计算机科学
COMPUTER SCIENCE

面向纵向图联邦学习的数据重构攻击方法

李荣昌, 郑海斌, 赵文红, 陈晋音

引用本文

李荣昌, 郑海斌, 赵文红, 陈晋音.面向纵向图联邦学习的数据重构攻击方法[J]. 计算机科学, 2023, 50(7): 332-338.

LI Rongchang, ZHENG Haibin, ZHAO Wenhong, CHEN Jinyin. Data Reconstruction Attack for Vertical Graph Federated Learning [J]. Computer Science, 2023, 50(7): 332-338.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900038

面向纵向图联邦学习的数据重构攻击方法

李荣昌[1]　郑海斌[1]　赵文红[2]　陈晋音[1,3]
1 浙江工业大学信息工程学院　杭州 310023
2 嘉兴南湖学院信息工程学院　浙江 嘉兴 314001
3 浙江工业大学网络空间安全研究院　杭州 310023
(lrcgnn@163.com)

摘　要　近年来,数据隐私保护法规限制了不同图数据拥有者之间的数据直接交换,出现了“数据孤岛”现象。为解决上述问题,纵向图联邦学习通过秘密交换嵌入表示的方式实现图数据分布式训练,在众多现实领域具有广泛应用,如药物研发、用户发掘以及商品推荐等。然而,纵向图联邦学习中的诚实参与方在训练过程中仍然存在隐私泄露的风险,为此提出了一个由诚实但好奇的参与方基于生成式网络发动嵌入表示重构攻击,通过范数损失函数使得生成式网络的输出结果向训练公布的置信度逼近,从而重构参与方的隐私数据。实验结果表明,所提嵌入表示重构攻击在Cora,Citeseer以及Pubmed数据集上均能完整地重构参与方的嵌入表示,凸显了纵向图联邦学习中参与方嵌入表示的隐私泄露风险。

关键词:图神经网络;隐私泄露;联邦学习;生成式网络;差分隐私
中图法分类号　TP391

Data Reconstruction Attack for Vertical Graph Federated Learning

LI Rongchang[1],ZHENG Haibin[1],ZHAO Wenhong[2] and CHEN Jinyin[1,3]
1 College of Information Engineering,Zhejiang University of Technology,Hangzhou 310023,China
2 College of Information Engineering,Jiaxing Nanhu University,Jiaxing,Zhejiang 314001,China
3 Institute of Cyberspace Security,Zhejiang University of Technology,Hangzhou 310023,China

Abstract　Recently,data privacy protection regulations restrict the direct exchange of raw data between different graph data owners,resulting in the phenomenon of “data silos”. To solve this problem,vertical federated learning graph neural network realizes distributed training of graph data by secretly exchanging embeddings,and has been widely used in many real-world fields,such as drug discovery,user discovery,and product recommendation. However,honest participants in vertical federated learning graph neural network still have the risk of privacy leakage during training. This paper proposes a private embedding representation reconstruction attack based on the generative network,and reconstructs the private data of the participant by the output of the generative network is approximated with the confidence published from server with the norm loss function. Experimental results show that the embedding representation reconstruction attack can completely reconstruct the embedding representation of the participants on the Cora,Citeseer and Pubmed datasets,which highlights the risk of leakage of the participant embedding representation in VFL-GNN.

Keywords　Graph neural network,Privacy leakage,Federated learning,Generative network,Differential privacy

1　引言

近年来,图数据作为一种重要的数据类型在众多领域具有广泛的应用,如推荐系统[1]、智能交通[2],以及文本识别[3]等。为了从图数据中提取丰富的信息,图神经网络(Graph Neural Network,GNN)将某个节点或者图映射为一个低维的嵌入表示向量,以捕获节点特征和节点之间的连边关系。现实场景中的图数据往往存在于不同的机构中,例如一家互联公司具有用户A的“消费记录”信息,另一家通信公司具有用户A的“社交关系”信息。由于数据隐私保护法律法规的

到稿日期:2022-09-05　返修日期:2022-12-05
基金项目:国家自然科学基金(62072406);信息系统安全技术重点实验室基金(61421110502);浙江省重点研发计划(2021C01117);2020年工业互联网创新发展工程项目(TC200H01V);浙江省“万人计划”科技创新领军人才项目(2020R52011)
This work was supported by the National Natural Science Foundation of China(62072406),National Key Laboratory of Science and Technology on Information System Security(61421110502),Key R&D Projects in Zhejiang Province(2021C01117),2020 Industrial Internet Innovation Development Project(TC200H01V) and “Ten Thousand Talents Program” in Zhejiang Province(2020R52011).
通信作者:陈晋音(chenjinyin@zjut.edu.cn)

限制，如欧盟发布的《通用数据保护条例》[4]和中国颁布的《中华人民共和国数据安全法》[5]，不同机构的数据无法直接集中训练图神经网络模型，从而出现“数据孤岛”现象[6]。

为了解决上述问题，Zhou 等[7]提出了满足数据隐私保护法规的图数据分布式训练技术——纵向图联邦学习（Vertical Federated Learning GNN，VFL-GNN）。不同于直接传输原始数据，纵向图联邦学习在不同的参与方之间秘密传输嵌入表示，实现快速且秘密的训练。具体而言，首先各个本地参与方利用图神经网络提取嵌入表示，并利用同态加密技术[8]上传嵌入表示至服务器，然后服务器聚合不同参与方的嵌入表示信息进行解密，并完成在服务器模型的前向传播，最后更新本地参与方的本地模型和服务器的顶部模型参数，以此迭代直至纵向图联邦学习完成训练。

纵向图联邦学习取得优异性能[8-11]的同时，该算法本身暴露出的隐私泄露风险也成为研究重点。文献[12]表明通过图神经网络模型生成的嵌入表示可以推测参与方的隐私属性、连边信息和子图隶属信息。嵌入表示作为参与方的重要知识产权，其泄露会对参与方的隐私造成严重破坏。因此，嵌入表示通常被视为纵向图联邦学习中参与方的重要隐私数据[11-15]。尽管纵向图联邦学习中参与方以加密的方式上传嵌入表示至服务器，然而我们发现诚实的参与方上传的嵌入表示仍然可能被诚实且好奇的攻击者完整重构。这是因为在正常的纵向图联邦学习协议中[7]，服务器在训练过程中会实时公布指导参与方本地模型参数更新的置信度信息。我们发现，攻击者可以利用该置信度指导参与方进行隐私嵌入表示的恢复。

基于上述发现，本文提出了一个基于生成式网络的嵌入表示重构攻击，攻击者在遵守纵向图联邦学习的协议基础上实现窃取参与方的隐私嵌入表示。具体而言，攻击者在纵向图联邦学习训练过程中，首先按照协议上传本地的嵌入表示数据，并记录服务器公布的当前轮次置信度信息；然后输入随机初始化噪声至生成式网络获得优化噪声；最后拼接优化噪声和参与方的本地嵌入表示，并输入服务器模型查询置信度信息，引入范数损失函数反向优化生成式网络模型参数；迭代训练直至生成式网络收敛。

本文的主要工作包括：

（1）首次评估了通用的纵向图联邦学习框架的隐私性，揭示诚实的参与方在训练过程中嵌入表示存在隐私泄露的风险。

（2）提出了一个基于生成式网络的嵌入表示重构攻击，攻击者在遵守正常的纵向图联邦学习协议基础上实现重构参与方的嵌入表示。

（3）在 3 个现实数据集上展开实验评估，实验结果表明，恶意参与方对纵向图联邦学习中诚实参与方的嵌入表示具有较好的重构效果，本文呼吁更多研究针对纵向图联邦学习提出有效的隐私保护策略。

2 相关工作

本节介绍了与 VFL-GNN 相关的现有工作，共分为 3 个部分：常见 VFL-GNN 方法研究、VFL 上的隐私泄露研究，以及 GNN 上的隐私泄露研究。

2.1 VFL-GNN 方法

谷歌首次提出联邦学习[16]，用以实现保护数据隐私的分布式训练，随后 Yang 等[6]根据参与方的数据分布特点将联邦学习扩展为 3 种常见的概念：横向联邦学习、纵向联邦学习，以及联邦迁移学习。随着现实中分布在不同机构的图数据联合满足数据保护法规情况下的分布式训练任务需求，纵向图神经网络成为技术研究热点。Ni 等[9]提出了一种图联邦学习框架，用于处理数据纵向分布情况下的节点分类任务，可以推广到现有的神经网络模型中。在文献[9]中，对于训练过程的每次迭代，参与方利用同态加密技术将中间结果进行加密并相互交换，用于训练图神经网络。He 等[10]提出了一个基于图神经网络的联邦学习系统，并对图联邦学习的训练效率展开了研究。Wu 等[17]提出了一个基于图神经网络模型的隐私保护联邦推荐框架，并针对推荐系统进行了隐私保护的优化。上述研究都致力于提高纵向图联邦框架的准确性和有效性，并在现实商用场景中得到了应用。

2.2 VFL 隐私泄露

在纵向联邦学习中，参与方和服务器之间交换的嵌入表示包含隐私数据和模型相关信息，其泄露将直接导致参与方的隐私泄露。通常来说，通过嵌入表示的方式会导致两类信息隐私泄露：标签信息和原始数据信息。一方面，攻击者直接通过参与方传输的嵌入表示推理标签信息，Li 等[18]使用参与方和服务器之间相互交换的梯度范数，揭露目标参与方的标签信息。Fu 等[19]提出利用本地模型信息来推断参与方的隐私信息。另一方面，Weng 等[20]针对传统的纵向联邦学习设计并实现了两种不影响学习模型准确性的推断攻击：反向求和攻击和反向乘法攻击。Jiang 等[21]提出了基于梯度的重建攻击框架，该框架可以应用于联邦学习逻辑回归模型。然而该研究针对的是基于传统机器学习的纵向联邦学习，并不适用于纵向图联邦学习。此外，Jin 等[22]针对图像数据在纵向联邦学习框架中提出了一种基于梯度的批量恢复图像数据方法。不同于上述研究，本文关注的是纵向图联邦学习中泄露的隐私数据为参与方上传给服务器的嵌入表示。

2.3 GNN 隐私泄露

Duddu 等[11]首次针对 GNN 提出 3 种推断攻击来评估其面临的隐私泄露风险。针对 GNN 提取的嵌入表示发动 3 种类型的推断攻击：成员推断攻击推断单个用户数据的图节点是否参与模型的训练；图重构攻击根据目标图嵌入表示重构图数据；属性推断攻击根据嵌入表示推断敏感属性信息。不同于文献[11]，Zhang 等[12]根据图嵌入表示推断图中的网络指标，如图密度、连边数量，以及节点数量；此外，他们还提出了分类网络，根据目标图嵌入表示推断子图是否存在目标图中。Wang 等[13]基于知识图谱中的嵌入表示提出了一种成员推断攻击。Wu 等[14]针对整个图提出了成员推断攻击来探索不同模型结构的鲁棒性。上述研究都表明图神经网络模型提取的嵌入表示包含重要的隐私信息，因此嵌入表示属于数据

拥有者的重要知识产权。然而上述研究都假设攻击者可以直接获得被攻击者上传的嵌入表示，这在纵向联邦学习场景中并不现实，因为嵌入表示往往是作为私有信息保存在参与方中。为此，本文研究纵向联邦学习中参与方嵌入表示泄露风险的问题。

3 通用 VFL-GNN 方法

本节介绍了一种通用的 VFL-GNN 方法，首先介绍基本的定义，然后展示基本框架的实现过程，最后介绍基于差分隐私技术保护隐私的 VFL-GNN 方法。

3.1 基本定义

定义 $\boldsymbol{U}=\{u_1,u_2,\cdots,u_m\}$ 和 $\boldsymbol{L}=\{l_1,l_2,\cdots,l_m\}$ 为纵向图联邦学习中参与方的数据集合以及参与方的标签集合，其中 m 是参与方的数量。记 $\boldsymbol{M}=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_m\}$ 为纵向图联邦学习中参与方的本地模型，$\boldsymbol{M}_0$ 为服务器的顶部模型。本文将纵向图联邦学习中用户之间的交互关系记为 $\boldsymbol{Y}\in\mathbb{R}^{m\times m}$。纵向图联邦学习中，为了保护参与方的数据隐私，每个参与方的本地图数据都不会离开参与方自身设备。本文提出的 VFL-GNN 实现下游任务的目标是为节点分类，即 VFL-GNN 基于参与者存储的用户交互数据来预测每个用户的所属类型。与集中式训练 GNN 的方式不同，VFL-GNN 中用户之间的交互信息分别存放于各自参与方本地设备中，而前者要求所有图数据都集中存放于一个设备中用于训练。

3.2 基本框架

本节介绍基于图卷积神经网络[23](Graph Convolutional Network，GCN)的纵向图联邦学习训练框架 VFL-GNN。算法 1 为纵向图联邦学习的训练流程。利用高度分散的用户交互数据，通过隐私保护的方式来学习 GCN 模型完成节点分类任务。VFL-GNN 的框架如图 1 所示，算法 1 展示了训练流程，它由一个服务器和多个参与方组成。参与方各自的本地数据由用户之间的部分交互历史组成，每个参与方采用 GCN 模型从其本地子图学习用户交互嵌入，并将嵌入上传到服务器。中央服务器负责聚合从多个参与方接收的嵌入表示，并将聚合的嵌入输入服务器顶部模型中，得到全局的预测输出并采用损失函数计算损失值。最后服务器将预测结果和损失值公布给每个参与方，指导各参与方更新各自的本地模型。接下来将详细介绍 VFL-GNN 的工作原理。

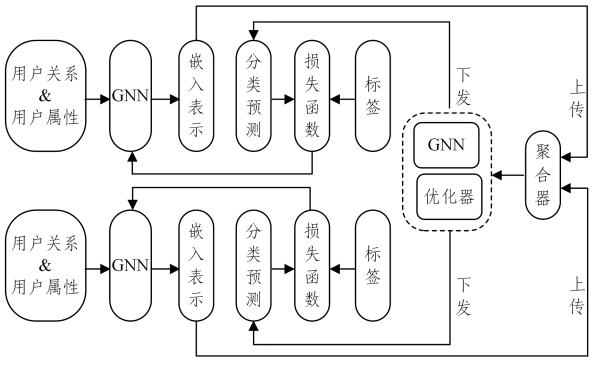


图 1 VFL-GNN 框架图

Fig. 1 VFL-GNN framework

算法 1 VFL-GNN

输入：客户端和服务器初始化模型 $\mathbf{M}_i$；客户端载入各自的数据集本地子图 $\mathbf{G}_i$

输出：预测结果 pred

1. 初始化本地 GCN 模型 $\mathbf{M}_i$
2. 初始化优化器与各项参数，模型进入训练模式
3. 得到指令 msg
4. if msg == “phase1”：
   4.1. 获得当前周期的本地嵌入 $\mathbf{e}_i^u$
   4.2. $\mathbf{e}_i^u \leftarrow \mathbf{e}_i^u$ + privacy(·)
   4.3. server $\leftarrow e_i^u$, time
   4.4. 传递指令 msg
5. if msg == “phase2”：
   5.1. $L_i^{epoch}, loss_i^{epoch} \leftarrow$ server
   5.2. $loss_i^u = loss_i^{epoch}$，并更新本地模型
   5.3. server $\leftarrow$ time，并传递指令 msg
6. if c == setting：
   6.1. 本地模型停止训练
   6.2. server $\leftarrow \mathbf{e}_i^u$ + privacy(·), $\mathbf{e}_i^u$
   6.3. $L_i^{test} \leftarrow \mathbf{G}_i$
   6.4. 计算准确率
   6.5. c $\leftarrow$ c+1
7. 初始化本地 GCN 模型 $\mathbf{M}_0$
8. 初始化优化器与参数，模型进入训练模式
9. if msg == “embedding”：
   9.1. $\mathbf{e}_i^u \leftarrow u_i$
10. if epoch == setting：
   10.1. $\mathbf{e}_s^u = \mathbf{e}_1^u \circ \mathbf{e}_2^u \circ \cdots \circ \mathbf{e}_m^u$
   10.2. $L_i^{epoch}, loss_i^{epoch} \leftarrow \mathbf{M}_0$
   10.3. 更新模型 $\mathbf{M}_0$
11. 当 msg 为“loss and pred”时，输出 L_i^{epoch} 以及 $loss_i^{epoch}$
12. Return pred

每个参与方的本地图数据均由用户与用户之间的部分交互数据构成。注意，在纵向图联邦学习中，每个参与方的 ID 信息都是自动进行匹配的，而用户间的交互数据是部分未知的。本地模型用于将用户节点 $[u_1,u_2,\cdots,u_m]$ 和每个用户 u_i 的交互信息 $[y_{i,1},y_{i,2},\cdots,y_{i,m}]$ 转换成对应的嵌入表示 $\boldsymbol{e}^u=[\boldsymbol{e}_1^u,\boldsymbol{e}_2^u,\cdots,\boldsymbol{e}_m^u]$，所有的嵌入 $\boldsymbol{e}^u$ 将被进一步上传到中央服务器进行聚合。

服务器的目标是协调所有参与方并计算全局损失，以更新模型并向这些参与方返回损失。在每一轮中，所有的参与方首先在本地计算嵌入并将其发送到服务器。服务器接收到来自参与方的嵌入表示 $\boldsymbol{e}^u=[\boldsymbol{e}_1^u,\boldsymbol{e}_2^u,\cdots,\boldsymbol{e}_m^u]$，服务器中的聚合器将这些本地嵌入聚合成一个统一的嵌入，即：

$$\boldsymbol{e}_s^u=\boldsymbol{e}_1^u\circ\boldsymbol{e}_2^u\circ\cdots\circ\boldsymbol{e}_m^u \tag{1}$$

其中，$\circ$ 代表按照特定维度的拼接操作。

然后，服务器将聚合后的嵌入 $\boldsymbol{e}_s^u$ 输入顶部模型 $\boldsymbol{M}_0$ 中完成前向传播，得到全局的预测结果 L_i^{epoch} 和损失 $loss_i^{epoch}$。该预测结果 L_i^{epoch} 和损失 $loss_i^{epoch}$ 将发送给每个参与方进行参与方本地模型的参数更新。将参与方本地模型中的损失记为

$loss_i^u$，它由 $loss_i^u = loss_i^{\text{epoch}}$ 来更新。迭代更新模型，直到全局模型收敛。

3.3 基于差分隐私的 VFL-GNN 框架

本节介绍基于差分隐私技术的 3 种常见的纵向图联邦学习框架，主要有 3 种不同的加噪方式，即同态加噪、拉普拉斯加噪和高斯加噪。

差分隐私保护，即向参与者输出的嵌入表示添加随机噪声，以达到保护隐私的目的。在联邦学习中，差分隐私可以隐藏参与者的大部分属性。设有算法 M，及其所有可能输出的集合 $\boldsymbol{R}$，若对于任意两个相邻集合 $\boldsymbol{D}$ 和 $\boldsymbol{D}'$，以及 $\boldsymbol{R}$ 的子集 $\boldsymbol{S}$，满足以下条件：

$$\Pr[M(\boldsymbol{D})\in\boldsymbol{S}]\leqslant e^{\varepsilon}\times\Pr[M(\boldsymbol{D}')\in\boldsymbol{S}] \tag{2}$$

则称该算法满足差分隐私。在实际中，由于上式过于严格，需要更多的隐私预算，导致主任务的性能下降。因此，为了算法的实用性，通常引入松弛版本的差分隐私，即：

$$\Pr[M(\boldsymbol{D})\in\boldsymbol{S}]\leqslant e^{\varepsilon}\times\Pr[M(\boldsymbol{D}')\in\boldsymbol{S}]+\delta \tag{3}$$

Abadi 等[24]提出差分隐私可保护参与者的隐私，可以用来防御属性推断攻击。依据添加噪声类型的不同，纵向图联邦学习框架可以分为 3 种方式。

(1)同态加噪

不考虑隐私预算，直接生成均值、标准差与嵌入相同的正态分布随机噪声。在介绍具有差分保护的 VFL-GNN 方法前，首先介绍差分隐私的敏感度 Δf 概念。给定两个孪生数据集 $\boldsymbol{D},\boldsymbol{D}'$，它们之间至多相差一条数据，而 Δf 代表的就是一个查询函数的最大变化范围。

(2)拉普拉斯加噪

$\Delta f=\max\limits_{\boldsymbol{D},\boldsymbol{D}'}\|f(\boldsymbol{D})-f(\boldsymbol{D}')\|_1$，噪声 $Y\sim L\left(0,\frac{\Delta f}{\varepsilon}\right)$ 满足 $(\varepsilon,0)-dp$[25]。

(3)高斯加噪

$\Delta f=\max\limits_{\boldsymbol{D},\boldsymbol{D}'}\|f(\boldsymbol{D})-f(\boldsymbol{D}')\|_2$，对于 $\forall\delta\in(0,1)$，$\sigma>\frac{\Delta f\sqrt{2\ln\left(\frac{1.25}{\delta}\right)}}{\varepsilon}$，有噪声 $Y\sim N(0,\sigma^2)$ 满足 $(\varepsilon,\delta)-dp$[25]。

算法 2 给出了基于差分隐私保护的 VFL-GNN 算法伪代码。

算法 2 基于差分隐私的 VFL-GNN 实现

输入：所需噪声种类 kind，真实嵌入表示 embedding，敏感度参数 Δf，隐私预算 ε，松弛项参数 δ

输出：噪声 n

1. 初始化噪声向量 **n**
2. if Kind == “normal”：
   2.1. 随机生成均值、标准差与 embedding 相同的正态分布噪声数据
3. if Kind == “Gaussian”：
   3.1. sigma $=\frac{\Delta f\sqrt{2\ln\left(\frac{1.25}{\delta}\right)}}{\varepsilon}$
   3.2. 随机生成均值为 0，标准差为 sigma 的噪声
4. if Kind == “Laplace”：
   4.1. 随机生成均值为 0，标准差为 $\frac{\Delta f}{\varepsilon}$ 噪声
5. Return **n**

4 VPATTACK 攻击方法

4.1 攻击定义

纵向图联邦学习中通常包含 1 个服务器和多个参与方。不失一般性，本文设定纵向图联邦学习中存在两个参与方，其中一个参与方为诚实但好奇的攻击者 P_{adv}；另一个参与方 P_{target} 是诚实的纵向图联邦学习参与成员。在上述纵向图联邦学习场景中，两个参与方都遵守纵向图联邦学习的正常训练协议。攻击者 P_{adv} 的攻击目标是推测诚实参与方 P_{target} 上传给服务器的嵌入表示 $\boldsymbol{x}_{\text{target}}$。假设预测样本 $x=(x_{\text{adv}},\boldsymbol{x}_{\text{target}})$，其中 x_{adv} 和 $\boldsymbol{x}_{\text{target}}$ 分别代表 P_{adv} 和 P_{target} 中所持有的发送给服务器的嵌入表示。此外，$\boldsymbol{D}_{\text{adv}}$ 和 $\boldsymbol{D}_{\text{pred}}$ 分别代表 P_{adv} 和 P_{target} 所持有的嵌入表示数据集合。

攻击者发动重构攻击的背景知识包含纵向图联邦学习参数 $\boldsymbol{\theta}$、预测输出 $\boldsymbol{v}$ 和自身的特征数据 $\boldsymbol{x}_{\text{adv}}$。在正常的纵向图联邦学习协议中，参与方已知纵向图联邦学习模型参数 $\boldsymbol{\theta}$，服务器公布主任务预测的置信度向量 $\boldsymbol{v}=(v_1,\cdots,v_c)$ 以及攻击者自身参与纵向图联邦学习的数据。

综上所述，攻击者 P_{adv} 的攻击目标为推断诚实的参与方 P_{target} 上传给服务器的嵌入表示 $\boldsymbol{x}_{\text{target}}$，即：

$$\hat{\boldsymbol{x}}_{\text{target}}=\boldsymbol{G}_A(\boldsymbol{x}_{\text{adv}},\boldsymbol{v},\boldsymbol{\theta}) \tag{4}$$

其中，$\hat{\boldsymbol{x}}_{\text{target}}$ 是攻击者推断出的嵌入表示，$\boldsymbol{G}_A$ 是由 P_{adv} 所执行的攻击模型。

4.2 攻击方法

本文采用的攻击方法是基于生成式网络(Generative Network，GN)进行的嵌入表示重构攻击。如前所述，攻击者 P_{adv} 在正常的纵向图联邦学习协议下具有以下背景知识：纵向图联邦学习参数 $\boldsymbol{\theta}$、预测输出 $\boldsymbol{v}$ 和攻击者的嵌入表示 $\boldsymbol{x}_{\text{adv}}$。

假设攻击方 P_{adv} 已经得到了 D_{pred} 中 n 个样本的预测输出，其中 $D_{\text{pred}}=(D_{\text{adv}},D_{\text{target}})$，使得第 t 个样本 $\boldsymbol{x}^t=(\boldsymbol{x}_{\text{adv}}^t,\boldsymbol{x}_{\text{target}}^t)$，$t\in\{1,\cdots,n\}$。

设 $\boldsymbol{V}=(v^1,\cdots,v^n)$ 是与之对应的 n 个预测输出。攻击方的目标是训练生成式网络参数 $\boldsymbol{\theta}_G$，使得在已知第 t 个样本 $\boldsymbol{x}_{\text{adv}}^t$ 嵌入表示的情况下，生成器网络可以输出被攻击者 $\boldsymbol{x}_{\text{target}}^t$ 的重构嵌入表示 $\hat{\boldsymbol{x}}_{\text{target}}^t$。$P_{\text{adv}}$ 采用小批量随机梯度下降法训练生成式网络模型，其中训练集由 D_{adv} 和 $\boldsymbol{V}$ 组成。目标函数如下：

$$\min_{\theta_G}\frac{1}{n}\sum_{t=1}^{n}l(f(\boldsymbol{x}_{\text{adv}}^t,f_G(\boldsymbol{x}_{\text{adv}}^t,r^T;\boldsymbol{\theta}_G);\boldsymbol{\theta}),\boldsymbol{v}^t)+\Omega f(f_G) \tag{5}$$

其中，$\boldsymbol{\theta}$ 和 $\boldsymbol{\theta}_G$ 分别是纵向图联邦模型和生成器模型的参数；f_G 是生成器模型的输出 $\hat{x}_{\text{target}}^t$；$f$ 是在指定生成样本(由 $\boldsymbol{x}_{\text{adv}}^t$ 和 $\hat{\boldsymbol{x}}_{\text{target}}^t$ 组成)的情况下，纵向图联邦模型的预测输出；$\Omega(\cdot)$ 是生成的未知特征值 $\{\hat{\boldsymbol{x}}_{\text{target}}^t\}_{t=1}^n$ 的正则化项，其目的是惩罚生成器模型参数。

算法 3 为重构攻击中生成式网络的训练方法。在每次生成式网络的训练迭代中，P_{adv} 首先从 D_{adv} 中选择一批训练样本 $(x_{\text{adv}}^1,\cdots,x_{\text{adv}}^{|B|})$，同时初始化一组数量为 $|B|$ 的随机向量 $\boldsymbol{R}=(r^1,\cdots,r^{|B|})$。单个随机向量的大小是攻击目标 P_{target} 持有的嵌入表示。对于训练样本 B 中的第 t 个样本，P_{adv} 将 $\boldsymbol{x}_{\text{adv}}^t$ 和

相应的$\boldsymbol{r}^T$输入生成器中获得目标预测嵌入表示$\hat{\boldsymbol{x}}^t_{\text{target}}$。

P_{adv}将$\boldsymbol{x}^t_{\text{adv}}$与$\hat{\boldsymbol{x}}^t_{\text{target}}$组合以获得完整的生成样本，并使用纵向图联邦学习模型进行预测，从而产生预测输出$\hat{v}^t$。最终可以利用真实预测输出v^t，通过损失函数$l(\cdot,\cdot)$，如均方误差(Mean Square Error，MSE)来计算该生成样本的损失。在获得训练样本B中所有样本的损失后，攻击方反向传播聚合损失来更新生成器模型$\boldsymbol{\theta}_G$的参数，最终在所有训练周期之后获得训练好的生成器模型。

当攻击者完成对生成器模型$\boldsymbol{\theta}_G$的训练后，可以利用$\boldsymbol{\theta}_G$推断$\boldsymbol{D}_{\text{pred}}$中的未知嵌入表示$\boldsymbol{D}_{\text{target}}$。具体而言，对于$\boldsymbol{D}_{\text{pred}}$中的任何样本$x$，攻击方生成随机向量$\boldsymbol{r}$，并直接计算$f_G(\boldsymbol{x}_{\text{adv}} \cup \boldsymbol{r}, \boldsymbol{\theta}_G)$作为推断的嵌入表示。攻击的样本是用于训练生成器模型的样本，这保证了攻击者可以积累模型预测来训练生成器，提升攻击性能。

算法 3 生成式网络的训练过程

输入：攻击方的特征值$\{\hat{\mathbf{x}}^t_{\text{target}}\}_{t=1}^n$；纵向联邦模型的参数$\boldsymbol{\theta}$；置信度$\{\mathbf{v}^t\}_{t=1}^n$；学习率$\alpha$

输出：生成器模型参数$\boldsymbol{\theta}_G^*$

1. $\boldsymbol{\theta}_G \leftarrow N(0,1)$ //初始化生成器模型参数
2. For 每个周期 do
   2.1. 批次中的每个样本输入
   2.2. Set loss=0
   2.3. 随机抽取一组样本 B
   2.4. $R \leftarrow N(0,1)$ //初始化随机值
   2.5. For $t \in \{1,\cdots,|B|\}$ do
      2.5.1. $\hat{\mathbf{x}}^t_{\text{target}} \leftarrow f_G(\mathbf{x}^t_{\text{adv}} \cup \mathbf{r}^t, \boldsymbol{\theta}_G)$
      2.5.2. $\hat{\mathbf{v}}^t \leftarrow f(\mathbf{x}^t_{\text{adv}} \cup \hat{\mathbf{x}}^t_{\text{target}}; \boldsymbol{\theta})$
      2.5.3. loss+=$l(\hat{\mathbf{x}}^t, \mathbf{x}^t)$
      2.5.4. $\boldsymbol{\theta}_G \leftarrow \boldsymbol{\theta}_G - \alpha \cdot \nabla \boldsymbol{\theta}_G^{\text{loss}}$
3. Return $\boldsymbol{\theta}_G$

5 实验

本节首先介绍本实验的数据集、模型、实验环境，以及评价指标，然后给出本文提出的攻击方法的实验结果，最后展示常见的隐私保护方法对攻击的防御效果。

5.1 实验设置

5.1.1 实验平台

实验的操作系统为 Linux，集成开发环境为 PyCharm 和 Python 3.7.0，运行环境的 CPU 为 i7-7700K 3.5 GHz，GPU 为 TITAN-Xp-12 GB，内存为 16 GB(DDR4)，开发及实验的操作系统为 Ubuntu 16.04(OS)。

5.1.2 数据集

本小节实验总共采用了 3 种被广泛使用的引文图数据集[1)]，包括 Cora，Citeseer 和 Pubmed。Cora 数据集由 2708 篇科学出版物组成，总共分为 7 个类别：基于案例、遗传算法、神经网络、概率方法、强化学习、规则学习，以及理论。Citeseer 数据集共包含 3312 篇科学出版物，分为 6 个类别。Citeseer 数据集由 4732 条连边(出版物之间构建的联系)组成，其中每个出版物都由一个 0 或 1 值的词向量表示字典中相应关键词的缺失/存在。Pubmed 数据集是关于糖尿病的 19717 个科学出版物，共包含 44338 条出版物之间的引用关系。数据集中的每个出版物都由字典中的加权词向量来描述，该字典由 500 个唯一词组成。表 1 列出了这些数据集的详细统计数据。

表 1 数据集统计指标

Table 1 Dataset statistics

数据集名称	节点数量	连边数量	特征数量	标签数量
Cora	2708	5429	1433	7
Citeseer	3327	4732	3703	6
Pubmed	19717	44338	500	3

5.1.3 数据集划分

针对纵向图联邦学习数据的分布特点，本文将数据集按照参与方数量进行随机均匀划分。实验中每个参与方之间具有不同的特征(不同的特征矩阵)和相同的交互关系(相同的邻接矩阵)。针对每个数据集，我们将数据集中 80%的网络节点作为训练节点，将 20%的网络节点作为测试节点，网络中的节点作为对应数据集的标签信息。本文的攻击背景知识和文献[11]保持相同设置。假设攻击者已知的节点数量占所有网络节点数量的 20%，攻击测试的节点数量占所有网络节点数量的 80%。实验结果默认给出 10 次重复随机实验结果的均值，其中每次重复实验的节点按比例保持随机划分。

5.1.4 纵向图联邦学习设置

纵向图联邦学习中包含 2 个参与方(一个诚实且好奇的参与方以及一个诚实参与方)和 1 个服务器。参与方的本地模型采用 2 层 GCN 模型[2)]，其中隐藏层的维度设置为 16，激活函数采用 ReLU 非线性函数。图卷积神经网络的随机丢弃率设置为 0.5。服务器的顶部模型具有 2 层全连接层，顶部模型最后一层输出层经过 Softmax 函数输出对主任务的预测结果。纵向图联邦学习中的损失函数设定为均方误差损失函数，其优化器采用 Adam，训练周期预设为 400。

5.1.5 评价标准

本文使用均方误差作为隐私泄露情况的主要评估指标，其计算式如下：

$$MSE=\frac{1}{n \cdot d_{\text{target}}} \sum_{t=1}^{n} \sum_{i=1}^{d_{\text{target}}} (\hat{\boldsymbol{x}}^t_{\text{target},i} - \boldsymbol{x}^t_{\text{target}})^2 \tag{6}$$

其中，n是预测数据的样本数，d_{target}是目标嵌入表示的节点数量，$\hat{\boldsymbol{x}}^t_{\text{target}}$和$\boldsymbol{x}^t_{\text{target}}$分别是第 t 个样本的推断嵌入表示和真实嵌入表示。

本文选择准确率作为节点分类主任务的评价指标。准确率(Accuracy)是常见的评价主任务性能的技术指标，其表示预测正确的样本占总样本的比例。准确率取值范围为[0,1]，取值越大，纵向图联邦学习预测能力越好。

$$Accuray=\frac{TP+TN}{P+N} \tag{7}$$

1) https://linqs.soe.ucsc.edu/data

2) https://github.com/tkipf/gcn

其中，TP 表示实际为正样本且被预测为正的样本数，TN 表示实际为负样本且被预测为负的样本数，P 为正样本数，N 为负样本数。

5.2 实验结果与分析

本节对本文提出的嵌入表示重构攻击方法展开实验验证，首先展示了在 3 个数据集上的嵌入表示重构攻击方法达到的攻击效果，然后测试了重构攻击对纵向图联邦学习节点分类任务的影响，最后评估了在一些先进的隐私保护机制下重构攻击的效果。

5.2.1 嵌入表示数据的重构攻击效果

首先在 3 个现实网络数据集上测试了基于生成式网络的嵌入表示重构攻击的攻击效果。如前文所述，重构攻击的均方误差值(MSE)越小表示攻击效果越好。表 2 列出了重构攻击在 3 个不同数据集上生成式网络训练不同周期情况下的攻击效果，结果表明本文提出的重构攻击取得了较好的攻击效果。例如，在 Cora 数据集上，攻击者利用生成式网络能够将嵌入表示的重构均方误差损失降低到 0.359，在 Citeseer 数据集上的重构均方误差损失值降低到 0.86 以下。上述结果表明现有的纵向图联邦学习往往能够被攻击者重构，因此存在参与方的嵌入表示数据泄露的风险。

表 2 不同周期下的攻击均方差误差数据

Table 2 Attack mean square error in different periods

数据集	200	500	1000	2000	3000	4000
Cora	1.280	1.260	0.710	0.350	1.120	0.990
Citeseer	0.850	0.850	0.940	0.680	0.365	0.200
Pubmed	1.520	1.520	1.520	1.550	1.537	1.560

图 2 的折线图展示了攻击者的生成式网络的训练周期对重构攻击的影响。我们发现攻击者的生成式网络的训练周期对攻击具有一定影响。

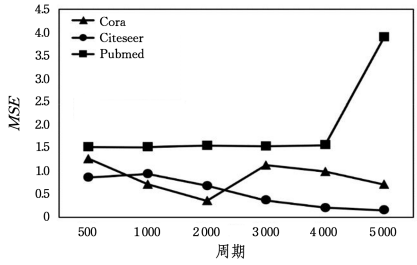


图 2 不同周期下的窃取攻击效果

Fig. 2 Steal attack effect in different periods

例如，在 Cora 数据集上的实验表明，攻击者生成式网络的训练周期达到 2000 次时，达到了最佳的攻击效果，这意味攻击模型的训练较为充分；随着周期的逐渐增加，攻击效果出现下降趋势，这可能是攻击模型出现过拟合现象导致的。

由于不同数据集的规模大小不同，攻击者的重构效果具有差异，其中针对规模较大的数据集进行攻击往往具有更大的重构误差。例如，针对 Pubmed 数据集的攻击效果相比 Cora 数据集的重构均方误差损失值提高了 0.57 左右，前者的节点数量以及连边数量相较于后者分别增加了 17009 和 38909。产生这种现象的原因是数据集的特征规模越大，纵向图联邦学习本地模型提取的嵌入表示越丰富，攻击者利用生成式网络进行重构攻击的难度越大。

5.2.2 攻击后的主任务性能

本文提出的嵌入表示重构攻击发生在参与方本地，攻击者在整个攻击过程中都遵守正常的纵向图联邦学习训练协议，因此攻击者发动攻击后的纵向图联邦学习在训练过程中的节点分类准确率表现正常。图 3 展示了在 3 个数据集上攻击者发动重构攻击后对纵向图联邦学习节点分类任务上的影响。结果如前文所述，3 个数据集上的实验结果表明攻击者发动重构攻击后的节点分类测试准确率提升，并且达到了较高的水平。例如，在 Pubmed 数据集上纵向图联邦学习在节点分类任务上的测试准确率达到 72%左右。由于攻击仅仅在攻击者本地进行，其他正常参与方难以检测到此类攻击，因此攻击具有较高的隐蔽性。

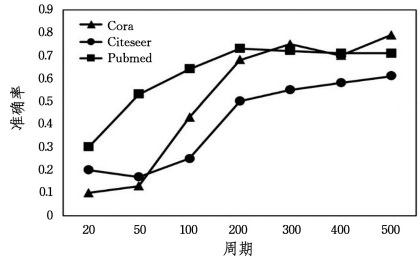


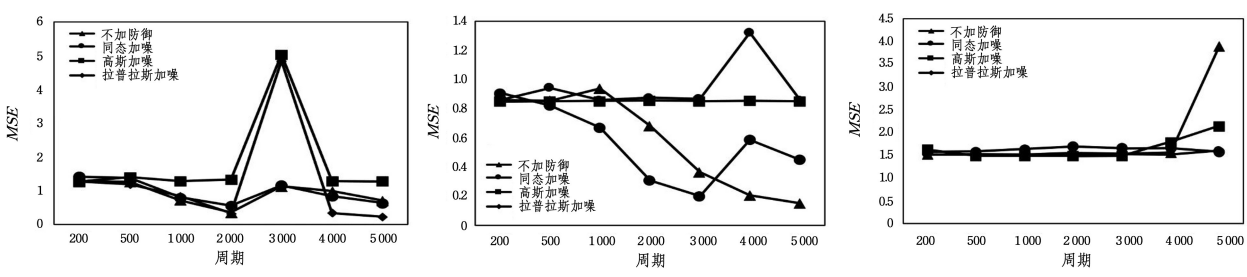
图 3 不同数据集的 VFL-GNN 准确率

Fig. 3 Accuracy of VFL-GNN in different datasets

5.2.3 隐私保护的 VFL-GNN

本小节评估纵向图联邦学习的隐私保护策略对重构攻击的影响。如前所述，基于差分隐私的隐私保护的纵向图联邦学习根据添加噪声的类型可以分为 3 种方式：同态加噪、拉普拉斯加噪和高斯加噪。

图 4 给出了 3 个数据集上纵向图联邦学习采取不同防御策略对攻击的影响。



(a)Cora (b)Citeseer (c)Pubmed

图 4 不同差分隐私方法对攻击的防御效果

Fig. 4 Defense effect of different differential privacy methods against attacks

表3列出了不同防御策略下的VFL-GNN在不同数据集上的测试准确率。观察图3可知，具有隐私保护的VFL-GNN相较于没有任何防御方法的VFL-GNN在一定程度上提升了攻击的重构均方误差值。例如，在Cora数据集上进行高斯加噪的防御策略使得攻击的重构损失误差MSE指标平均达到1.2左右。然而现有的这些隐私保护方法对VFL-GNN完成节点分类任务的准确率存在较大影响。例如，在Cora数据集上采用高斯加噪的方式使得准确率降低了9.8%。因此，现有的防御方法存在以下不足：1)难以同时权衡隐私保护和主任务，使得防御方法不实用；2)现实中难以针对不同的数据集确定最佳的噪声添加方式。

表3 不同防御方法下VFL-GNN在不同数据集上的测试准确率

Table 3 VFL-GNN test set accuracy with different defense methods

(单位：%)

数据集	不加防御	同态噪声	高斯噪声	拉普拉斯噪声
Cora	70.90	71.70	61.10	71.60
Citeseer	59.00	58.90	51.00	59.10
Pubmed	74.80	74.20	67.80	73.40

结束语 随着纵向图联邦学习被广泛地应用于各种实际场景中，其性能和效率也受到了广泛关注。本文从隐私安全的角度，提出了一种基于生成式网络的嵌入表示数据重构攻击方法，用于深入评估现有的纵向图联邦学习算法的隐私性。一方面，本文提出的重构攻击取得了较好的嵌入表示重构效果；另一方面，攻击者的攻击过程对主任务的性能影响较小，具有较强的隐蔽性。此外，本文评估了纵向图联邦学习采用一些现有的隐私保护方法后对攻击的影响，然而这些隐私保护方法由于难以权衡隐私保护和主任务而无法实现有效保护。

参 考 文 献

[1] FAN W Q, MA Y, LI Q, et al. Graph neural networks for social recommendation[C]// The World Wide Web Conference. ACM, 2019: 417-426.

[2] WANG X Y, MA Y, WANG Y Q, et al. Traffic flow prediction via spatial temporal graph neural network[C]// The World Wide Web Conference. ACM, 2020: 1082-1092.

[3] XIAO C, XU L L. Loosely Coupled Graph Convolutional Neural Network for Text Classification[J]. Journal of Chinese Computer Systems, 2021, 42(3): 449-453.

[4] VOIGT P, AXEL VON DEM B. The EU General Data Protection Regulation(Gdpr)[M]. Cham: Springer International Publishing, 2017.

[5] Data Security Law of the People's Republic of China [J]. Bulletin of the Standing Committee of the National People's Congress of the People's Republic of China, 2021(5): 951-956.

[6] YANG Q, LIU Y, CHENG Y, et al. Federated learning[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2019, 13(3): 1-207.

[7] CHEN C C, ZHOU J, ZHENG L F, et al. Vertically Federated Graph Neural Network for Privacy-Preserving Node Classification[C]// Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. 2022: 1959-1965.

[8] ZHOU H K, HUA B. Homomorphic Encryption Offloading and Its Application in Privacy-preserving Computing[J]. Journal of Chinese Computer Systems, 2021, 42(3): 595-600.

[9] NI X, XU X L, LYU L J, et al. A Vertical Federated Learning Framework for Graph Convolutional Network[J]. arXiv: 2106.11593, 2021.

[10] HE C Y, KESHAV B, EMIR C, et al. Fedgraphnn: A federated learning system and benchmark for graph neural networks[J]. arXiv: 2104.07145, 2021.

[11] DUDDU V, BOUTET A, SHEJWALKAR V. Quantifying Privacy Leakage in Graph Embedding[C]// MobiQuitous'20: Computing, Networking and Services. ACM, 2020: 76-85.

[12] ZHANG Z K, CHEN M, MICHAEL B, et al. Inference Attacks Against Graph Neural Networks[C]// Proceedings of the 31th USENIX Security Symposium. USENIX, 2022: 1-18.

[13] WANG Y, SUN L. Membership inference attacks on knowledge graphs[J]. arXiv: 2104.08273, 2021.

[14] WU B, YANG X W, PAN S, et al. Adapting membership inference attacks to gnn for graph classification: Approaches and implications[C]// IEEE International Conference on Data Mining. IEEE, 2021: 1421-1426.

[15] LIAO P Y, ZHAO H, XU K, et al. Informationobfuscation of graph neural networks[C]// Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021: 6600-6610.

[16] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]// Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR, 2017: 1273-1282.

[17] WU C H, WU F Z, CAO Y, et al. Fedgnn: Federated graph neural network for privacy-preserving recommendation[J]. arXiv: 2102.04925, 2021.

[18] LI O, SUN J K, YANG X, et al. Label leakage and protection in two-party split learning[J]. arXiv: 2102.08504, 2021.

[19] FU C, ZHANG X H, JI S L, et al. Label inference attacks against vertical federated learning[C]// 31st USENIX Security Symposium. USENIX, 2022: 1-18.

[20] WENG H Q, ZHANG J, XUE F, et al. Privacy leakage of real-world vertical federated learning[J]. arXiv: 2011.09290, 2020.

[21] JIANG X, ZHOU X B, GROSSKLAGS J. Comprehensive analysis of privacy leakage in vertical federated learning during prediction[J]. Proceedings of Privacy Enhancing Technologies, 2022(2): 263-281.

[22] JIN X, CHEN P Y, HSU C Y, et al. CAFE: Catastrophic data leakage in vertical federated learning[C]// Advances in Neural Information Processing Systems. NeurIPS, 2021: 994-1006.

[23] KIPF T N, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[C]// 5th International Conference on Learning Representations. 2017: 1-14.

[24] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016: 308-318.

[25] DWORK C, ROTH A. The Algorithmic Foundations of Differential Privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014: 9(3/4): 211-407.

LI Rongchang, born in 1998, postgraduate. His main research interests include federated learning and graph neural network.

CHEN Jinyin, born in 1982, Ph.D, professor. Her main research interests include artificial intelligence security, data mining and intelligent computing.

(责任编辑：何杨)