基于差分隐私的数据世系发布方法

倪巍伟 沈 涛 闫 冬
(东南大学计算机科学与工程学院 南京 211189)
(东南大学计算机网络和信息集成教育部重点实验室 南京 211189)

摘 要 数据世系描述数据产生、演化的机理和流程,对数据质量评估、数据恢复、数据分析有重要意义.伴随着数据共享的日益深化,对数据世系的主要表现结构世系工作流进行共享的需求也日益迫切.世系工作流中包含的节点模块,以及节点间的时序关系可能涉及数据所有者的隐私,对其进行共享不可避免地会带来隐私保护问题.已有研究侧重世系工作流局部映射关系的维持,对世系工作流可用性的重要表现——工作流时序约束关系维持效果较弱;也缺少对工作流相邻节点有向度分布隐私的保护.针对上述问题,引入输入/输出度序列(Input and Output Degree Sequence with Scale i,IO-iD)模型,在描述世系工作流节点度分布的同时,兼顾对工作流方向特性的提取;提出 Previous-Next 时序序列结构,描述工作流中节点与其邻接节点的子结构特征;在此基础上,提出基于差分隐私的隐私保护世系工作流发布算法 DpriPP,实现弱背景知识依赖的隐私保护世系工作流发布与工作流时序依赖关系可用性的有效维持.理论分析和实验结果表明,所提算法在保护世系工作流局部相邻节点有向度分布隐私的同时,能有效维持世系工作流节点局部与整体时序依赖关系的可用性.

关键词 隐私保护;世系工作流;差分隐私;IO-iD 序列模型;Previous-Next 时序序列
中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2020.00573

Differential Privacy Based on Data Provenance Publishing Method

NI Wei-Wei SHEN Tao YAN Dong
(*Department of Computer Science and Engineering, Southeast University, Nanjing* 211189)
(*Key Laboratory of Computer Network and Information Integration in Southeast University, Ministry of Education, Nanjing* 211189)

Abstract Data provenance describes the mechanism and process of data generation and evolution, which records information about the node module executions used to produce concrete data items, as well as those intermediate data items acting as parameters passed between nodes' executions. Data provenance plays an important role in research and applications of data quality assessment, data recovery and data analysis. With increasingly deepening of data sharing, the need for publishing and sharing workflow of provenance, which is the main representation structure of data provenance, becomes increasingly urgent. However, the provenance workflow often contains private or confidential data. For instance, the node modules included in the provenance workflow, as well as the temporal relations among those nodes, may involve the privacy of the data owner. Direct release of provenance workflow will inevitably bring privacy protection issues. Privacy-preserving data provenance publication becomes an urgent problem to be solved. That is to say, the utility of published provenance workflow needs well maintaining under the premise that data privacy should not be disclosed. The existing research mainly focuses on the maintenance of the local mapping relationship of provenance workflow. For example, privacy protection process is implemented to the provenance workflow, which ensures that the degree of a single node would not be leaked, or

收稿日期:2018-12-07;在线收稿日期:2019-09-06.本课题得到国家自然科学基金(61772131,61370077)资助.**倪巍伟**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为数据挖掘、数据隐私安全保护.E-mail:wni@seu.edu.cn.**沈 涛**,硕士研究生,主要研究方向为数据隐私安全保护.**闫 冬**,硕士研究生,主要研究方向为数据隐私安全保护.

sensitive mapping relationship would not be leaked in parallel with effective maintaining of provenance workflow's overall input and output mapping relations. For another important manifestation of provenance workflow's utility, i. e. temporal dependence among those front and corresponding back task nodes of the provenance workflow, the maintenance effect is relatively poor. As for the privacy of adjacent nodes distribution in the provenance workflow, the protection ability of existing methods is also far insufficient. To solve the above problems, the definition of input and output degree sequence with scale i model is introduced to describe the degree distribution of provenance workflow nodes. It provides a carrier for describing the utility and privacy-sensitive information of provenance workflow. As a by-product, it can also accommodate the extraction of directional characteristics of the provenance workflow well. The definition of Previous-Next sequence is further devised to describe the substructure distribution characteristics of the workflow. This structure can reduce the possible loss of workflow's temporal relations in adding differential noise. By constructing schema of Previous-Next sequence, the substructure characteristics of those nodes and their adjacent nodes in workflow are captured, and the temporal constraints in workflow are maintained during the reconstruction process. On this basis, a differential privacy-based on privacy-preserving provenance workflow publishing method DpriPP is proposed to implement a weak background knowledge-dependent privacy-preserving provenance workflow publication, it can also provide well maintenance to temporal dependance relations of the provenance workflow. Targeted experiments are designed to verify the effectiveness and privacy protection effect of our proposal. The theoretical analysis and experimental results demonstrate that the proposed algorithm can effectively maintain both the local and global temporal dependence relations of nodes in the workflow, while protecting the directional distribution privacy of the locally adjacent nodes in the provenance workflow well.

Keywords privacy protection; provenance workflow; differential privacy; Input and Output Degree sequence with scale i; Previous-Next sequence

1 引 言

数据世系[1-2]用于描述数据产生、演化流程(工作流)和数据源信息,主要以有向工作流图(Directed Acyclic Graph,DAG)形式展现.数据世系在数据质量评估、数据恢复、信息安全领域发挥着日益重要的作用.随着人们对数据质量、溯源要求的日益提高,世系共享需求愈加迫切,而世系中包含的数据的生成演化机理,一些敏感处理逻辑的公开,可能导致数据所有者隐私的泄露,这也带来了数据世系共享发布中的隐私安全保护问题.世系安全研究主要分为两类:世系安全访问(Secure access for data provenance)和世系隐藏发布(Privacy in data provenance),前者侧重世系不被破坏,主要采用加密和安全计算技术,处理后世系为密文无法直接使用;后者强调兼顾隐私保护的世系可用性最大化和隐藏后的直接可用.近年来,世系安全访问得到了研究者的持续关注,提出了一系列保护方法[3-6],而世系隐藏发布研究仍处于起步阶段.

在隐私保护世系发布方面,结合世系工作流图的层次结构特点,Davidson 等人[7-8]将世系隐私保护分为数据隐私保护、模块隐私保护和结构隐私保护三个层面.目前主要研究工作集中于模块和结构世系工作流隐私保护.在模块隐私保护方面,提出Γ-隐私模型,采用泛化思想,对世系工作流输入/输出映射关系进行隐匿,以保证模块的特定映射关系被逆推的概率小于$1/\Gamma$,该方法对输入/输出数据集的限制较大,仅适用于有限值域情况.在结构隐私保护方面,主要采用泛化思想对特定结构进行隐匿,实现多模块节点间结构信息保护,其缺陷是删除边和节点可能造成世系可用性的损失过多.

已有方法主要存在世系工作流可用性局限于工作流局部映射特征,以及世系隐私保护有效性对背景知识的约束较强的不足.近几年,差分隐私以其严格的数学基础和对背景知识的弱依赖,在数据隐私

保护领域得到了长足的关注.目前,在社交网络隐私保护发布方面,文献[9-10]提出基于 dK-序列的社交网络隐匿方法,通过提取社交网络节点度信息构建 dK-序列,再对节点度序列进行差分隐匿,重构满足差分隐私的隐匿图发布,实现保护敏感节点度信息和社交网络基于节点度的可用性维持.然而世系工作流有别于传统社交网络图结构,世系工作流采用有向图结构,且整体有输入和输出,对世系工作流的应用也更为注重节点模块间时序依赖关系.上述基于 dK-序列的社交网隐匿方法直接应用于世系工作流隐匿,主要存在以下问题:

(1) dK-序列适用于无向图结构,缺乏对世系工作流可用性所依赖节点流向特征的兼顾;在隐私保护方面,局限于对单个节点度分布信息的保护,缺少对局部节点对(即有向边)度特征泄露可能导致隐私泄露的关注.

如图 1 所示,即便隐藏后工作流满足度为 5 的节点至少为 3 个,若攻击者获知满足邻接节点的入度为 3、出度为 2 的节点对只有一组,仍将导致节点 A、B 及其邻接边信息的泄露.

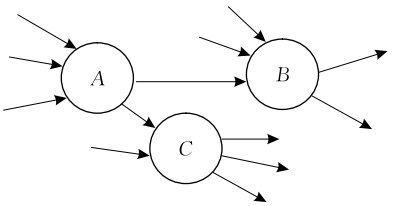


图 1　工作流局部示例

(2)已有方法难以支持世系可用性的重要表现——工作流整体时序依赖关系的维持.若要维持图结构局部或整体时序,需要提取并维护高层次 dK 序列,随着 dK 序列层次增加,差分隐私敏感度亦增大,造成数据可用性显著缺失.

已有方法侧重单个节点基于度的可用性维持,缺少对节点对方向及其出入度分布这一世系工作流时序可用性重要组成结构的维持.

针对上述问题,引入输入/输出度序列(Input and Output Degree Sequence with Scale i, IO-iD)模型,在描述世系工作流节点度分布的同时,兼顾对工作流方向特性的提取;通过构建 Previous-Next 时序序列,维护工作流中节点与其前后邻接节点的子结构特征,兼顾工作流重构中时序约束的维持;在此基础上,提出基于差分隐私的世系工作流隐匿发布算法 DpriPP(Differential Privacy based Data Provenance Publishing),实现弱背景依赖的隐私保护世系工作流发布.

论文主要贡献包括:

(1)提出输入/输出度序列模型(IO-iD),基于 IO-2D 模型扩展构建 IO-2D 边序列模型,作为描述世系工作流可用性与隐私敏感信息的载体;

(2)提出基于差分的隐私保护世系工作流发布方法,通过对 IO-2D 边序列添加差分噪声,并进行工作流重构,避免基于节点对有向度的隐私泄露;

(3)通过引入 Previous-Next 时序序列减少工作流时序关系在添加差分噪声中可能受到的损失,实现工作流重构中时序特征的高效维持;

(4)设计实验验证所提方法的有效性和隐私保护效果.

本文第 2 节对隐私保护数据世系发布领域相关工作进行概述;第 3 节描述问题并介绍相关定义;第 4 节介绍基于差分的隐私保护世系工作流发布方法 DpriPP,并进行算法分析;第 5 节对 DpriPP 算法进行实验分析;最后,总结全文.

2　相关工作

近年来,世系工作流的隐私保护得到了研究者的持续关注,提出了一系列保护方法,本节对隐私保护世系工作流发布相关技术进行概述.

2.1　隐私保护世系工作流发布

在隐私保护世系发布方面,已有研究侧重于工作流整体与局部敏感映射关系的保护. Davidson 等人[8]提出了 Γ-隐私模型,要求对工作流中给定模块的任意输入 I,至少存在一个真实值 $O=f(I)$ 和 $\Gamma-1$ 个不同于 O 的隐匿输出值,使得攻击者逆推猜测出正确输出 O 的概率不大于 $1/\Gamma$. 该模型采用 k 匿名思想,解决独立敏感模块隐私保护问题.考虑噪声数据对世系工作流可用性的影响,算法采用限制发布数据策略,存在对节点输入/输出数据集限制较强约束,同时单纯依赖限制发布也导致世系工作流隐私保护强度的不足.

基于结构隐藏的目标是保持节点间敏感连接关系安全,即保护特定节点间是否有连接及其边的权值.主要采用删除边或节点以及将多个节点组合为复合节点的方法[7],保护敏感节点及路径的隐私不被泄露,但存在破坏原世系工作流可用性的不足.

2.2　差分保护技术

差分隐私是 Dwork[11]在 2006 年针对统计数据库隐私泄露提出的一种隐私保护方法.通过向目标

数据集添加特定噪声保证任意个体记录是否在数据集,对发布数据集的查询结果近乎无影响.

目前,差分技术已从关系型数据聚集查询应用扩展到诸如社交网络等图结构数据隐私保护[12-13].已有研究主要关注差分隐私应用于图形式的边权值、度与结构信息等统计特征的发布,利用差分隐私机制以直方图等形式进行发布.文献[14]提出MB-CI(Merging Barrels and Consistency Inference)方法,保护社交网络图的边权重信息,通过将边权重序列看作无属性的直方图,利用直方图实现差分隐私机制.MB为合并桶操作,为将相同计数的桶合并以减少敏感度;CI为一致性推断,通过序列的原始顺序进行一致性的约束,从而使得发布信息满足差分隐私.这种方式采取合并操作可能会带来新的隐私泄露问题.文献[15-16]使用投影技术解决节点差分隐私下图数据带来的高敏感度问题,通过给定度数阈值,将输入图投影到最大度低于阈值的图像集上,并使其满足节点差分隐私条件.上述工作可归结为对图的特定统计值进行差分机制的扰动和发布,未对图本身进行修改.针对数据生成过程或者工作流本身使用过程可能造成的隐私泄露问题,文献[17]提出量化每个敏感数据源相对业务流程中涉及的隐私泄露量度量方法,通过敏感度和差分隐私广义组合规则,实现链式业务工作流的高适用性.文献[18]设计并实施了一种结合差分隐私和互信息来分析工作流中隐私泄露的方法,输出基于互信息的工作流隐私泄露量.文献[19]提出基于差分的隐私保护世系工作流发布方法,将世系工作流的路径作为多维空间中的一组点进行建模,并且使用空间表示方法(如:多维网格方法或 N-gram 树方法),将世系查询映射到空间域.利用差分机制对经过路径数目即映射到空间域中的计数进行保护,使得攻击者通过大量数据样本进行预测的情况下,难以确定特殊路径是否存在,从而实现世系工作流中路径查询的隐私保护.该方法的优点在于通过差分机制实现无背景知识的保护,但应用场景局限于工作流使用过程中路径查询的保护,针对特定的路径查询,保证单记录的隐私安全与统计结果的可用.

随着差分隐私应用的不断深入,研究人员开始关注图数据在满足隐私保护机制过后如何进行重构图发布的问题.文献[20]提出一种满足边-差分隐私的合成图发布方法 NoiseGraph,保证了在较低的隐私预算下,仍能实现图数据中 k-三角形计数的精确维持.文献[9-10]提出基于 dK-度模型的无向图隐私保护发布方法,利用 dK-度模型提取无向图特征,通过对特征信息进行差分机制隐匿,重构出满足差分约束的图进行发布,使得攻击者对发布后图进行隐私特征信息攻击时,无法确定该特征在原始图中存在与否.实现无关背景知识的保护,既保证图中单个节点不被攻击者推测,又兼顾图整体结构特征的维持.然而,dK 模型存在以下不足:(1)未考虑图的方向性,采用完全随机的重构方式,不适应世系工作流有向的特征;(2)dK 序列难以维持工作流整体时序结构.已有方法采用的 dK 序列层次较低,局限于对个体节点度分布的维持;若要兼顾有向图整体时序,需提升 dK 序列层次,将导致差分隐私敏感度增大,使得所添加差分噪声对数据可用性的破坏较大.

3 问题描述及相关概念

3.1 问题描述

随着数据共享的日趋迫切,数据世系共享发布中的隐私保护问题日益严峻,已有的基于限制发布保护模块隐私的方法不管在隐私保护保护强度、世系工作流的可用性,还是保护方法对背景知识的依赖方面都无法满足应用需要.在图数据隐私保护发布研究中,基于度相关的差分隐私保护方法缺少对工作流节点方向性的考虑,特征提取模型也仅关注节点模块的总体度分布,难以维持工作流可用性中重要的时序结构可用性.

针对上述问题,考虑设计基于差分的隐私保护世系工作流发布方法,充分考虑世系工作流有向性和时序特征的影响,实现不依赖背景知识的世系工作流局部节点对敏感特征保护和工作流整体时序结构约束的维持.

具体思路如图 2 所示:引入 Previous-Next 时序序列模型,描述工作流各节点的输入/输出度及其直接连接边信息,以便维护工作流时序特征;提取工作流有向结构信息,获得 IO-iD 序列,基于 IO-2D 边序列模型实现工作流图特征提取,通过添加 La-

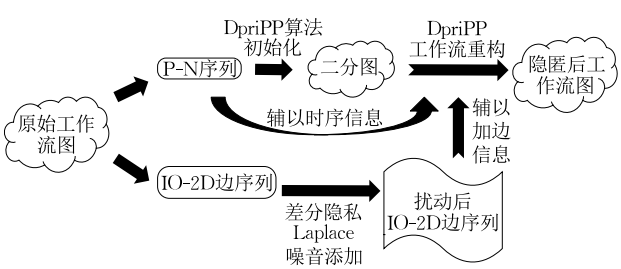


图 2 方法思路图

place 噪声，获得满足差分约束的IO-2D边序列；提出有向工作流重构算法，构建满足IO-2D边序列特征与P-N序列时序约束的隐匿后工作流图，兼顾局部节点对度特征的保护和工作流整体时序关系的维持.

3.2　相关定义

目前普遍采用DAG图描述世系工作流，图中的节点表示功能模块. 节点间的边表示模块间的数据传递关系. 节点可以视作由输入/输出及功能函数组成. 例如：节点 M_1 描述为 (i_1, o_1, f_1)，其中 i_1 为模块输入，o_1 为输出，f_1 为对应该模块的功能函数.

定义 1.　世系工作流. 一个世系工作流定义为 $G(M,E,F)$，M 表示工作流中节点模块集合，E 表示工作流中边的集合，F 表示节点模块功能函数集合. 单模块 M_i 可以描述为 (M_i^+, M_i^-, M_i^f)，其中 M_i^+ 表示模块 M_i 的输入度，M_i^- 表示模块 M_i 的输出度，M_i^f 表示 M_i 对应功能函数.

定义 2.　差分隐私[21-22]. 假设 $\mathcal{R}$ 为随机图分析算法，$Range(\mathcal{R})$ 表示 $\mathcal{R}$ 所有可能的输出构成的集合. $\mathcal{R}$ 算法以图 G 作为输入，以 $\mathcal{R}(G)$ 作为输出. 对图 G 和 $G'(G'\in\Gamma(G))$，以及满足 $S\subseteq Range(\mathcal{R})$ 的任意子集 S，若算法 $\mathcal{R}$ 满足下式：

$$\Pr[\mathcal{R}(G)\in S]\leq \exp(\varepsilon)\cdot\Pr[\mathcal{R}(G')\in S]+\delta.$$

称算法 $\mathcal{R}$ 提供 (ε,δ)-差分隐私保护，其中参数 δ 为松弛因子，若 $\delta=0$，算法 $\mathcal{R}$ 提供严格的 ε-差分隐私. 参数 ε 称为隐私保护参数，用来平衡隐私保护性和数据可用性. 参数值越小，隐私保护性越强，数据可用性越低. 其中，

$$\Gamma(G)=\{G'(V,E')\mid \exists!(u,v)\in G \text{ but } (u,v)\notin G'\}$$

表示两相邻图 G' 与 G 的节点集合相同，有且仅有一条边的差异.

定义 3.　差分隐私全局敏感度[23]. 函数 $f: D\to\mathbb{R}^d$，输入为图 G，输出为 d 维向量. 对任意满足 $d(G,G')=1$ 的 G 和 $G'(G,G'\in D)$，得出 f 的全局敏感度是：

$$GS_f(G)=\max_{G,G';d(G,G')=1}\|f(G)-f(G')\|_1,$$

其中 $d(G,G')=1$ 表示图 G 与图 G' 仅存在一条边的差异.

定义 4.　差分隐私局部敏感度[24]. 函数 $f: D\to\mathbb{R}^d$，输入为图 G，输出为 d 维向量. 对给定图 $G(G\in D)$ 和任意满足 $d(G,G')=1$ 的邻近图 $G'(G'\in D)$，则 f 对于 G 的局部敏感度是：

$$LS_f(G)=\max_{G';d(G,G')=1}\|f(G)-f(G')\|_1,$$

其中 $d(G,G')=1$ 表示图 G 与图 G' 仅存在一条边的差异.

局部敏感度由查询函数 f 以及给定图 G 计算生成. 因为在一定程度上利用了数据集的分布特征，所以局部敏感度通常比全局敏感度小得多，但也因此导致添加的噪音会揭露数据集的特征，带来数据集敏感信息泄露的风险. 为了满足严格差分隐私，Nissim 等人[24]提出 β-平滑敏感度，通过给定局部敏感度的平滑上界，来确定平滑敏感度的大小和噪音量的添加. 因此，局部敏感度通常与平滑上界共同使用，确定平滑敏感度的大小.

定义 5.　平滑上界[24-25]. 对于 $\beta>0$，若有函数 $S: D\to\mathbb{R}^+$ 满足以下两个条件：

$$\forall G\in D: S(G)\geq LS_f(G);$$

$$\forall G,G'\in D, d(G,G')=1: S(G)\leq e^{\beta}S(G').$$

则称 S 为函数 f 的局部敏感度的 β-平滑上界.

所有满足定义5的函数都可被定义为 f 的平滑上界. 一般以满足条件的最小函数 $S_{f,\beta}^*$ 作为平滑敏感度.

定义 6.　平滑敏感度[24-25]. 设有函数 $f: D\to\mathbb{R}^d$，给定输入图 $G(G\in D)$. 对 $\beta>0$，函数 f 的 β-平滑敏感度为

$$S_{f,\beta}^*(G)=\max_{G'\in D}(LS_f(G')\cdot e^{-\beta d(G,G')}),$$

其中 $d(G,G')$ 表示图 G 与 G' 间存在的差异(即存在不同边的数量).

定义 7.　计算平滑敏感度[24]. 在距离为 k 时，函数 f 的敏感度为

$$A^{(k)}(G)=\max_{G'\in D;d(G,G')\leq k}LS_f(G').$$

平滑敏感度可以用 $A^{(k)}$ 的形式来表示：

$$S_{f,\beta}^*(G)=\max_{k=0,1,\cdots,n}e^{-k\beta}\left(\max_{G';d(G,G')=k}LS_f(G')\right)$$
$$=\max_{k=0,1,\cdots,n}e^{-k\beta}A^{(k)}(G).$$

4　隐私保护工作流发布算法 DpriPP

DpriPP 算法主要包括以下步骤：提取世系工作流中节点分布以及节点间的时序特征；向特征中添加满足差分约束的噪声数据实现隐藏处理；设计图重构算法，实现基于隐匿后节点分布及时序特征的世系工作流重构，发布重构后的工作流图.

4.1　IO-iD 模型及 P-N 序列

有向无环图是世系工作流的主流描述结构，相较传统的无向图结构，DAG 图中边具有方向性，除了常见的节点间出入度分布外，基于 DAG 图的世系工作流的很多应用依赖于 DAG 图节点间整体和

局部时序关系和输入/输出特征.考虑首先建立工作流结构与时序特征提取模型.

基于上述思路,提出工作流 IO-iD 模型,描述整个工作流的概要信息.

定义 8. IO-iD 序列模型. 对工作流图 $G(M,E,F)$,其 IO-iD 序列是组合 i 个连续节点模块,进行 I/O 度统计生成的序列,其中 $i\in\{0,1,2,\cdots,|M|\}$,$|M|$ 为图 G 的模块节点数.G 的 IO-iD 序列是一组序列集合,以节点模块 M_j 为起始的一条 IO-iD 序列形式如下:

$\langle M_j^+,M_j^-;M_{j\_1}^+,M_{j\_1}^-;\cdots;M_{j\_i-1}^+,M_{j\_i-1}^-\rangle:count$,

其中 $M_{j\_l}^+$ 表示与 M_j 经过 l 条边相连的节点模块的输入度,$M_{j\_l}^-$ 表示与 M_j 经过 l 条边相连的节点模块的输出度,其中 $count$ 表示具有相同度特征的序列个数.

对工作流 G,$i=0$ 时表示工作流平均输入/输出度的分布,$i=1$ 时,表示单节点输入/输出度分布,$i=2$ 时,表示有边相连的邻接节点整体输入/输出度分布,$i=3$ 时表示三节点结构局部输入/输出度分布,依次类推.以图 3 为例,IO-iD 序列如下:

IO-0D:$\bar{d}=(1,1)$

IO-1D:$\langle 0,1\rangle:1,\langle 1,1\rangle:2,\langle 1,2\rangle:1,\langle 2,0\rangle:1$

IO-2D:$\langle 0,1;1,1\rangle:1,\langle 1,2;1,1\rangle:1,\langle 1,2;2,0\rangle:1,$
$\langle 1,1;2,0\rangle:1,\langle 1,1;1,2\rangle:1$

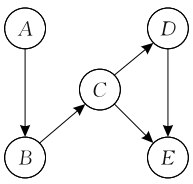


图 3 工作流图 G

其中,$M_A=(M_A^+,M_A^-)=(0,1)$,$M_B=(1,1)$,$M_c=(1,2)$,$M_D=(1,1)$,$M_E=(2,0)$.

当 $i=1$ 时,IO-1D 描述工作流中节点输入/输出度特征,例如:$\langle 1,1\rangle:2$ 表示节点输入/输出度皆为 1 的节点数为 2,即 B 和 D 模块.当 $i=2$ 时,IO-2D 描述工作流中节点对间的关系,以输入/输出度特征向量表示,例如:$\langle 1,2;1,1\rangle:1$ 表示满足第一个节点有 1 条输入边和 2 条输出边,第二个节点为 1 条输入边和 1 条输出边的局部结构模式数为 1,对应 $C\rightarrow D$ 结构概要.

IO-iD 序列模型用于获取工作流中节点或局部结构的输入/输出度,而结构规模与涉及节点数量(即序列模型的层次 i)有关.IO-1D 以单节点输入/输出关系进行工作流建模,对每个节点以(输入度,输出度)形式描述,获取工作流的 1D 形式的统计特征.IO-1D 局限于对工作流单个节点输入/输出度的提取,难以实现工作流局部节点时序关系的描述.

分析易知,IO-iD 模型的层次 i 越高,对工作流整体时序关系的提取效果越好;但模型层次越高,利用差分实现保护的难度也剧增,导致实现差分隐私所需加入噪声数据的数值幅度变大,使得发布后工作流与原始工作流在边的规模和分布上存在较大差异,丧失可用性.

结合世系工作流可用性的重要表现——节点间时序关系主要依赖相邻节点特征的维持,而 IO-2D 加入了模块间基于输入/输出度的依赖关系,相较 IO-1D 能更好地保留模块间依赖关系,因此考虑采用 IO-2D 序列模型提取节点对间的依赖关系.在此基础上,为了提升模型对工作流整体时序关系的提取效果,引入 Previous-Next 时序序列定义.

定义 9. Previous-Next 时序序列(简称 P-N 序列). 对工作流图 $G(M,E,F)$,其 Previous-Next 时序序列为对 G 进行拓扑排序,获得线性遍历序列,对序列中每个节点模块 M_j 添加度特征 $\langle M_{\rightarrow j}^+;M_j^+,M_j^-;M_{j\rightarrow}^-\rangle$,$M_{\rightarrow j}^+$ 表示 M_j 的所有直接前序节点入度之和,$M_{j\rightarrow}^-$ 表示 M_j 的所有直接后继节点出度之和.

如图 3 所示,Previous-Next 时序序列可以保护各节点(A,B,C,D,E)的前后关系,拓扑排序得到序列 $A\rightarrow B\rightarrow C\rightarrow D\rightarrow E$,其 P-N 序列表示为 $(-1;0,1;1)\rightarrow(0;1,1;2)\rightarrow(1;1,2;1)\rightarrow(1;1,1;0)\rightarrow(2;2,0;-1)$.

A 节点为输入,将其 Previous-Next 时序定义为 $(-1;0,1;1)$,-1 表示 A 节点的 previous 节点为空,1 表示 A 的 next 节点 B 的输出度为 1,$(0,1)$ 为 A 的输入/输出度描述;B 为中间节点,定义成 $(0;1,1;2)$,其中 0 表示 B 的 previous 节点 A 的输入度为 0,2 表示 B 的 next 节点 C 的输出度为 2,$(1,1)$ 为节点 B 的输入/输出度描述;节点 E 表示为 $(2;2,0;-1)$,E 的 previous 节点有输入度为 1 的 D 和输入度为 1 的 C,总和为 2;E 的 next 节点为空,表示为 -1,$(2,0)$ 表示 E 的输入/输出度.

4.2 基于差分的 IO-2D 序列隐匿方法

IO-iD 序列模型提取了工作流图中局部结构时序与度特征.如前所述,其中诸如局部节点对度特征等可能涉及工作流所有者的隐私.所以我们考虑利用差分技术对 IO-iD 序列模型中的敏感信息进行保护.

定义 10. 基于全局敏感度添加 Laplace 噪音

的差分机制[6,9]. 算法 M 以图 G 作为输入，设置参数 $\varepsilon>0$，查询 Q 满足计算函数 $f:D^n\to\mathbb{R}^d$，则其输出

$$M(G)=f(G)+(X_1,\cdots,X_d),$$

其中 X_i 满足 $\mathrm{Lap}(GS_f(G)/\varepsilon)$ 的独立同分布，称算法 M 满足 ε-差分隐私.

定义 11.　基于平滑敏感度添加 Laplace 噪音的机制[9,24]. 对一个函数 $f:D\to\mathbb{R}^d$，给定图 $G(G\in D)$ 和算法 M(算法 M 以图 G 作为输入)，通过如下机制：

$$M(G)=f(G)+\frac{S^*_{f,\beta}(G)}{\alpha}(Z_1,\cdots,Z_d)$$

可以实现 (ε,δ)-差分隐私保护($\varepsilon>0,\delta\in(0,1)$).

其中 $\alpha=\varepsilon/2,\beta=\dfrac{\varepsilon}{4(d+\ln(2/\delta))},Z_i(i=1,\cdots,d)$ 服从拉普拉斯分布 $\mathrm{Lap}(0,1)$. 当 $d=1$ 时，β 可被表示为 $\beta=\dfrac{\varepsilon}{2\ln(2/\delta)}$.

采用 IO-2D 序列模型，对工作流特征进行提取. 考虑采用边差分思想，为了降低差分隐私应用于 IO-2D 模型时的敏感度，对 IO-2D 序列模型的边结构改进，形成 IO-2D 边序列模型.

IO-2D 边序列模型. 根据图 G 中每条边进行相连节点的 I/O 度特征提取，即每次选取图 G 一条边 $(i\to j)$，记录其前驱节点 i 的输入度和后继节点 j 的输出度，以 $\langle M_i^+;M_j^-\rangle$:*count* 形式表示.

以图 3 为例，原 IO-2D 序列如下：

$\langle 0,1;1,1\rangle$:1,$\langle 1,2;1,1\rangle$:1,$\langle 1,2;2,0\rangle$:1,

$\langle 1,1;2,0\rangle$:1,$\langle 1,1;1,2\rangle$:1.

改进后的 IO-2D 边序列为

$\langle 0,1\rangle$:1,$\langle 1,1\rangle$:1,$\langle 1,0\rangle$:2,$\langle 1,2\rangle$:1.

其中，$\langle 1,2\rangle$:1 对应图 3 中 $B\to C$ 边，$\langle 1,2\rangle$ 表示 $B\to C$ 的边特征，其中 1 表示节点 B 的输入度，2 表示节点 C 的输出度. 基于有向边提取节点度特征能有效降低差分隐私的敏感度，减小需要加入的噪音数据的幅度，改善差分隐私实现工作流隐私保护的效果.

性质 1.　对图 G 采用 IO-2D 序列模型，可计算其局部敏感度为 $LS_f(G)=4d_{\max}-3$.

证明.　当从图 G 中删除任意边 $i\to j$，节点模块 i 的输出度减 1，节点模块 j 的输入度减 1，即节点模块 i 的输出度从 d_i^- 变为 d_i^--1；节点模块 j 的输入度从 d_j^+ 变化为 d_j^+-1. 上述图的转换操作在图 G 的 IO-2D 序列模型中造成了以下的改变：i 模块的前驱节点有 d_i^+ 个，用 x 表示其中的某一节点，由于 i 模块的输出度变化，原有的序列 $\langle d_x^+,d_x^-;d_i^+,d_i^-\rangle$:*count* 的 *count* 值减少 1，这样的序列一共有 d_i^+ 个；i 模块的后继节点有 d_i^- 个，同样也存在由于 i 模块输出度变化导致的原有序列 $\langle d_i^+,d_i^-;d_x^+,d_x^-\rangle$:*count* 的 *count* 值减少 1 的情况，一共有 d_i^- 个；则与 i 模块相关会导致 $d_i^++d_i^-$ 个变化. 同理与 j 模块相关也会导致 $d_j^++d_j^-$ 个变化. 然而这两组变化中，对同一个序列 $\langle d_i^+,d_i^-;d_j^+,d_j^-\rangle$ 的变化计算了两次，因此总共导致的原有序列减少变化为 $d_i^++d_i^-+d_j^++d_j^--1$，即为 d_i+d_j-1 项，d_i 表示节点 i 的输入/输出度之和. 原有序列数目发生了减少，同时会有新的序列数目增加. 考虑序列数目增加的变化，与 i 模块相关的有 d_i^+ 个前驱节点和 d_i^--1 个后继节点，一共有 $d_i^++d_i^--1$ 项变动；与 j 模块相关的有 d_j^+-1 个前驱节点和 d_j^- 个后继节点，一共有 $d_j^++d_j^--1$ 项变动. 总共导致的原有序列增加变化为 $d_i^++d_i^--1+d_j^++d_j^--1$，即为 d_i+d_j-2 项. 总结，删除一条边对 IO-2D 序列造成的总差异为 $2*(d_i+d_j)-3$. 考虑到最糟糕的情况，两个节点度数最大为 $d_{\max}$，图 G 以上述模型建模的局部敏感度为 $LS_f(G)=4d_{\max}-3$.　　证毕.

性质 2 分析 IO-2D 边序列模型的局部敏感度.

性质 2.　对图 G，IO-2D 边序列模型的局部敏感度为 $LS_f(G)=2(d^+_{\max}+d^-_{\max})+1$.

证明.　当从图 G 中删除任意一条边 $i\to j$，节点模块 i 的输出度 d_i^- 值变化为 d_i^--1，节点模块 j 的输入度 d_j^+ 变化为 d_j^+-1. 上述操作在图 G 的 IO-2D 边序列模型中造成了以下的改变：由于 i 模块的输出度产生变化，则仅有 i 模块的前驱节点与其构成的序列 $\langle d_x^+,d_i^-\rangle$:*count* 会产生变动，原有 $\langle d_x^+,d_i^-\rangle$ 的 *count* 值会减少 d_i^+ 项，并增加 $\langle d_x^+,d_i^--1\rangle$，其总共的 *count* 值会增加 d_i^+ 项；同理由于 j 模块的输入度产生变化，原有的 $\langle d_j^+,d_x^-\rangle$:*count* 序列的 *count* 值一共减少 d_j^- 项，新增 $\langle d_j^+-1,d_x^-\rangle$:*count* 序列的 *count* 值一共增加 d_j^- 项；最后考虑序列 $\langle d_i^+,d_j^-\rangle$:*count*，由于删除边 $i\to j$，导致该 *count* 值减去 1，产生 1 项变化. 因此，对图 G 删除一条边，其 IO-2D 边序列可能造成的总差异为 $2*(d_i^++d_j^-)+1$. 在最坏情况下，i 为输入度最大 $d^+_{\max}$ 的节点且 b 为输出度最大 $d^-_{\max}$ 的节点，则图 G 应用 IO-2D

边序列模型的局部敏感度为 $LS_f(G)=2(d^+_{\max}+d^-_{\max})+1$. 证毕.

IO-2D 边序列模型隐匿方法如下：

算法 1. IO-2D 边序列模型隐匿方法.

输入：图 G，隐私参数 ε,δ

输出：P-N 序列及满足(ε,δ)-差分隐私的 IO-2D 边属性表

1. 计算原始图 G 中的 IO-2D 序列
2. 改进提出 IO-2D 边序列并生成属性表
3. 计算图 G 的 Previous-Next 时序序列，即 P-N 序列
4. 利用隐私参数(ε,δ)计算(α,β) //定义 11
5. 利用 β 计算图 G 在 IO-2D 边序列模型中的 β-平滑敏感度 $S^*_{f,\beta}(G)$ //定义 7
6. 利用 $S^*_{f,\beta}(G)$，α 参数进行 Laplace 噪音机制的添加，获得满足(ε,δ)-差分隐私的 IO-2D 边序列并生成属性表 //定义 11

算法 1 描述了本方法依据原始图 G 进行 IO-2D 边序列模型的概要提取，并对提取出的原 IO-2D 边序列加入 Laplace 噪音机制扰动的过程. 为降低敏感度，采用基于平滑敏感度对 IO-2D 边序列添加 Laplace 噪音，文献[24-25]已证明采用平滑敏感度，能满足(ε,δ)-差分隐私. 即经过扰动的 IO-2D 边序列能达到对原 IO-2D 边序列的(ε,δ)-差分保护.

4.3 工作流重构

Tillman 等人[26]提出基于有向度序列的 DAG 图重构方法，通过提取 DAG 图的有向度序列构建二分图，并根据连接度属性表进行随机建边. 该方法应用于工作流图时，缺少约束的随机建边会破坏工作流的时序特性. 考虑基于上述重构思想，利用 P-N 序列对添加边操作进行约束，提出基于 P-N 序列和 IO-2D 边序列的工作流重构方法. 步骤如下：

(1) 根据 previous-next 时序序列，构建初始二分图对工作流图中节点进行描述，每个节点划分成输入和输出部分，分别记录输入度和输出度，记为该部分的存根，表示该部分可连接的边数.

二分图有助于描述和记录工作流图的有向性，文献[27]已证明任意有向无环图可以与唯一的二分图互相转换. 例如，图 4 为图 3 中工作流 G 的二分图，图中每个节点模块被划分成 in 部分和 out 部分，分别记录下模块的输入度和输出度信息.

(2) 根据隐匿后的 IO-2D 边序列，对第(1)步获得的初始二分图进行加边操作. 添加边过程兼顾 previous-next 时序进行约束，挑选连接节点，直到 IO-2D 边序列全部遍历完为止. 添加边过程可能存在挑选的两个节点 u,v，其中一个或者两个都没有空余存根的情况，此时采用相邻节点交换策略，实现对连接节点的重新挑选. 若满足约束的节点集中找不到具有空余存根的节点，则从节点集中挑选节点增加存根，实现加边.

(3) 对生成的二分图，保留剩余存根，并将其唯一转换为对应工作流图，实现工作流图的重构发布.

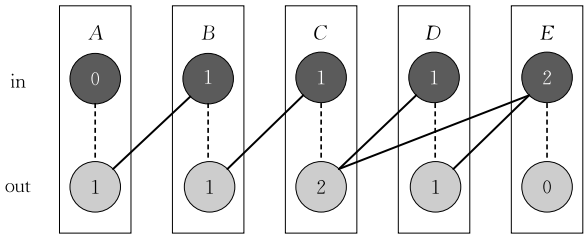


图 4 工作流 G 对应二分图

IO-2D 工作流重构算法具体的流程如下：

算法 2. IO-2D 工作流重构算法.

输入：P-N 序列，IO-2D 属性表

输出：有向工作流图

初始化：

利用 P-N 序列创建一个带有节点、分区和存根的初始二分图，把每一节点分为输入、输出节点并添加非弦边(同一节点 in 与 out 间连线)，节点按拓扑时序排列.

添加边：

1. FOR 分区中每一对($\{i,\text{in}\},\{j,\text{out}\}$)：
2. WHILE $Count(\{i,\text{in}\},\{j,\text{out}\})<Count^*(\{i,\text{in}\},\{j,\text{out}\})$
3. 挑选任意节点 u (from $V\{i,\text{in}\}$)，v (from $V\{j,\text{out}\}$)且(u,v)不是非弦边或者已存在边，且 u,v 满足 P-N 序列时序
4. IF u 没有空余存根：
5. 选择 u'：from $V\{i,\text{in}\}$，有空余存根
6. 使用 u' 与 u 进行邻居交换
7. IF 交换失败，$u=u'$
8. IF v 没有空余存根：
9. 选择 v'：from $V\{j,\text{out}\}$，有空余存根
10. 使用 v' 与 v 进行邻居交换
11. IF 交换失败，$v=v'$
12. 对于所有满足上述条件的 u,v 节点对，优先选择拓扑序中相近的添加为边(u,v)
13. 若找不到空余的 u,v 节点，对于拓扑序最近的 u,v 节点增加存根并添边(u,v)

图转化：

1. 保留二分图中剩余存根
2. 将二分图唯一转化为有向工作流图 G^*

算法 2 描述了依据算法 1 生成的满足(ε,δ)-差分隐私的 IO-2D 边属性表进行工作流图重构的过

程.本算法利用 P-N 序列实现原始图时序方向维持,并以此为基础,按照扰动后 IO-2D 边属性表,进行工作流图重构,获得重构图 G^*.对重构图 G^* 采取同样方法提取 IO-2D 边序列,该边序列与原图添加差分噪声的 IO-2D 边序列是相同的.

以图 3 为例,获取输入的 P-N 序列〈−1;0,1;1〉→〈0;1,1;2〉→〈1;1,2;1〉→〈1;1,1;0〉→〈2;2,0;−1〉,以及通过差分隐私机制处理后获得的一种 IO-2D 边序列:〈0,1〉:1,〈1,1〉:1,〈1,0〉:1,〈1,2〉:1,其转换后的输入/输出度属性表见表 1(表中行表示输入度,列对应输出度,单元格中数据表示以〈输入度;输出度〉形式的边的数目).

表 1　IO-2D 边序列的输入/输出度属性表

	0	**1**	**2**
0	0	1	0
1	1	1	1

按照 P-N 序列的存储顺序,依次获取每个节点 v 的输入/输出度信息,将其拆分成 v_{in}(表示节点 v 的输入度)和 v_{out}(表示节点 v 的输出度),并在其二分图的 v_{in} 和 v_{out} 节点上利用输入度和输出度来创建空余的存根,最后在同一节点的 in 节点和 out 节点间添加非弦边.可获得初始化二分图形式见图 5.

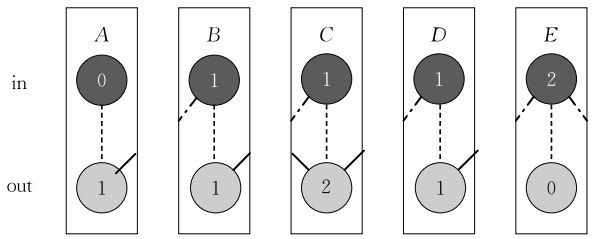


图 5　初始二分图

根据属性表,对初始化二分图进行有选择的加边操作,例如:输入为 0,输出为 1 存在一种情况,从初始化二分图中从前往后寻找符合条件的节点(A,B 节点),将其与 P-N 序列中前后节点信息比对,若添加边 $A\rightarrow B$,A 节点不存在前序节点(即其入度和为−1),节点 A 的后继节点出度和为 1;B 节点的前序节点入度和为 0,添加 $A\rightarrow B$ 满足 P-N 序列.若选择 A,D 节点,添加 $A\rightarrow D$ 边,节点 A 不存在前序节点(即其入度和为−1),节点 A 的后继节点出度和为 1;节点 D 的前序节点入度和为 0,将与原 P-N 序列中节点 D 的前序节点入度和为 1 冲突,添加 $A\rightarrow D$ 边不满足 P-N 序列,需重新寻找.

按上述过程不断迭代添加边,最终可获得图 6 所示二分图.

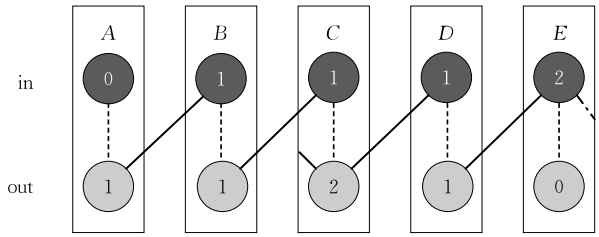


图 6　加边操作后的二分图

通过保留二分图节点中的多余存根,并依据二分图形式将其转化为图 7 所示重构工作流图 G^*,进行发布,G^* 是原工作流图 G 满足差分隐私约束的重构图.

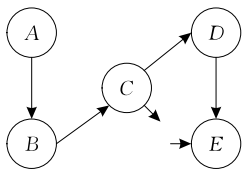


图 7　重构工作流图 G^*

5　实验分析

本节对 DpriPP 算法的隐私保护效果和数据可用性维持效果进行实验分析.选取来源于斯坦福大学 Stanford Large Network Dataset Collection 数据平台的社交网络数据集 as20000102、ego-Facebook 和 wiki-Vote 作为实验数据.由于世系工作流图表现为整体连续有向且单节点输入/输出度较小的有向无环图形式,对上述开源的社交网络数据进行处理,随机抽取其中满足工作流形式的有向无环子图用于本算法的实验验证.具体实验数据信息见表 2.

表 2　实验数据集

数据集	节点数	边数	平均入度	平均出度
as20000102	76	122	1.61	1.61
ego-Facebook	177	250	1.41	1.41
Wiki-Vote	624	799	1.28	1.28

5.1　隐私保护效果及工作流可用性度量指标

基于局部节点对的有向度特征攻击是图结构数据发布中常见的攻击模式.DP-2K 算法[9]提出抵御局部节点对总体度特征攻击(即无向边特征识别攻击)的图数据隐私保护发布方法.本节将 DpriPP 算法与 DP-2K 方法[9]进行对比,验证本文所提方法抵御边特征识别攻击的有效性.采用差异度对隐私保护性能进行量化对比.假设世系工作流 G 隐匿处理并重构出的工作流为 G^*,S 为 G 和 G^* 的 IO-2D 边序列集合,s 表示集合 S 中的某一序列.

差异度 $Diff$ 定义为对原工作流和重构后工作

流采用相同特征提取方法进行特征提取，得出的所有特征差异的总和：

$$Diff=\frac{\sum_{s\in S}(|s|_{G^*}-|s|_G)}{\sum_{s\in S}(|s|_G)}.$$

在满足相同差分隐私要求情况下，重构后工作流与原工作流差异度越大，表示对原工作流的隐私保护效果越好，但同时也表明对原图添加的噪音规模越大，对原工作流图时序特征可用性的维持效果越差；反之，$Diff$ 值越小，表示 G^* 与 G 越相似，该方法对 G 的隐私保护性能越差，对其时序特征的维持效果越好.

DpriPP 算法主要关注世系工作流邻接节点子结构以及工作流时序约束关系可用性的维持，即实现差分处理后重构发布的工作流在较好地保持原世系工作流节点局部结构特征的同时兼顾工作流的时序约束关系的维持. 考虑选择整体时序匹配率 R 来度量工作流时序约束的可用程度. 假设世系工作流 G 隐匿为 G^*，P' 为 G^* 中局部时序路径，P 为 G 中相应的局部时序路径. 对于 P' 路径上的每一个节点 n' 与 P 路径上对应位置的节点 n 进行匹配.

时序匹配率 R 定义如下：

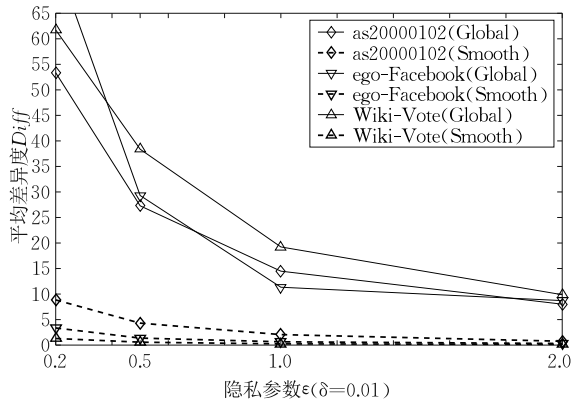
$$R=\frac{\sum_{P\in G\&P'\in G^*}|\{n|n\in P\&n'\in p'\&n'=n\}|}{|\{n|n\in G\&n'\in G^*\}|}.$$

R 表示扰动后工作流 G^* 与原始工作流 G 在整体时序特征上的相似比例. R 值越大，说明 G^* 整体时序与 G 整体时序差异越小；R 值越小，说明 G^* 整体时序与 G 整体时序差异越大.

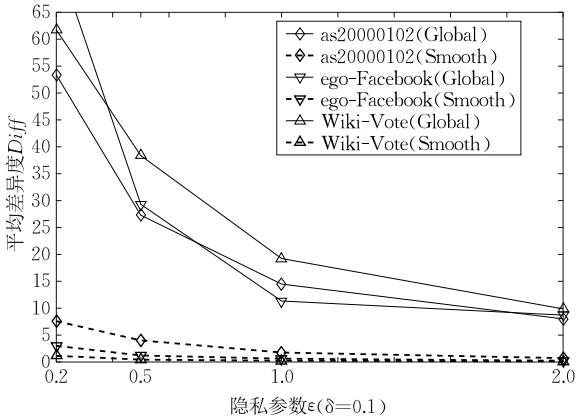
5.2 全局敏感度与平滑敏感度实验效果对比

本节对DpriPP方法分别采用全局敏感度和平滑敏感度实现隐私保护的效果进行实验分析. 采用两种满足差分隐私保护的 Laplace 噪音添加机制，分别基于全局敏感度和平滑敏感度实现对工作流图的概要信息提取扰动和满足相应差分的新工作流图重构发布. 实验中隐私预算参数 ε 分别取 0.2，0.5，1.0，2.0. 对全局敏感度，满足的是严格的参数 ε-差分隐私，即松弛因子 $\delta=0$；对平滑敏感度，满足的是近似的 (ε,δ)-差分隐私，松弛因子 δ 分别取 0.01，0.1，0.5.

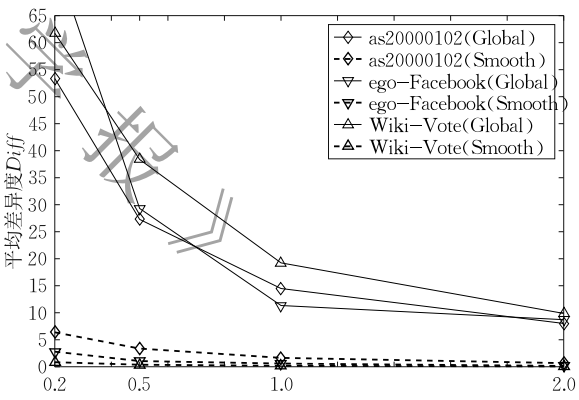
实验 1. 利用扰动前后工作流图在边特征方面的平均差异度，对基于全局敏感度和平滑敏感度的 Laplace 噪音添加机制进行比较. 图 8(a)、(b)、(c) 给出 δ 分别取 0.01、0.1、0.5 情况下，as20000102、ego-Facebook 和 Wiki-Vote 数据集的实验结果. 由图可知，当 δ 相同时，随着隐私预算参数 ε 的增加，平均差异度 $Diff$ 均在减小，这表明隐私预算参数 ε 越大，原数据集中添加的噪音数据规模就越小，隐私保护效果越差；扰动导致的变化越小，结构特征的维持效果也越好.



(a) 当δ=0.01时，各数据集边特征的平均差异度



(b) 当δ=0.1时，各数据集边特征的平均差异度



(c) 当δ=0.5时，各数据集边特征的平均差异度

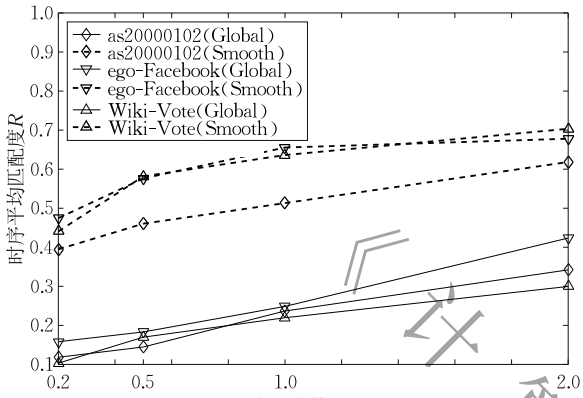
图 8 两种噪音机制下边特征差异度的对比

由图 8 明显可见，基于全局敏感度添加噪音的机制，平均差异度显著高于基于平滑敏感度添加噪音的机制，表明全局敏感度机制下添加的噪音数据幅度是较大的，敏感度激增，扰动给原图带来的破坏严重影响工作流可用性，无法有效维持原工作流图信息.

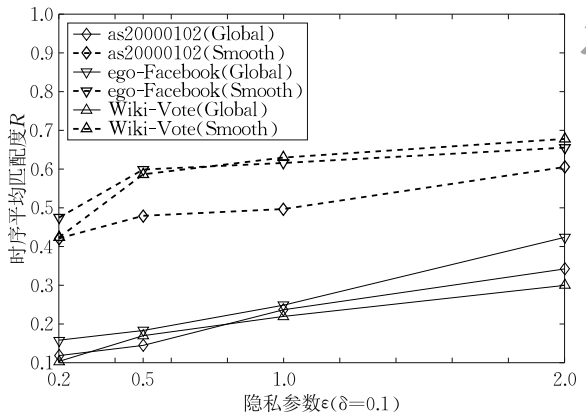
观察比较图 8(a)、(b)、(c)，对 as20000102 数据

集，对相同隐私预算 ε，用基于平滑敏感度噪音添加机制，随着 δ 取值的增大，其平均差异度缓慢的减小，ego-Facebook 和 Wiki-Vote 数据集也表现出同样的减小趋势，表明随着松弛因子 δ 变大，隐私保护强度要求降低，实现 (ε,δ)-差分隐私所需添加噪声规模也越小，差异度也减小.

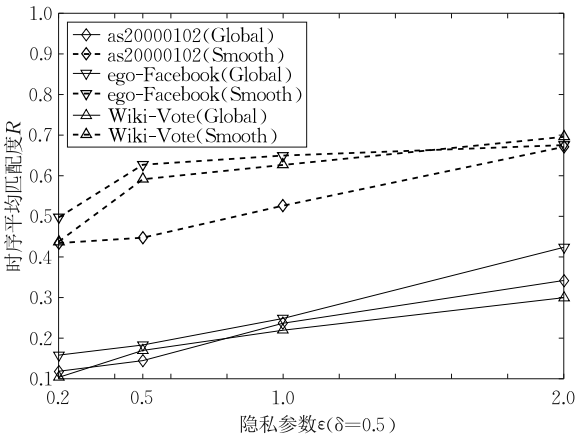
实验 2. 利用时序平均匹配率对基于全局敏感度和基于平滑敏感度的噪音添加机制的工作流时序可用性维持效果进行分析. 图 9 对应 δ 分别取 0.01、0.1、0.5 时，算法时序平均匹配度的变化趋势. 由三组实验可知，δ 相同时，随着隐私预算参数 ε 的增加，时序匹配率亦呈增加趋势，扰动后工作流整体时序维持效果也越好. 因为隐私参数 ε 控制差分隐私扰动程度，随着 ε 增大，扰动程度减小，扰动后数据特征也越接近原始工作流特征，工作流的时序特征维持效果也越好. 由图 9 可知，隐私预算参数 ε 相同时，采取平滑机制会带来更高的时序平均匹配率 R，表明利用平滑机制能够在对原图扰动尽量小的情况下实现工作流时序可用性的有效维持.



(a) 当δ=0.01时，各数据集的时序平均匹配率



(b) 当δ=0.1时，各数据集的时序平均匹配率



(c) 当δ=0.5时，各数据集的时序平均匹配率

图 9　两种噪音机制下时序匹配率的对比

如图 9 所示，给定隐私预算 ε，随着 δ 增大，三个数据集的时序平均匹配率相对稳定，呈现轻微涨幅，表明随着 δ 变大，时序平均匹配率缓慢增长，但增长幅度较小.

对比图 8、图 9 可发现，采用平滑机制可有效减少添加噪音的幅度，较好的维持工作流结构和时序可用性. 因此，本文采用基于平滑敏感度的噪音添加机制实现差分隐私保护. 后续实验，亦采用平滑机制对 DpriPP 算法和 DP2K 算法进行实验分析.

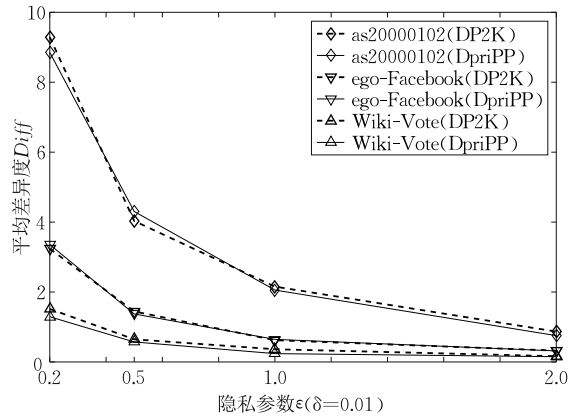
实验 3.　利用扰动前后工作流图在边特征方面的平均差异度，对 DpriPP 算法和 DP2K 算法进行实验分析. 实验结果如图 10 所示，当 δ 值固定时，随着隐私预算 ε 增大，DP2K 算法和 DpriPP 算法的平均差异度均显著下降，当 ε 大于 0.5 时，随着 ε 增加，平均差异度减小幅度趋缓. 表明 ε 越大，隐私保护的要求越低.

从图 10 的三组实验中容易发现对相同隐私预算 ε 和松弛因子 δ，DpriPP 算法能达到或者近似达到 DP2K 算法所实现的差异度，表明两算法在满足相同的差分隐私保护情形下，添加与 DP2K 相同甚至更少的噪音，即能满足与 DP2K 近乎相同的结构特征保护效果，同时 DpriPP 算法能有效维持工作流图整体时序特性，达到较高的时序匹配率.

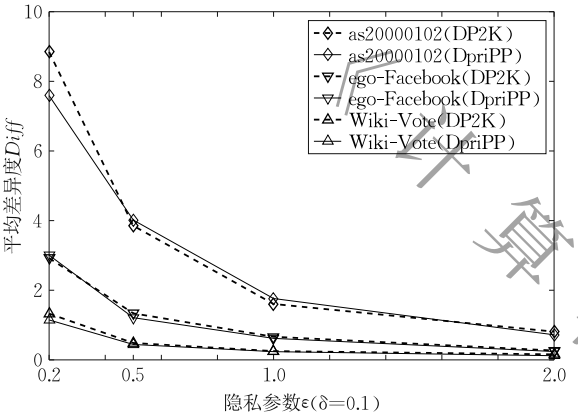
比较图 10(a)、(b)、(c)可知，基于平滑机制进行噪音添加扰动，松弛因子 δ 会影响隐私保护效果. 无论是 DP2K 算法还是 DpriPP 算法，给定隐私预算 ε，松弛因子 δ 增大，平均差异度均会减小. 即随着 δ 的增大，(ε,δ)-差分隐私约束亦减弱.

综上所述，DpriPP 算法在 DP2K 基础上引入了有向度特征，将无向图结构特征保护扩展为有向图结构特征保护，在达到与 DP2K 算法相同的隐私保护强度的条件下，实现与 DP2K 算法相同甚至更优的结构特征的维系和保护. 在此基础上，DpriPP 算法进一步实现了工作流方向性和时序特征可用性的有效维持，可以弥补已有的基于工作流度分布扰动

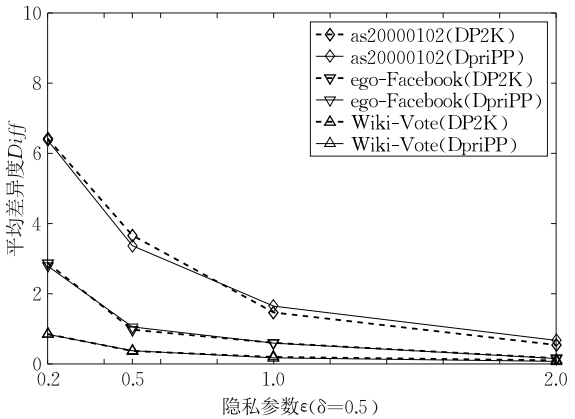
方法难以维系世系工作流方向以及整体时序关系的不足.



(a) 当δ=0.01时,DP2K与DpriPP的平均差异度



(b) 当δ=0.1时,DP2K与DpriPP的平均差异度



(c) 当δ=0.5时,DP2K与DpriPP的平均差异度

图 10 DP2K 与 DpriPP 差异度的对比

6 总 结

目前隐私保护世系工作流发布侧重局部映射关系的维持,对世系工作流可用性的重要表现——工作流时序依赖关系,维持效果较弱;也缺少对工作流局部节点对有向度分布隐私保护问题.我们提出基于差分的隐私保护世系工作流发布方法,该方法采用输入/输出度序列模型提取工作流方向性和时序约束关系,对特征边添加差分噪声,基于特征序列进行世系工作流重构发布,有效保护世系工作流局部节点对度分布隐私的同时兼顾工作流局部与整体时序约束关系的可用性.

下一步,将对兼顾世系工作流中更复杂节点约束关系隐私保护及可用性维持的工作流共享发布技术进行研究.

参 考 文 献

[1] Glavic B, Dittrich K. Data provenance: A categorization of existing approaches//Proceedings of the 6th MMC Workshop of BTW 2007. Aachen, Germany, 2007: 227-241

[2] Gao Ming, Jin Che-Qing, Wang Xiao-Ling, et al. A survey on management of data provenance. Chinese Journal of Computers, 2010, 33(3): 373-389(in Chinese)
(高明, 金澈清, 王晓玲等. 数据世系管理技术研究综述. 计算机学报, 2010, 33(3): 373-389)

[3] Davidson S B, Milo T, Roy S. A propagation model for provenance views of public/private workflows//Proceedings of the International Conference on Database Theory. Genoa, Italy, 2013: 165-176

[4] Bertino E, Ghinita G, Kantarcioglu M, et al. A roadmap for privacy-enhanced secure data provenance. Journal of Intelligent Information Systems, 2014, 43(3): 481-501

[5] Sultana S, Ghinita G, Bertino E, Mohamed S. A lightweight secure scheme for detecting provenance forgery and packet drop attacks in wireless sensor networks. IEEE Transactions on Dependable & Secure Computing, 2015, 12(3): 256-269

[6] Wang Changda, Hussain S R, Bertino E. Dictionary based secure provenance compression for wireless sensor networks. IEEE Transactions on Parallel & Distributed Systems, 2016, 27(2): 405-418

[7] Davidson S B, Khanna S, Roy S, et al. On provenance and privacy//Proceedings of the Database Theory-ICDT 2011. Uppsala, Sweden, 2011: 3-10

[8] Davidson S B, Khanna S, Milo T, et al. Provenance views for module privacy//Proceedings of the 11th ACM SIGMOD Conference on Principles of DB Systems. Athens, Greece, 2011: 175-186

[9] Wang Y, Wu X. Preserving differential privacy in degree-correlation based graph generation. Transactions on Data Privacy, 2013, 6(2): 127-145

[10] Sala A, Zhao X, Wilson C, et al. Sharing graphs using differentially private graph models//Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference. Berlin, Germany, 2011: 81-98

[11] Dwork C. Differential privacy//Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming. Venice, Italy, 2006: 1-12

[12] Zhang Xiao-Jian, Meng Xiao-Feng. Differential privacy in data publication and analysis. Chinese Journal of Computers, 2014, 37(4): 927-949(in Chinese)
(张啸剑，孟小峰. 面向数据发布和分析的差分隐私保护研究. 计算机学报，2014, 37(4): 927-949)

[13] Zhang Sen, Ni Weiwei. Graph embedding matrix sharing with differential privacy. IEEE Access, 2019, (7): 89390-89399

[14] Li Xiaoye, Yang Jing, Sun Zhenlong, Zhang Jianpei. Differential privacy for edge weights in social networks. Security and Communication Networks, 2017, 4(8): 1-10

[15] Day W Y, Li N, Lyu M. Publishing graph degree distribution with node differential privacy//Proceedings of the 2016 International Conference on Management of Data—SIGMOD. San Francisco, USA, 2016: 123-138

[16] Kasiviswanathan S P, Nissim K, Raskhodnikova S, Smith A. Analyzing graphs with node differential privacy//Proceedings of the 10th Theory of Cryptography Conference. Tokyo, Japan, 2013: 457-476

[17] Dumas M, García-Bañuelos L, Laud P. Differential privacy analysis of data processing workflows//Proceedings of the International Workshop on Graphical Models for Security. Lisbon, Portugal, 2016: 62-79

[18] Pettai M, Laud P. Combining differential privacy and mutual information for analyzing leakages in workflows//Proceedings of the Principles of Security and Trust - 6th International Conference. Uppsala, Sweden, 2017, 14: 298-319

[19] Maruseac M, Ghinita G, Rughinis R. Privacy-preserving publication of provenance workflows//Proceedings of the ACM Conference on Data and Application Security and Privacy. San Antonio, USA, 2014: 159-162

[20] Sun Y, Zhao H, Han Q, et al. Composite graph publication considering important data//Proceedings of the 3rd International Conference of Pioneering Computer Scientists, Engineers and Educators. Changsha, China, 2017: 207-219

[21] Dwork C. Differential privacy: A survey of results//Proceedings of the International Conference on Theory and Applications of Models of Computation. Berlin, Germany, 2008: 1-19

[22] Dwork C. A firm foundation for private data analysis. Communications of the ACM, 2011, 54(1): 86-95

[23] Dwork C, Mcsherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis//Proceedings of the 3rd Conference on Theory of Cryptography. New York, USA, 2006: 265-284

[24] Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis//Proceedings of the 39th Annual ACM Symposium on Theory of Computing. San Diego, USA, 2007: 75-84

[25] Xiong Ping, Zhu Tian-Qing, Wang Xiao-Feng. Differential privacy protection and its application. Chinese Journal of Computers, 2014, 37(1): 101-122(in Chinese)
(熊平，朱天清，王晓峰. 差分隐私保护及其应用. 计算机学报，2014, 37(1): 101-122)

[26] Tillman B, Markopoulou A, Butts C T, et al. Construction of directed 2K graphs//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 1115-1124

[27] Gale D. A theorem on flows in networks. Pacific Journal of Mathematics, 1956, 7(2): 1073-1082

NI Wei-Wei, Ph. D., professor, Ph. D. supervisor. His current research interests include data mining and data privacy protection.

SHEN Tao, M. S. candidate. His main research interest is data privacy protection..

YAN Dong, M. S. candidate. His main research interest is data privacy protection.

Background

Data provenance describes the mechanism and process of data generation and evolution, which records information about the node module executions used to produce concrete data items, as well as those intermediate data items acting as parameters passed between nodes' executions. Data provenance plays an important role in research and applications of data quality assessment, data recovery and data analysis. With increasingly deepening of data sharing, the need of publishing and sharing workflow of provenance becomes increasingly urgent. However, the provenance workflow often contains private or confidential data. In recent years, privacy problem in provenance workflow sharing attracts increasingly attention. Existing work mainly focus on the maintenance of the local mapping relationship of provenance workflow. For example,

privacy protection process is implemented to the provenance workflow, which ensures that the degree of a single node would not be leaked, or sensitive mapping relationship would not be leaked in parallel with effective maintaining of provenance workflow's overall input and output mapping relations. As for provenance workflow's utility, i. e. the time-sequence dependence of workflow, most of them does not work. To solve these problems, the definition of input and output degree sequence with scale i model is introduced to describe the degree distribution of provenance workflow nodes. It provides a carrier for describing the utility and privacy-sensitive information of provenance workflow. As a by-product, it can also accommodate the extraction of directional characteristics of the provenance workflow well. The definition of Previous-Next sequence is further devised to describe the substructure distribution characteristics of the workflow. This structure can reduce the possible loss of workflow's temporal relations in adding differential noise. By constructing schema of Previous-Next sequence, the substructure characteristics of nodes and their adjacent nodes in workflow are captured, and the temporal constraints in workflow are maintained during the reconstruction process. On this basis, a differential privacy-based on privacy-preserving provenance workflow publishing method DpriPP is proposed to implement a weak background knowledge-dependent privacy-preserving provenance workflow publication, it can also provide well maintenance to temporal dependance relations of the provenance workflow.

Our work is mainly supported by the Natural Science Foundation of China (No. 61772131 and No. 61370077). The project with No. 61772131 just focuses on the problem of privacy preserving problem in data provenance sharing. Our proposal in this paper is an important research content of the project and will provide direct support for the solution of key issues in the project. The project with number 61370077 concentrates on the problem of privacy preference supporting in privacy preserving location based query.