.** The reported results are mean and standard deviation over three different runs. The statistically significant performances ($p > 0.05$) are emphasized in bold.

	Cora			**CiteSeer**			**Pubmed**			-
Methods	**10 Clients**	**30 Clients**	**50 Clients**	**10 Clients**	**30 Clients**	**50 Clients**	**10 Clients**	**30 Clients**	**50 Clients**	-
Local	73.98 ± 0.25	71.65 ± 0.12	76.63 ± 0.10	65.12 ± 0.08	64.54 ± 0.42	66.68 ± 0.44	82.32 ± 0.07	80.72 ± 0.16	80.54 ± 0.11	-
FedAvg	76.48 ± 0.36	53.99 ± 0.98	53.99 ± 4.53	69.48 ± 0.15	66.15 ± 0.64	66.51 ± 1.00	82.67 ± 0.11	82.05 ± 0.12	80.24 ± 0.35	-
FedProx	77.85 ± 0.50	51.38 ± 1.74	56.27 ± 9.04	69.39 ± 0.35	66.11 ± 0.75	66.53 ± 0.43	82.63 ± 0.17	82.13 ± 0.13	80.50 ± 0.46	-
FedPer	78.73 ± 0.31	74.18 ± 0.24	74.42 ± 0.37	69.81 ± 0.28	65.19 ± 0.81	67.64 ± 0.44	85.31 ± 0.06	84.35 ± 0.38	83.94 ± 0.10	-
GCFL	78.84 ± 0.26	73.41 ± 0.27	76.63 ± 0.16	69.48 ± 0.39	64.92 ± 0.18	65.98 ± 0.30	83.59 ± 0.25	80.77 ± 0.12	81.36 ± 0.11	-
FedGNN	70.63 ± 0.83	61.38 ± 2.33	56.91 ± 0.82	68.72 ± 0.39	59.98 ± 1.52	58.98 ± 0.98	84.25 ± 0.07	82.02 ± 0.22	81.85 ± 0.10	-
FedSage+	77.52 ± 0.46	51.99 ± 0.42	55.48 ± 11.5	68.75 ± 0.48	65.97 ± 0.02	65.93 ± 0.30	82.77 ± 0.08	82.14 ± 0.11	80.31 ± 0.68	-
FED-PUB (Ours)	**79.60** ± 0.12	**75.40** ± 0.54	**77.84** ± 0.23	**70.58** ± 0.20	**68.33** ± 0.45	**69.21** ± 0.30	**85.70** ± 0.08	**85.16** ± 0.10	**84.84** ± 0.12	-

	Amazon-Computer			**Amazon-Photo**			**ogbn-arxiv**			**All**
Methods	**10 Clients**	**30 Clients**	**50 Clients**	**10 Clients**	**30 Clients**	**50 Clients**	**10 Clients**	**30 Clients**	**50 Clients**	**Avg.**
Local	88.50 ± 0.20	86.66 ± 0.00	87.04 ± 0.02	92.17 ± 0.12	90.16 ± 0.12	90.42 ± 0.15	62.52 ± 0.07	61.32 ± 0.04	60.04 ± 0.04	76.72
FedAvg	88.99 ± 0.19	83.37 ± 0.47	76.34 ± 0.12	92.91 ± 0.07	89.30 ± 0.22	74.19 ± 0.57	63.56 ± 0.02	59.72 ± 0.06	60.94 ± 0.24	73.38
FedProx	88.84 ± 0.20	83.84 ± 0.89	76.60 ± 0.47	92.67 ± 0.19	89.17 ± 0.40	72.36 ± 2.06	63.52 ± 0.11	59.86 ± 0.16	61.12 ± 0.04	73.38
FedPer	89.30 ± 0.04	87.99 ± 0.23	88.22 ± 0.27	92.88 ± 0.24	91.23 ± 0.16	90.92 ± 0.38	63.97 ± 0.08	62.29 ± 0.04	61.24 ± 0.11	78.42
GCFL	89.01 ± 0.22	87.24 ± 0.09	87.02 ± 0.22	92.45 ± 0.10	90.58 ± 0.11	90.54 ± 0.08	63.24 ± 0.02	61.66 ± 0.10	60.32 ± 0.01	77.61
FedGNN	88.15 ± 0.09	87.00 ± 0.10	83.96 ± 0.88	91.47 ± 0.11	87.91 ± 1.34	78.90 ± 6.46	63.08 ± 0.19	60.09 ± 0.04	60.51 ± 0.11	73.66
FedSage+	89.24 ± 0.15	81.33 ± 1.20	76.72 ± 0.39	92.76 ± 0.05	88.69 ± 0.99	72.41 ± 1.36	63.24 ± 0.02	59.90 ± 0.12	60.95 ± 0.09	73.12
FED-PUB (Ours)	**89.98** ± 0.08	**89.15** ± 0.06	**88.76** ± 0.14	**93.22** ± 0.07	**92.01** ± 0.07	**91.71** ± 0.11	**64.18** ± 0.04	**63.34** ± 0.12	**62.55** ± 0.12	**79.53**

Thanks