软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(11):2828−2842 [doi: 10.13328/j.cnki.jos.005099]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向关系-事务数据的数据匿名方法*

龚奇源, 杨 明, 罗军舟

(东南大学 计算机科学与工程学院,江苏 南京 211189)

通讯作者: 杨明, E-mail: yangming2002@seu.edu.cn

摘 要: 在发布同时包含关系和事务属性的数据(简称为关系-事务数据)时,由于关系数据和事务数据均有可能受到链接攻击,需要同时匿名这两部分的数据.现有的数据匿名技术在匿名化关系-事务数据时会造成严重的数据缺损,无法保障数据可用性.针对此问题,提出了(*k*,*l*)-多样化模型,通过等价类上的 *l*-多样化约束和事务数据上的 *k*-匿名约束来保证用户隐私不被泄露.在此基础上,设计并实现了 APA 和 PAA 两种满足该模型的匿名算法,以不同的顺序对关系-事务数据进行匿名,并提出了相应的数据缺损评估方法.实际公开数据集上的实验结果表明,与现有的数据匿名技术相比,APA 和 PAA 能够在保护用户隐私的前提下,以更低的数据缺损和更高的效率完成对关系-事务数据的匿名.

关键词: 数据匿名;隐私泄露;*k*-匿名;*l*-多样化;关系-事务数据

中图法分类号: TP311

中文引用格式: 龚奇源,杨明,罗军舟.面向关系-事务数据的数据匿名方法.软件学报,2016,27(11):2828−2842. http://www.jos.org.cn/1000-9825/5099.htm

英文引用格式: Gong QY, Yang M, Luo JZ. Data anonymization approach for microdata with relational and transaction attributes. Ruan Jian Xue Bao/Journal of Software, 2016,27(11):2828−2842 (in Chinese). http://www.jos.org.cn/1000-9825/5099.htm

Data Anonymization Approach for Microdata with Relational and Transaction Attributes

GONG Qi-Yuan, YANG Ming, LUO Jun-Zhou

(School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

Abstract: When publishing datasets that contain relational and transaction attributes, referred to as RT-data for briefness, either type of data may suffer from linking attacks. Anonymizing both of them is essential. However, previous approaches suffer from huge information loss during anonymizing RT-data, and they fail to preserve the utility of datasets. To address this problem, an anonymization model, (*k*,*l*)-diversity is proposed to ensure privacy by guaranteeing *l*-diversity on each equivalence class and *k*-anonymity on transaction data. In addition, two heuristic algorithms named APA and PAA, which anonymize RT-data in different orders, are also provided to achieve (*k*,*l*)-diversity. Extensive experiments based on real-world dataset show that APA and PAA outperform existing approaches in terms of execution time and information loss.

Key words: data anonymization; privacy breach; *k*-anonymity; *l*-diversity; RT-data

* 基金项目: 国家自然科学基金(61272054, 61572130, 61632008, 61320106007, 61502100, 61402104); 江苏省自然科学基金(BK20150628, BK20140648, BK20150637); 中央高校基本科研业务费专项资金(2242014R30010); 江苏省科技支撑项目(BE2014603); 江苏省青蓝工程; 江苏省网络与信息安全重点实验室资助项目(BM2003201); 教育部网络与信息集成重点实验室资助项目(93K-9)

Foundation item: National Natural Science Foundation of China (61272054, 61572130, 61632008, 61320106007, 61502100, 61402104); Jiangsu Provincial Natural Science Foundation (BK20150628, BK20140648, BK20150637); Fundamental Research Funds for the Central Universities (2242014R30010); Jiangsu Provincial Key Technology R&D Program (BE2014603); Qinglan Project of Jiangsu Province; Program of Jiangsu Provincial Key Laboratory of Network and Information Security (BM2003201); Program of Key Laboratory of Computer Network and Information Integration of the Ministry of Education of China (93K-9)

收稿时间: 2015-11-09; 修改时间: 2016-02-23; 采用时间: 2016-04-21; jos 在线出版时间: 2016-05-03

CNKI 网络优先出版: 2016-05-04 08:44:18, http://www.cnki.net/kcms/detail/11.2560.TP.20160504.0844.010.html

为了在数据共享过程中保护用户隐私,同时尽可能地保持数据完整性,研究人员提出了数据匿名技术[1].数据匿名技术主张在不泄露用户隐私的前提下,对数据进行尽可能少的、不可逆的匿名化操作,降低攻击者获取用户敏感信息的概率,同时保证数据的真实性.匿名后的数据可以安全地共享给第三方,甚至发布到网络中.

大部分现有数据匿名研究主要针对关系数据或者事务数据,不能有效处理同时包含两类数据的关系-事务数据.其中,关系数据指包含若干关系型属性的数据,这些属性中联合起来能够唯一标识用户身份的属性被称为准标识符;事务数据指包含若干取值的集合型数据,其取值中通常包含用户的敏感信息.在真实的数据应用中,绝大部分的数据集都属于关系-事务数据,例如诊疗数据中的年龄、性别和邮编属于关系数据,而病人的多次诊疗记录属于包含用户敏感信息的事务数据.在关系-事务数据匿名发布中,由于关系数据和事务数据均有可能受到链接攻击,需要同时匿名这两部分数据.为此,文献[2]首次针对这个问题提出了(k,k^m)-匿名模型,通过对等价类同时施加 k-匿名和 k^m-匿名约束来保护用户隐私.但是,该文献提出的匿名模型过于严格,按照该模型匿名关系-事务数据会造成严重的数据缺损,无法保障数据可用性.

针对关系-事务数据匿名中的数据可用性问题,本文提出(k,l)-多样化模型,通过等价类上的 l-多样化约束和事务数据上的 k-匿名约束,保证用户隐私不被泄露.在此基础上,本文设计并实现了满足该模型的 APA (anatomize and partition anonymization)和 PAA(partition and anatomize anonymization)算法,用真实的公开数据集对其进行测试和评估,并与相关工作进行对比.本文的主要贡献如下:

(1) 提出了(k,l)-多样化匿名模型,通过等价类上的 l-多样化约束和事务数据上的 k-匿名约束,保证关系-事务数据中的用户隐私不被泄露.

(2) 针对(k,l)-多样化模型,设计并实现了两种满足该模型的匿名算法 APA 和 PAA,并给出了相应的数据缺损评估方法;用真实的公开数据集对 APA 和 PAA 算法进行了评估和分析,并与相关工作进行对比.

本文第 1 节介绍相关工作.第 2 节介绍数据匿名基本概念和定义.第 3 节介绍关系-事务数据匿名技术,提出(k,l)-多样化匿名模型,并给出满足该模型的 APA 和 PAA 算法.第 4 节分析和评估实验结果.第 5 节总结全文.

1 相关工作

近年来,能够保证数据集真实性的数据匿名技术得到了许多研究者的关注.根据匿名对象不同,可以将现有研究工作分为三大类:关系数据匿名、事务数据匿名和关系-事务数据匿名.3 类研究工作都包含两方面的内容:数据匿名模型和数据匿名算法[3].其中,数据匿名模型研究主要针对特定的泄露风险,在理论上建立约束模型,为数据匿名算法提供理论依据和指导;数据匿名算法研究主要在数据匿名模型约束下设计高效的匿名算法,力求在不泄露用户隐私的前提下,以尽可能低的数据缺损完成匿名.在特定模型约束下,以最小数据缺损代价实现的匿名被称为最优匿名.但是在现有数据匿名模型约束下,实现最优匿名均为 NP-hard[4],所以常用的数据匿名算法均为近似算法.

(1) 关系数据匿名

在关系数据匿名中,一般将数据属性分为两类:联合起来能够唯一标识用户身份的准标识符和包含用户敏感信息的敏感属性.Sweeney 和 Samarati 首先指出准标识符上链接攻击带来的隐私泄露问题,并提出 k-匿名模型[1]——通过保证每条记录都有至少 k–1 条记录与它在准标识符上无法区分,来确保数据受到链接攻击时不会泄露隐私信息.Meyerson 等人在文献[4]中证明了当 $k\geqslant 2$ 时,通过最小的泛化实现数据 k-匿名的问题是 NP-hard.文献[5,6]分别讨论了高维 k-匿名问题和多约束 k-匿名问题.k-匿名没有对敏感属性进行约束,当大部分记录具有相同的敏感属性取值时,攻击者能够以较高的概率推断出用户的敏感信息.因此,Raymond 等人[7]在 k-匿名的基础上提出了(α,k)-匿名,保证发布数据在满足 k-匿名的同时,每个等价类中与任意敏感属性取值相关的记录不超过 1/α.文献[8]提出了安全性更高的 l-多样化,保证等价类中任意敏感属性取值的出现频率不高于 1/l,使得攻击者获取到用户敏感属性的概率不高于 1/l;文中还提出约束更强的 entropy-l-多样化,要求每个等价类中的敏感属性信息熵不低于 l.文献[9]在 l-diversity 的基础上考虑敏感属性的分布问题,并提出了 t-Closeness,保证不同等价类中的敏感属性分布尽量接近于全局分布.文献[10]指出在数据增量发布的过程中,现有数据匿名模型会造

成隐私泄露,并提出了 m-Invariance 模型.文献[11]提出了个性化隐私保护(personalized privacy preservation)的匿名模型,为用户提供不同粒度的隐私保护.文献[12]提出了面向匿名查询的数据匿名模型——差分隐私(differential privacy),通过保证相差一条记录的不同集合有较大的概率具有相同的查询结果来保护匿名查询中的用户隐私.需要注意的是,数据匿名模型并不是越严格越好.数据匿名模型要求越严格,匿名后数据集的安全性越高,但是实现匿名的代价也越大,导致数据可用性越差.

与数据匿名模型不同,数据匿名算法的研究主要侧重于针对特定的数据匿名模型设计高效的数据匿名算法,在满足匿名模型的基础上保证匿名后数据的可用性.文献[13]给出了基于全局泛化空间完全搜索的 k-匿名算法 MinGen.文献[14]给出了近似比为 $O(\log k)$和 $O(\beta\log k)$的近似算法($\beta\geqslant 1$),其中,近似比为 $O(\beta\log k)$的算法通过牺牲一定的近似比实现了更高的匿名效率.文献[15]给出了基于全局泛化空间剪枝的 k-匿名算法 Incognito,该算法通过动态规划的方式可以有效减少搜索空间.虽然全局泛化方法的搜索空间较小,生成的匿名化数据格式统一、便于分析,但是会造成较高的数据缺损.为此,文献[16]提出了局部泛化技术,通过进一步细化泛化粒度,降低数据缺损.文献[17]在之前泛化算法的基础上提出了多维度泛化技术 Mondrian,通过多维度划分,将泛化粒度再次缩小.随着数据维度的增加,泛化技术造成的数据缺损会迅速增加[5],为此,文献[18]提出了基于有损分解的数据匿名方法 Anatomy,通过有损分解弱化准标识符和敏感属性的关系,降低攻击者获取到敏感信息的概率.文献[19]提出了基于泛化和有损分解的匿名方法 ANGEL,通过结合泛化和有损分解,大幅度降低数据缺损.

在匿名查询方面,差分隐私技术得到了广泛的关注[20,21].针对现有差分隐私方法中数据可用性差的问题,文献[22]提出了基于 Wavelet 的 Privelet 方法,提高了匿名数据集的范围查询精度.针对差分隐私无法有效保护数据关系的问题,文献[23]提出了面向非交互式网络数据的差分隐私方法.

(2) 事务数据匿名

与关系型数据不同,事务数据由于其维度不固定,匿名方式有很大的不同.Ghinita 等人[24]将事务数据划分为公开的事务数据和敏感的事务数据,并提出利用带状矩阵来压缩高维稀疏事务数据;在此基础上,他们设计了一种基于泛化和置换的数据匿名算法.文献[25]针对发布敏感事务数据导致隐私泄露的问题提出了(h,k,p)-coherent 模型,通过同时约束敏感和非敏感事务数据保证用户隐私不被泄露.在攻击者获取的背景知识维度不超过 m 的前提假设下,文献[26]提出了 k^m-匿名,保证攻击者背景知识维度不超过 m 时,获取到的组合数大于 k;在此基础上,他们给出了 Apriori Anonymization 算法,并通过 count tree 结构降低了算法复杂度.文献[27]发现,k^m-匿名无法保护长度大于 m 的组合;同时,基于全局泛化的 Apriori Anonymization 算法会造成严重的数据缺损.为了解决这个问题,文中提出了满足 k-匿名的局部泛化的算法 Partition,通过自顶向下划分和回溯来保证数据被近似最优的分组,从而降低数据缺损.文献[28]在 Apriori Anonymization 算法的基础上设计并实现了基于格雷码的局部泛化算法,降低了数据缺损.

(3) 关系-事务数据匿名

关系-事务数据同时包含关系数据和事务数据,并普遍存在于关系数据库中.发布该类型数据能够为更复杂的数据应用(如并发症分析、购物偏好分析等)提供支持.但是,上述研究工作主要针对单一型数据匿名,无法直接应用到关系-事务数据中,否则会造成隐私泄露[2].为此,Poulis 等人在文献[2]中首次针对关系-事务数据的匿名问题展开研究,提出了(k,k^m)-匿名模型,通过保证每个等价类在关系数据上满足 k-匿名,同时在事务数据上满足 k^m-匿名的方式来保护用户隐私,并给出了满足该模型的 RMR 等算法.但是,该文献提出的(k,k^m)-匿名模型和算法存在 3 个方面的问题:

1) 仅能保护事务数据中长度不超过 m 的组合,对于长度超过 m 的组合不施加任何保护;
2) 该模型没有多样性约束,无法防御同质攻击[8];
3) 对于等价类要求过严格,导致严重的数据缺损.

其中,第 3 个问题尤为严重.例如,表 1 为诊疗数据,表 2 为按照(2,2^2)-匿名要求发布的表 1 数据.虽然攻击者无法通过关系数据和事务数据唯一标识 Bob 并获取他的疾病信息,但是,为了形成满足(2,2^2)-匿名的等价类,匿名算法泛化了大量的数据,导致事务数据全部缺损.此外,由于采用了基于聚类的匿名算法和 Apriori

Anonymization 算法,RMR 算法运行效率非常低.

Table 1 Relational and transaction data

表 1 关系-事务型数据

ID	Name	Age	Sex	Zipcode	Disease
1	Bob	18	M	12 000	$\langle a_1,b_2\rangle$
2	Alice	25	F	21 000	$\langle a_2,b_2\rangle$
3	Tom	19	M	14 000	$\langle c_1\rangle$
4	Simon	21	M	18 000	$\langle c_2\rangle$
5	Jim	24	M	19 000	$\langle b_2,c_2\rangle$
6	Alen	29	F	22 000	$\langle b_2,c_1\rangle$

Table 2 Anonymized data by (k,k^m)-anonymous

表 2 (k,k^m)-匿名后的数据

Age	Sex	Zipcode	Disease
[15,20)	M	[10000,15000)	⟨,⟩
[15,20)	M	[10000,15000)	⟨,⟩
[20,25)	M	[15000,20000)	⟨,⟩
[20,25)	M	[15000,20000)	⟨,⟩
[25,30)	F	[20000,25000)	⟨,⟩
[25,30)	F	[20000,25000)	⟨,⟩

针对上述问题,本文拟从两方面展开研究:

1) 更合理的数据匿名模型:通过保护任意长度的事务数据组合,并增加多样性约束来增强匿名模型安全性;通过放宽对等价类的约束来提高数据可用性.
2) 更高效的数据匿名算法:采用更高效的匿名子算法,并采用有损分解的方式匿名关系-事务数据,提高匿名效率.

2 基本定义

本文针对关系-事务数据展开研究,该类型数据普遍存在于关系数据库系统中.为了简化问题,我们将关系-事务数据(RT-data)中的所有关系属性作为准标识符(quasi-identifier,简称 QI),将事务数据均作为敏感属性(sensitive attribute,简称 SA).同时,我们按照文献[2]的方式,假设每个用户对应一条关系-事务数据记录,且在该记录中,事务数据中的取值不重复[26−28].本文用 T 表示需要匿名的数据集,n 表示 T 中的记录数,T^*代表匿名化之后的数据,t 表示 T 中某一条数据记录,t_r 表示 t 的关系数据取值.由于用户在事务数据上的取值没有先后顺序且不会重复,可以用集合表示,我们称其为事务数据组合,记为 c,并用 t_c 表示 t 中的事务数据组合.我们用 A 表示属性,d 代表关系数据属性数目,则 T 中共有 d+1 个属性,记为 $A_1,A_2,\ldots,A_d,A_{d+1}$;其中,$A_1,A_2,\ldots,A_d$ 表示关系数据属性,A_{d+1} 表示事务数据属性.本文所用符号见表 3.

Table 3 Symbols

表 3 符号表

Symbol	Explanation
T, T^*	原始数据和匿名化后数据
n, d	数据集规模,关系数据属性数量
$A_1,A_2,\ldots,A_d$	关系数据属性
A_{d+1}	事务数据属性
t	T 中某一条记录
c	事务数据组合
t_r, t_c	t 的关系数据和事务数据组合
QI, *SA*	准标识符和敏感属性
EC, *TBucket*	等价类,事务数据桶

定义 1(等价类(equivalence class))[17]. 数据表 T 中,在关系数据上具有相同取值的所有记录形成等价类,记

为 *EC*.

定义 2(事务数据桶(transaction data bucket)). 数据表 T 中,在事务数据上具有相同取值组合的所有记录构成事务数据桶,记为 *TBucket*.

EC 和 *TBucket* 是分别根据关系数据(准标识符)取值和事务数据组合是否相同对数据集 T 进行的划分.假设 T 中有 x 个 *EC* 和 y 个 *TBucket*,则 $\bigcup_{i=1}^{x} EC_i = \bigcup_{j=1}^{y} TBucket_j = T$, 且 $\forall i \neq j, EC_i \cap EC_j = \varnothing, TBucket_i \cap TBucket_j = \varnothing$.

定义 3(泄露风险(disclosure risk)). 攻击者通过关系数据或者事务数据取值,唯一标识用户身份并获取到敏感信息的概率.

根据 Zhou 等人[3]的研究,数据的隐秘性由其泄露风险决定:泄露风险越高,隐秘性越差.本文将泄露风险分为关系数据泄露风险和事务数据泄露风险,并取较大值作为泄露风险.由于 *EC* 内的记录在关系数据上无法区分,故关系数据泄露风险与|*EC*|成反比.另一方面,由于 *TBucket* 内的记录在事务数据上无法区分,故事务数据泄露风险与|*TBucket*|成反比.根据这个性质,我们定义 T 中任意记录 t 的关系数据泄露风险 $Risk(t_r)$为

$$Risk(t_r) = \frac{1}{|EC_i|} \tag{1}$$

定义 t 的事务数据泄露风险 $Risk(t_c)$为

$$Risk(t_c) = \frac{1}{|TBucket_j|} \tag{2}$$

定义 t 的泄露风险 $Risk(t)$为

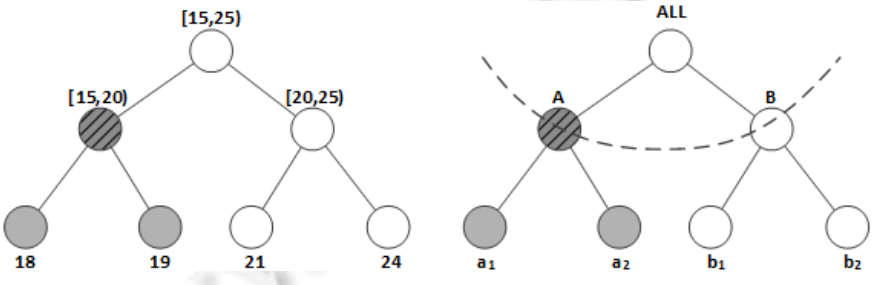
$$Risk(t)=\max(Risk(t_r),Risk(t_c)) \tag{3}$$

其中,$t_r \in EC_i, t_c \in TBucket_j$.以表 1 为例,根据公式(1)~公式(3),由于 Bob 的事务数据组合$\langle a_1,b_2\rangle$只出现一次,*TBucket* 大小为 1,故事务数据泄露风险为 100%,即攻击者通过事务数据唯一标识 Bob 的概率是 100%.同时,由于 Bob 在关系数据上的取值⟨18,M,12000⟩只出现一次,故其在关系数据上的泄露风险为 100%,即攻击者通过关系数据上的背景知识唯一标识 Bob 的概率为 100%.最终,Bob 在表 1 中的泄露风险为 100%.为了保护用户隐私,需要保证数据集的泄露风险低于某个阈值.

目前,实现数据匿名的方法主要有两类:数据泛化和有损分解.前者通过一定的方式模糊数据取值,使得原有等价类不断合并,最终满足匿名要求;后者则通过分组和表切分的方式,将原有的一一对应关系破坏为一对多关系,使得每个等价类至少对应多个不同的 *SA* 取值.

定义 4(数据泛化(generalization)). 在数据处理过程中,用模糊的、范围的取值取代精确取值的过程被称为数据泛化.

泛化实质上是对数据的粗粒度化,会造成不可逆的数据缺损.为了降低数据缺损,同时保留数据的语义,需要建立泛化层次结构(generalization hierarchy),如图 1 所示.



(a) 关系数据泛化 (b) 事务数据泛化

Fig.1 Relational data generalization vs. transaction data generalization

图 1 关系数据泛化和事务数据泛化

现有的数据泛化技术主要分为两类:关系数据泛化[12−16]和事务数据泛化[23−26],分别如图 1(a)和图 1(b)所示.关系数据泛化中,每个属性对应不同的泛化层次,泛化过程实质是寻找泛化层次中能够覆盖相应取值的最低公共子节点[1],例如,将 18 和 19 岁泛化为[15,20);与之相反,事务数据泛化一般只有一个泛化层次,泛化过程实质为寻求泛化层次中能够覆盖所有取值的最低公共切割[26−28],例如将$\langle a_2,b_2\rangle$泛化为$\langle A,B\rangle$,将$\langle a_1,a_2\rangle$泛化为$\langle A\rangle$.随着数据的维度增加,数据泛化造成的缺损会急剧增加,最终可能会丢失所有信息[5].为此,Xiao 等人[18]提出了有损分解技术.

定义 5(有损分解[18]). 通过将数据表 $T(QI,SA)$划分为 $QIT(QI,GID)$表和 $ST(GID,SA,Count)$表,将原有的 QI 与 SA 的一一对应关系转化为通过 GID 维持的一对多关系,保证攻击者无法通过准标识符获取到用户敏感属性取值.

有损分解不会改变原有等价类,但是通过切分数据表 T,将原有的 QI 与 SA 的一一对应关系打破,使得每个等价类对应多个 SA 取值.攻击者只能通过自然连接 QIT 和 ST 的方式来重构数据,但是根据自然连接的性质,重构会产生很多并不存在于 T 中的记录,导致获取敏感信息的几率大为降低.Xiao 等人[18]还提出了一种基于有损分解的匿名算法 Anatomize,通过保证每个 GID 对应的敏感属性取值不小于 l,使得数据集满足 l-多样化.

3 关系-事务数据匿名

本节中,我们首先提出(k,l)-多样化模型,通过等价类上的 l-多样化约束和事务数据上的 k-匿名约束,保证用户隐私不被泄露.在此基础上,本文设计并实现了满足该模型的匿名算法.

3.1 (k,l)-多样化匿名模型

为了在关系-事务数据匿名中保护用户隐私,需要同时约束关系数据和事务数据.此外,为了防御同质攻击[8],需要增加多样性约束.因此,匿名后的关系-事务数据集 T^*至少需要满足以下 3 个条件:

(1) 事务数据组合不能唯一标识用户,避免事务数据上的链接攻击;

(2) 关系数据不能唯一标识用户,避免关系数据上的链接攻击;

(3) 任意等价类中的事务数据组合满足多样化约束,避免同质攻击.

目前,常用的匿名模型[1,8,27]只能满足条件(1)或者条件(2).文献[2]虽然提出了满足条件(1)、条件(2)的模型,但是该模型无法保护长度超过 m 的事务数据组合[27],也无法防御同质攻击;并且,该模型对等价类的要求过于严格,会造成严重的数据缺损.为了解决上述问题,本文在前人研究的基础上,提出满足条件(1)~条件(3)的(k,l)-多样化模型,通过同时约束关系数据和事务数据,保证用户隐私不被泄露;在此基础上,适当放宽对等价类的约束,提高数据可用性.为了便于理解,我们首先介绍(k,l)-多样化模型的两个子模型.

定义 6(事务数据 k-匿名(k-anonymity on transaction data))[27]. 如果数据表 T^*中的每个 $TBucket$ 中至少包含 k 条记录,则称该数据表 T^*满足事务数据 k-匿名.

在事务数据 k-匿名约束下,每个 $TBucket$ 均不小于 k,保证事务数据组合上的泄露风险不高于 $1/k$.T^*满足条件(1),但不满足条件(2)、条件(3),攻击者仍然可以通过关系数据唯一标识用户.

定义 7(关系-事务数据 l-多样化(l-diversity on RT-data)). 如果数据表 T^*中,攻击者通过事务数据组合关联到等价类中任意事务数据组合的概率不高于 $1/l$,则称该数据表 T^*满足关系-事务数据 l-多样化.

在关系-事务数据 l-多样化约束下,每个等价类至少包含 l 个不同的事务数据组合,且任意事务数据组合在等价类中的出现频率不高于$|EC|/l$,故关系数据上的泄露风险不高于 $1/l$.T^*满足条件(2)、条件(3),但不满足条件(1),攻击者仍然可以通过事务数据组合唯一标识用户.

定义 8((k,l)-多样化((k,l)-diversity)).如果 T^*满足以下条件,则称 T^*满足(k,l)-多样化:

(1) T^*中任意 $TBucket$ 中至少包含有 k 条记录;

(2) T^*中任意 EC 中,攻击者通过事务数据组合关联到等价类中任意事务数据组合的概率不高于 $1/l$.

如上所述,(k,l)-多样化对关系-事务数据集施加了双重约束:约束(1)保证每个事务数据组合至少对应 k 个用户,使得 T^*满足条件(1);约束(2)保证攻击者通过事务数据组合关联到等价类中任意事务数据组合的概率不高

于 1/l,使得 T^*满足条件(2)、条件(3).从而保证 T^*中的关系数据和事务数据不会泄露用户隐私.根据以上性质,我们还可以得到引理 1 和引理 2.

引理 1. 如果数据集 T^*满足(k,l)-多样化,则 T^*中任意事务数据组合必然满足事务数据 k-匿名.

证明:为了证明引理 1,我们假设 T^*满足(k,l)-多样化,且 T^*中存在事务数据组合 c 不满足事务数据 k-匿名,则 c 所在 $TBucket_c$ 中的记录数必然小于 k,$TBucket_c$ 不满足定义 8(1),与假设矛盾. □

引理 2. 如果数据集 T^*满足(k,l)-多样化,则 T^*中任意等价类必然满足关系-事务数据 l-多样化.

证明:为了证明引理 2,我们假设 T^*满足(k,l)-多样化,且 T^*中存在等价类 x 不满足关系-事务数据 l-多样化,则 x 中必然存在某个事务数据组合 c,并且 c 在 x 中的出现频率超过|x|/l,使得攻击者能够以高于 1/l 的概率关联到 c,不满足 8(2),与假设矛盾. □

定理 1. 如果数据集 T^*满足(k,l)-多样化,则攻击者通过事务数据唯一标识 T^*中任意用户身份的概率低于 1/k,通过关系数据唯一标识用户的概率低于 1/l.

证明:根据引理 1,攻击者通过事务数据组合唯一标识用户身份的概率低于 1/k;根据引理 2,T^*中的任意等价类满足 l-多样化.故攻击者通过关系数据唯一标识用户身份的概率低于 1/l. □

在定理 1 的保证下,即使攻击者获取到了用户的所有事务数据取值或者关系数据取值,也无法唯一标识用户身份,攻击者通过事务数据或者关系数据获取用户隐私的概率低于 max{1/k,1/l}.参数 k 和 l 没有约束关系,故可以根据数据集的实际特性分别选取.表 4 为满足(k,l)-多样化的表 1 匿名后版本(由 PAA 或 APA 匿名表 1 可得),其中,k=2,l=3.故攻击者通过关系数据和事务数据唯一标识用户的概率不高于 1/2,用户隐私得到保护.

Table 4 Dataset satisfies (k,l)-diversity

表 4 满足(k,l)-多样化的数据

(a) RT

RID	Age	Sex	Zipcode	Group ID
1	18	M	12000	1
2	25	F	21000	2
3	19	M	14000	1
4	21	M	18000	2
5	24	M	19000	1
6	29	F	22000	2

(b) TT

Group ID	Disease
1	$\langle A,b_2\rangle$
1	$\langle C\rangle$
1	$\langle b_2,C\rangle$
2	$\langle A,b_2\rangle$
2	$\langle C\rangle$
2	$\langle b_2,C\rangle$

需要注意的是,(k,l)-多样化并不要求每个等价类均满足事务数据 k-匿名.因此,即使 T^*满足(k,l)-多样化,其中的等价类也未必满足事务数据 k-匿名.通过适当放宽对等价类的约束,(k,l)-多样化能够更好地保障数据可用性.与文献[2]中的(k,k^m)-匿名相比,(k,l)-多样化不存在 m 的限制,对事务数据的约束更强;其次,通过多样化约束,可以防御同质攻击;最后,由于(k,l)-多样化对于等价类的约束更宽松,能够更好地保证数据可用性.

3.2 APA和PAA算法

现有的大部分匿名算法都不支持关系-事务数据,无法同时保证关系数据和事务数据不泄露用户隐私.仅有的匿名算法 RMR 因为采用了基于聚类的匿名算法和 Apriori Anonymization 算法,计算开销非常大.为了高效地匿名关系-事务数据,本文在现有研究的基础上,基于 Partition 算法[27]和 Anatomize 算法[18]设计了 APA 和 PAA 算法.通过融合关系数据和事务数据匿名技术,实现关系-事务数据匿名.最终发布的数据存储在 RT(relational table)和 TT(transaction table)中,RT 中包含关系数据和 GID,TT 中包含事务数据和 GID.其中,Partition 算法通过自顶向下的方式对数据集进行划分,再通过局部泛化对数据进行泛化,可以保证泛化后的数据满足事务数据 k-匿名,算法复杂度为 $O(n^2)$;Anatomize 算法则通过对分桶后的敏感属性进行多样化分组,可以确保每个等价类至少对应 l 个不同的敏感属性取值,使整个数据集满足 l-多样化,算法复杂度为 $O(n)$.

为使这两种算法符合关系-事务数据匿名需求,我们对它们进行了以下优化:为 Anatomize 算法增加了对事务数据的支持,按照事务数据组合建立不同的 *TBucket*,并优化了剩余记录处理机制;为 Partition 增加了对关系-事务数据的支持,并在原有算法的基础上优化了数据分组模块,提高了算法效率.限于篇幅,本文不对改进后的

Anatomize 和 Partition 做过多的描述.

由于 Anatomize 和 Partition 在匿名过程中会互相影响,直接调用两种算法并不能实现关系-事务数据匿名.为此,我们设计的协调机制如下:

- 首先,为了避免子算法覆盖彼此的分组结果,我们参照 ANGEL[19]的设计思路,同时保留两种数据划分方式,即 *EC* 和 *TBucket* 并存的方式,使得两种子算法能够相对独立的匿名关系或者事务数据部分.
- 其次,我们有效地利用了两种算法的中间结果,提高了算法效率.例如,Partition 算法的中间结果 *B* 能够被 Anatomize 算法直接利用,省去了根据事务数据组合建立桶的阶段.
- 最后,为了避免 Partition 算法降低 Anatomize 分组的多样性,导致匿名度降低,我们设计了分组合并阶段,通过分组合并,使所有分组满足关系-事务数据 *l*-多样化,保证算法的匿名度.

在此基础上,我们根据不同的匿名顺序设计了 APA 和 PAA 算法,其伪代码如下所示:

算法 1. APA (anatomize and partition anonymization).

Input:*k*,*l*,原始数据集 *T*.

Output:包含关系数据和 GID 的 RT 和包含事务数据和 GID 的 TT.

//利用改进的 Anatomize 对数据进行匿名化

将事务数据组合相同的记录划分到同一个 *TBucket* 中,得到 *TBucket* 组,记为 *B*;

利用 Anatomize 从 *B* 中生成分组 *G*;

利用 Partition 对事务数据组合进行 *k*-匿名化.

//分组合并

取出 *G* 中事务数据组合小于 *l* 的组合,放置于 *CG* 中;

While *CG* 不为空

从 *CG* 中 Pop 一个分组 *g*;

If *CG* 中存在与 *g* 合并之后事务数据组合大于 *l* 的 *g*′

从 *CG* 中 Pop *g*′;

$g=g\cup g'$;

将 *g* 加入到 *G*;

Else

将 *g* 合并到 *G* 中任意分组中;

//有损分解

按照 Anatomize 的方式对 *G* 内数据进行有损分解,得到 RT 和 TT;

Return RT 和 TT

算法 2. PAA (partition and anatomize anonymization).

Input:*k*,*l*,原始数据集 *T*.

Output:包含关系数据和 GID 的 RT 和包含事务数据和 GID 的 TT.

//利用改进的 Partition 对事务数据匿名

利用 Partition 对事务数据组合进行 *k*-匿名化,使得 *T* 中的记录自然地形成 *TBucket* 组,记为 *B*;

利用 Anatomize 从 *B* 中生成分组 *G*;

//有损分解

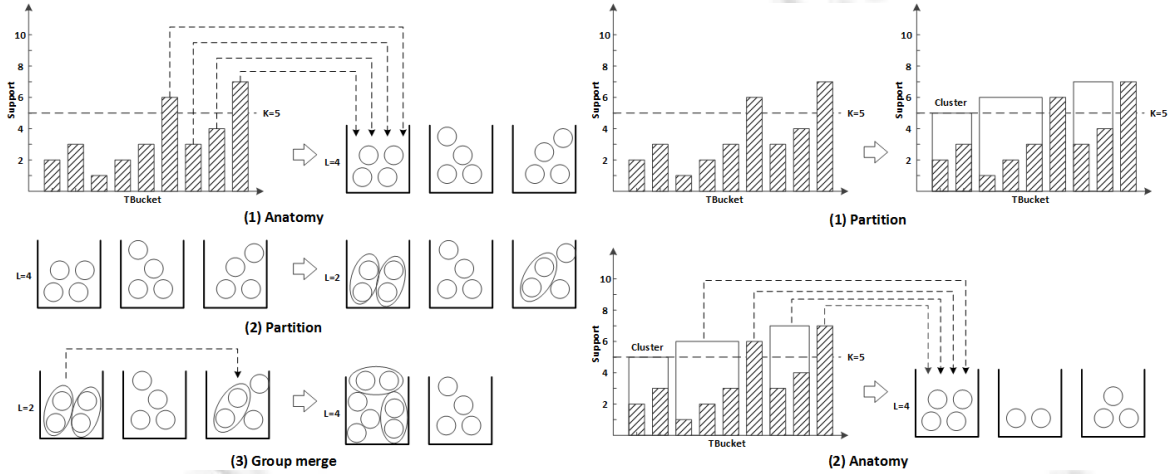
按照 Anatomize 的方式对 *G* 数据进行有损分解,得到 RT 和 TT;

Return RT 和 TT

如算法 1 和图 2(a)所示,APA 分为 3 个阶段:有损分解、事务数据匿名和分组合并.首先,利用改进的 Anatomize 算法对 *T* 进行划分,保证每个 *EC* 至少包含 *l* 个不同的事务数据组合,复杂度为 $O(n)$;随后,利用改进后的 Partition 算法对事务数据进行 *k*-匿名化,保证所有 *TBucket* 中的记录数大于 *k*,使其满足事务数据 *k*-匿名,复杂

度为 $O(n^2)$.但是 Partition 会降低数据集中事务数据组合的多样化特性,导致第 1 阶段划分的分组内事务数据组合减少,部分分组的事务数据组合数目会小于 l(不再满足 l-多样化).例如图 2(a)中,Partition 后数据集由 4-多样化变为了 2-多样化.所以,最后一个阶段,我们利用分组合并算法对事务数据组合数目小于 l 的分组进行合并,使得所有分组均满足 l-多样化,其复杂度为 $O(n^2)$.最终 APA 的复杂度为 $O(n^2)$,远低于 RMR 的 $O(n^4)$.

如算法 2 和图 2(b)所示,PAA 分为两个阶段:事务数据匿名和有损分解.首先,利用改进的 Partition 算法对 T 中所有事务数据组合进行 k-匿名化,保证所有 *TBucket* 中的记录数大于 k,使其满足事务数据 k-匿名,复杂度为 $O(n^2)$.由于 Partition 算法使用了事务数据泛化,匿名化后事务数据组合相同的记录会自动地形成 *TBucket*.所以,可以直接利用改进后的 Anatomize 算法对 Partition 算法形成的 *TBucket* 进行 l-多样化划分,复杂度为 $O(n)$.与 APA 不同,按照这个顺序进行匿名化,两种算法可以有机地结合起来,各司其职,互不干扰.最终,PAA 的复杂度为 $O(n^2)$,远低于 RMR 的 $O(n^4)$.



(a) APA (b) PAA

Fig.2 Stages of APA and PAA

图 2 APA 和 PAA 的基本步骤

与文献[2]中的 RMR 相比,APA 和 PAA 选择了线性的 Anatomize 算法和自上而下的 Partition 算法,计算开销远远低于 RMR 中基于聚类的关系数据匿名算法和 Apriori Anonymization 算法.因此,APA 和 PAA 的复杂度更低,运行效率更高.同时,由于 APA 和 PAA 没有采用 k^m-匿名,能够更好地保护用户隐私.

4 实验验证和分析

本节中,我们利用真实的公开数据集来评估和分析 APA 和 PAA 的有效性及其性能,并与文献[2]中提出的 RMR 进行对比.

4.1 实验设置

实验所使用的数据集为 INFORMS(https://sites.google.com/site/informsdataminingcontest/)数据集.INFORMS 数据集包含 16 个属性、102 578 条用户记录(包含 58 568 个用户)、317 872 条诊疗记录.实验中,我们保留其中 6 个属性:Month of birth,Year of birth,race,Years of education,Income,Diagnosis codes.其中,Years of education 和 Income 属性为有序属性,其他属性均为无序属性.实验中,我们取前 5 个属性作为关系数据,Diagnosis codes 作为事务数据.实验数据集中,各个属性的特性见表 5.由于无法获取 Diagnosis codes 的泛化层次结构,本文参照文献[27]的方式将泛化空间平均划分,以 5 作为单位,每 5 个节点上升一层(INFORMS 中事务数据

取值域为1 297,由于5^4<1297≤5^5,故最终层次为5).按照这种划分方式,不需要人为地指定泛化层次即可实现匿名化.实验中,k和l的默认取值为10和5,qd和s的默认取值为5和2.为了公平起见,RMR算法的阈值δ设定为0.65,m设定为2.

实验的硬件环境为HP ProLiant DL 580 G5,16G RAM,操作系统为Ubuntu 13.04.实验中的所有算法均由Python实现(APA,PAA和RMR代码已上传至Github(https://github.com/qiyuangong)).

Table 5 Attributes selected for experiment

表5 实验所选属性

分类	Quai-Identifier (relation)					Sensitive attribute (transaction)
属性	Month of birth	Year of birth	race	Years of education	Income	Diagnosis codes
不同取值	12	88	6	22	16 532	632

4.2 数据缺损评估

为了对数据匿名算法进行客观评估,我们需要量化数据缺损[3].现有的信息缺损衡量技术主要分为两类:(1) 通过缺损度公式来衡量缺损,如NCP[16]缺损矩阵[15];(2) 通过聚集查询和计数查询的方式来衡量数据缺损[18].但是,这两类数据缺损衡量方法不适用于关系-事务数据,直接应用现有技术会造成事务数据部分无法 衡量.

为此,本文基于文献[18]的计数查询技术,提出面向关系-事务数据的计数查询,通过匿名化前后的计数查询结果差异度来衡量数据匿名造成的数据缺损.在实际应用中,计数查询是很多数据应用的基本操作,例如数据统计、关联规则挖掘和OLAP.匿名数据集的计数查询结果与原始数据集越接近,则其数据应用的结果与原始数据集的差异也越小.

计数查询基本的查询语句如下:

SELECT COUNT(*) **FROM** T^*

WHERE $pred_1(A_1)$ AND $pred_1(A_2)$ AND … AND $pred_1(A_{qd})$ AND $pred_2(A_{d+1})$

查询子条件$pred_1$和$pred_2$如下:

$pred_1$: (A=x_1 OR A=x_2 OR … OR A=x_b)

$pred_2$: (A_{n+1} *contains*($pred_3(c_1)$ OR $pred_3(c_2)$ OR … OR $pred_3(c_b)$))

$pred_3$: (v_1 AND v_2 AND … AND $v_{|c|}$)

其中,A代表属性(关系数据中的准标识符或者事务数据),x代表A中的具体取值,qd代表选择的准标识符个数,b表示该属性上需要选取的取值数目.b的计算方法如下:

$$b=\lceil |A|s^{1/(qd+1)} \rceil \tag{4}$$

其中,s表示选取比例,0≤s≤100%.每次计数查询时,我们先确定qd和s的取值.随后,随机选取qd个准标识符,并根据公式(4)计算每个属性上b的取值.之后,按照b从关系数据的属性上选取一定数量的取值$\{x_1,..,x_b\}$来形成$pred_1$.再按照b从T中所有的事务数据组合中选取一定数量的事务数据组合$\{c_1,..,c_b\}$来形成$pred_2$.需要注意的是,每个事务数据组合c中包含若干事务数据取值v,所有取值必须都出现在记录中,查询$pred_3$才为True.最后,利用查询语句分别查询原始数据集T和匿名后的数据集T^*.

通过查询T,可以得到精确的计数结果act;通过查询T^*,并按照文献[18]的方式计算概率,可以得到est.利用act和est,我们可以计算T和T^*在本次计数查询中的差异度,即相对误差RE(relative error).

$$RE=\frac{|act-est|}{act} \tag{5}$$

例如,我们随机生成了一个查询(Age=18 or 19,Sex=M,$Disease$=$\langle a_1,b_2\rangle$),通过查询表1,我们得到act=1,通过查询表2,我们得到est=0.4,查询表4得到est=0.66.所以,表2的RE为60%,而表4的RE为34%.通常情况下,由于数据集T和T^*中的取值不满足均匀分布,RE会随着随机选取取值的不同而波动.为了客观反映T和T^*在计数查询中的差异度,需要计算多次RE求平均值,即平均相对误差ARE(average relative error).

$$ARE=\frac{\sum_{i=1}^{qt}RE_i}{qt} \tag{6}$$

其中,qt 表示计数查询次数.qt 取值越大,ARE 越稳定,计算 ARE 的时间也越长.实际实验中,需要根据数据集选取合适 qt.在实验中,我们将 qt 设置为 1 000,即进行 1 000 次随机的计数查询求一组 ARE.

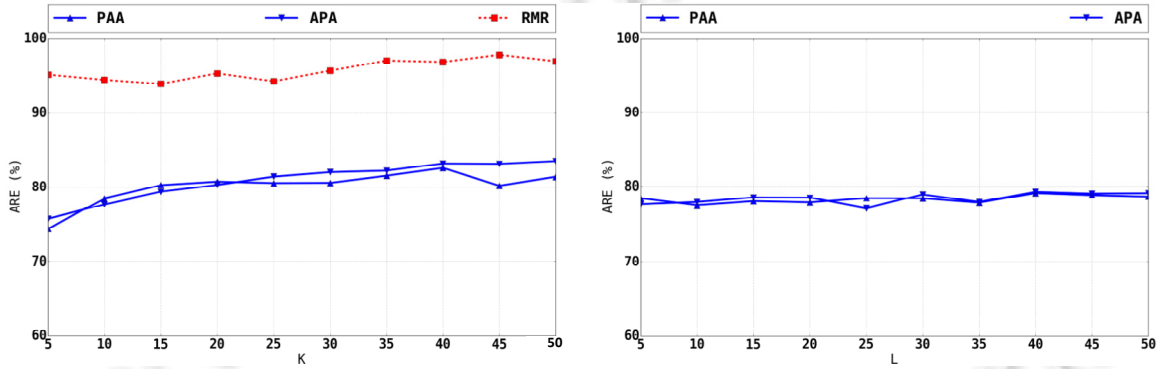
为了方便理解,一般用百分比来表示 ARE,其取值范围为 $ARE\geqslant 0\%$;ARE 可以评估匿名后数据集 T^*和原始数据集 T 之间的计数查询差异:ARE 越大,act 和 est 的差异度越大,数据集缺损越严重.当 est 与 act 差异较大时,ARE 有可能超过 100%.

4.3 实验分析

实验中,我们通过 ARE 来衡量 APA,PPA 和 RMR 的有效性,通过运行时间来衡量 3 种算法的性能.需要注意的是,由于关系-事务数据中的事务数据维度不固定,且部分事务数据维度较高,造成匿名后的 ARE 普遍偏高.

(1) 数据缺损分析

为了分析算法造成的数据缺损,我们选取 k,l,qd,s 和数据集规模共 5 个属性.每次选择其中某个属性并改变其取值,用以分析 3 种算法对这些属性的敏感程度.由于 RMR 算法中没有参数 l,故图 3(b)中没有 RMR 的曲线.



(a) ARE vs. K (b) ARE vs. L

Fig.3 Information loss of APA and PAA when varying k and l

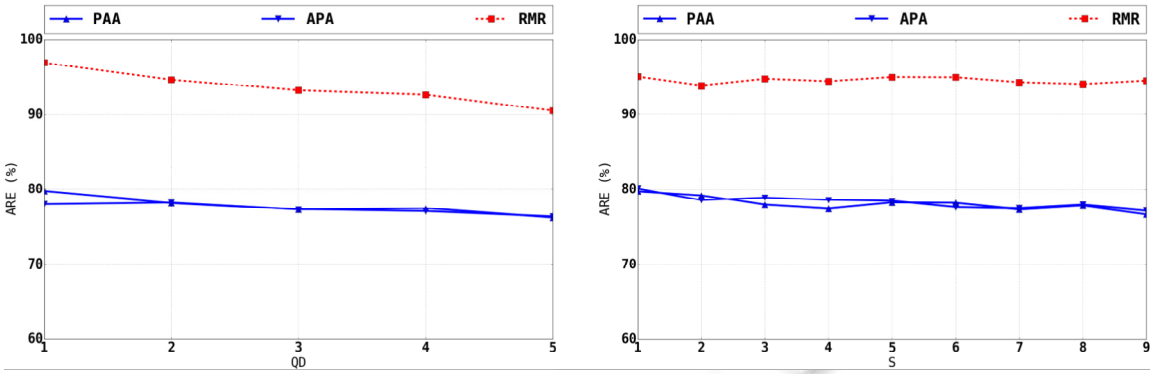
图 3 k 和 l 变化对 APA 和 PAA 数据缺损的影响

如图 3(a)所示,我们发现,随着 k 的增加,3 种算法造成的数据缺损都会缓慢增加,这是因为 k 越大,等价类中包含的记录越多,实现匿名的代价越大;同时,在所有的 k 值上,APA 和 PAA 造成的缺损远远低于 RMR,当 k 值较大时,RMR 算法的 ARE 甚至接近 100%.如图 3(b)所示,由于子算法 Anatomize 对 l 不敏感,所以 l 的增加不会对 APA 和 PAA 造成太大影响,ARE 只是因为随机生成的查询出现稍许的波动.

通过分析图 4 我们发现,qd 和 s 两个计数查询的属性对于 3 种算法的 ARE 有一定的影响.随着 qd 的增加,引入的属性会增加;而随着 s 增加,选取的查询空间会变大.随着引入属性的增加和查询空间的增加,事务数据上的查询误差被稀释,导致 3 种算法的 ARE 均有所下降.同时,在任意 qd 和 s 上,APA 和 PAA 的数据缺损都低于 RMR 算法.

为了分析数据集规模对两种算法的影响,我们以 5K 为单位,顺序选取 12 个数据集.通过对比实验我们可以看到,在所有测试数据集上,APA 和 PAA 的数据缺损都低于 RMR 算法.另外,如图 5 所示,随着数据集规模的增加,APA 和 PAA 造成的数据缺损逐步降低.这是因为 Partition 算法和 Anatomize 算法造成的数据缺损都会随着数据集规模增加而减少.随着数据集规模的增加,等价类中的记录数越来越多,满足 k-匿名和 l-多样化的概率也会随之增加,从而使得缺损度大为降低.但是,数据集规模的增加会造成 RMR 的 ARE 增加.这是因为随着数据集

的增加,RMR 算法会更趋向于合并分组,从而造成数据缺损增加.



(a) ARE vs. *QD* (b) ARE vs. *S*

Fig.4 Information loss of APA and PAA when varying *qd* and *s*

图 4 *qd* 和 *s* 变化对 APA 和 PAA 数据缺损的影响

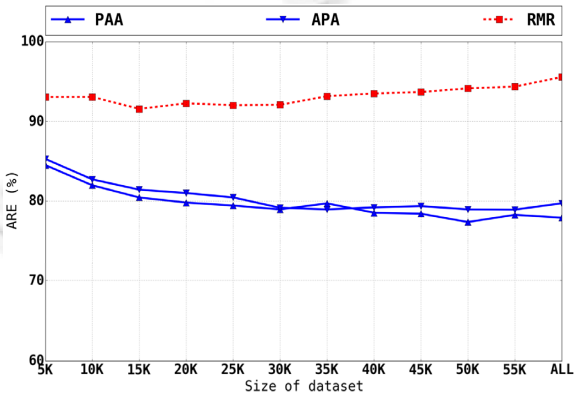


Fig.5 Information loss of APA and PAA when varying dataset size

图 5 数据集规模对 APA 和 PAA 数据缺损的影响

从图 3~图 5 中我们可以看到,APA 和 PAA 造成的数据缺损远远低于 RMR;同时,PAA 算法造成的数据缺损比 APA 算法略低.通过分析中间结果我们发现,APA 形成的分组比 PAA 少,平均分组大小高于 PAA,造成一定程度的过度匿名,导致缺损度增加.

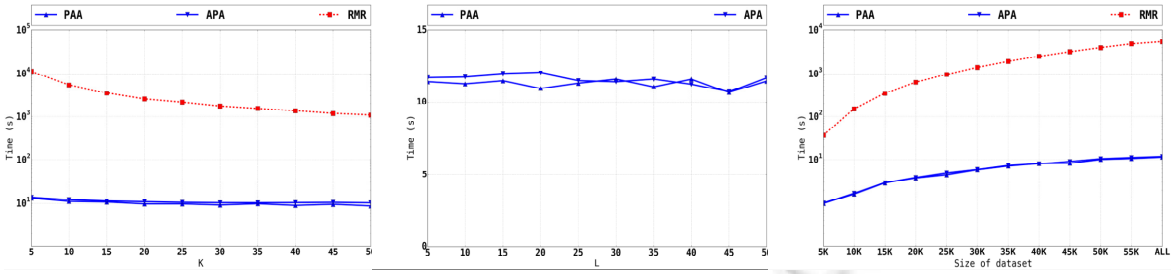
(2) 性能分析

为了分析 APA,PAA 和 RMR 的执行效率,我们从 *k*,*l* 和数据集规模这 3 个角度出发进行分析.由于 RMR 算法中没有参数 *l*,故图 6(b)中没有 RMR 的曲线.

如图 6(a)所示,随着 *k* 的增加,3 种算法的效率均会提高.这是因为随着 *k* 的增加,算法的划分次数减少,运行时间随之减少.但是 *l* 的取值变化对于 APA 和 PAA 的运行效率影响很小,如 6(b)所示.因为 Partition 算法和 Anatomize 算法的执行效率本身对 *l* 不敏感,改变 *l* 对总体的运行时间没有太大影响.如图 6(c)所示,随着数据集规模的增大,3 种算法的运行时间直线上升.这是因为,数据匿名算法的效率受数据集规模(n)影响,数据集规模越大运行时间越长.同时,随着数据集规模的增加,*EC* 和 *TBucket* 数量都会增加,造成运行时间增加.

我们可以看到,APA 和 PAA 的运行效率远远高于 RMR.这与我们之前的分析相符:APA 和 PAA 的复杂度为 $O(n^2)$,远低于 RMR 的 $O(n^4)$.因此在同等条件下,APA 和 PAA 能够比 RMR 更快地完成匿名任务.在所有的实验

中,APA 和 PAA 均能在 12s 内完成匿名化;而 RMR 的运行时间则长达上千秒,当 k 值较小时,RMR 甚至需要上万秒的运行时间.同时,PAA 算法的效率普遍高于 APA 算法.通过分析中间结果我们发现,两种算法在前 2 个阶段的运算时间基本相同,但是 PAA 算法比 APA 算法少一个分组合并阶段,所以运行效率较高.



(a) Time vs. K (b) Time vs. L (c) Time vs. size of dataset

Fig.6 Running time of APA and PAA when varying dataset size, k and l

图 6 数据集规模、k 和 l 变化对 APA 和 PAA 运行时间的影响

(3) 实验小结

在上述实验中,APA 和 PAA 都表现出了极高的运行效率和相对较低的数据缺损度.在所有参数组合上,APA 和 PAA 都可以在 12s 内完成对 INFORMS 数据集的匿名化,并保证计数查询的平均相对误差低于 85%.与文献[2]中 RMR 算法相比,我们的算法具有更高的运行效率和更低的数据缺损.通过对比实验我们发现,PAA 算法比 APA 算法执行效率要高,造成的数据缺损也更少.

5 结束语

针对关系-事务数据匿名中的数据可用性问题,本文提出了(k,l)-多样化模型以保护发布数据中的用户隐私,并基于该模型,以 Anatomize 和 Partition 为基础设计了 APA 和 PAA 算法.与现有的技术相比,两种算法能够以更低的数据缺损和更高的运行效率完成关系-事务数据匿名.但是引入 Anatomize 对于属性间的关联关系具有一定的破坏作用,原有的数据关系变为概率性的结果,造成关联分析的难度增加.因此,如何将其他匿名方法应用到(k,l)-多样化中,将是我们下一阶段的研究方向.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是东南大学的倪巍伟教授和美国马萨诸塞大学洛威尔分校的李晓白教授表示感谢.

References:

[1] Sweeney L. k-Anonymity: A model for protecting privacy. Int'l Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002,10(5):557−570. [doi: 10.1142/S0218488502001648]

[2] Poulis G, Loukides G, Gkoulalas-Divanis A, Skiadopoulos S. Anonymizing data with relational and transaction attributes. In: Proc. of the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD). Berlin, Heidelberg: Springer-Verlag, 2013. 353−369. [doi: 10.1007/978-3-642-40994-3_23]

[3] Zhou SG, Li F, Tao YF, Xiao XK. Privacy preservation in database applications: A survey. Chinese Journal of Computers, 2009, 32(5):847−861 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2009.00847]

[4] Meyerson A, Williams R. On the complexity of optimal K-anonymity. In: Proc. of the 23rd ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems. ACM Press, 2004. 223−228. [doi: 10.1145/1055558.1055591]

[5] Aggarwal CC. On k-anonymity and the curse of dimensionality. In: Proc. of the 31st Int'l Conf. on Very Large Data Bases. VLDB Endowment, 2005. 901−909.

[6] Yang XC, Liu XY, Wang B, Yu G. *K*-Anonymization approaches for supporting multiple constraints. Ruan Jian Xue Bao/Journal of Software, 2006,17(5):1222−1231 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1222.htm [doi: 10.1360/jos171222]

[7] Wong RCW, Li J, Fu AWC, Wang K. (α,*k*)-Anonymity: An enhanced *k*-anonymity model for privacy preserving data publishing. In: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2006. 754−759. [doi: 10.1145/1150402.1150499]

[8] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. *L*-Diversity: Privacy beyond *k*-anonymity. ACM Trans. on Knowledge Discovery from Data, 2007,1(1):Article 3. [doi: 10.1145/1217299.1217302]

[9] Li N, Li T, Venkatasubramanian S. *t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity. In: Proc. of the IEEE 23rd Int'l Conf. on Data Engineering. IEEE, 2007. 106 −115. [doi: 10.1109/ICDE.2007.367856]

[10] Xiao XK, Tao YF. *M*-Invariance: Towards privacy preserving re-publication of dynamic datasets. In: Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2007. 689−700. [doi: 10.1145/1247480.1247556]

[11] Xiao XK, Tao YF. Personalized privacy preservation. In: Proc. of the 2006 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2006. 229−240. [doi: 10.1007/978-0-387-70992-5_19]

[12] Dwork C. Differential privacy. In: Automata, Languages and Programming. Berlin, Heidelberg: Springer-Verlag, 2006. 1−12. [doi: 10.1007/11787006]

[13] Sweeney L. Achieving *k*-anonymity privacy protection using generalization and suppression. Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002,10(5):571−588. [doi: 10.1142/S021848850200165X]

[14] Park H, Shim K. Approximate algorithms for *K*-anonymity. In: Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2007. 67−78. [doi: 10.1145/1247480.1247490]

[15] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full-domain *K*-anonymity. In: Proc. of the 2005 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2005. 49−60. [doi: 10.1145/1066157.1066164]

[16] Xu J, Wang W, Pei J, Wang X, Shi B, Fu AWC. Utility-Based anonymization using local recoding. In: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2006. 785−790. [doi: 10.1145/1150402.1150504]

[17] LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional *K*-anonymity. In: Proc. of the 22nd Int'l Conf. on Data Engineering. IEEE Computer Society, 2006. 25−36. [doi: 10.1109/ICDE.2006.101]

[18] Xiao XK, Tao YF. Anatomy: Simple and effective privacy preservation. In: Proc. of the 32nd Int'l Conf. on Very Large Data Bases. VLDB Endowment, 2006. 139−150.

[19] Tao YF, Chen H, Xiao XK, Zhou SG, Zhang D. ANGEL: Enhancing the utility of generalization for privacy preserving publication. IEEE Trans. on Knowledge and Data Engineering, 2009,21(7):1073 −1087. [doi: 10.1109/TKDE.2009.65]

[20] Zhang XJ, Meng XF. Differential privacy in data publication and analysis. Chinese Journal of Computers, 2014,37(4):927−949 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2014.00927]

[21] Xiong P, Zhu TQ, Wang XF. A survey on differential privacy and applications. Chinese Journal of Computers, 2014,37(1):101−122 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2014.00101]

[22] Xiao XK, Wang G, Gehrke J. Differential privacy via wavelet transforms. IEEE Trans. on Knowledge and Data Engineering, 2011, 23(8):1200−1214. [doi: 10.1109/TKDE.2010.247]

[23] Chen R, Fung BC, Yu PS, Desai BC. Correlated network data publication via differential privacy. The VLDB Journal, 2014,23(4): 653−676. [doi: 10.1007/s00778-013-0344-8]

[24] Ghinita G, Tao YF, Kalnis P. On the anonymization of sparse high-dimensional data. In: Proc. of IEEE 24th Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2008. 715 −724. [doi: 10.1109/ICDE.2008.4497480]

[25] Xu Y, Fung B, Wang K, Fu A, Pei J. Publishing sensitive transactions for itemset utility. In: Proc. of 8th IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society, 2008. 1109 −1114. [doi: 10.1109/ICDM.2008.98]

[26] Terrovitis M, Mamoulis N, Kalnis P. Privacy-Preserving anonymization of set-valued data. In: Proc. of the VLDB Endow. VLDB Endowment, 2008. 115−125. [doi: 10.14778/1453856.1453874]

[27] He Y, Naughton JF. Anonymization of set-valued data via top-down, local generalization. In: Proc. of the VLDB Endow. VLDB Endowment, 2009. 934−945. [doi: 10.14778/1687627.1687733]

[28] Terrovitis M, Mamoulis N, Kalnis P. Local and global recoding methods for anonymizing set-valued data. The VLDB Journal, 2011,20(1):83−106. [doi: 10.1007/s00778-010-0192-8]

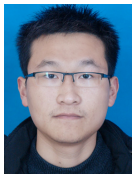
附中文参考文献:

[3] 周水庚,李丰,陶宇飞,肖小奎.面向数据库应用的隐私保护研究综述.计算机学报,2009,32(5):847−861. [doi: 10.3724/SP.J.1016.2009.00847]

[6] 杨晓春,刘向宇,王斌,于戈.支持多约束的 *K*-匿名化方法.软件学报,2006,17(5):1222−1231. http://www.jos.org.cn/1000-9825/17/1222.htm [doi: 10.1360/jos171222]

[20] 张啸剑,孟小峰.面向数据发布和分析的差分隐私保护.计算机学报,2014,37(4):927−949. [doi: 10.3724/SP.J.1016.2014.00927]

[21] 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. [doi: 10.3724/SP.J.1016.2014.00101]



龚奇源(1986－),男,江苏江阴人,博士生,主要研究领域为数据匿名.

罗军舟(1960－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为下一代网络体系结构,协议工程,网络安全,云计算,无线局域网.

杨明(1979－),男,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为网络安全,无线网络.