

一种有效的差分隐私事务数据发布策略

欧阳佳 印 鉴 刘少鹏 刘玉葆  
(中山大学信息科学与技术学院 广州 510006)  
(ouyangjial@163.com)

An Effective Differential Privacy Transaction Data Publication Strategy

Ouyang Jia, Yin Jian, Liu Shaopeng, and Liu Yubao  
(School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006)

**Abstract** For the past few years, privacy preserving data publishing which can securely publish data for analysis purpose has attracted considerable research interests in database community. However, the sparsity of the transaction data burdens the trade-off between privacy protection and enough utility maintaining. Most existing data publishing methods for transaction data are based on partition-based anonymity models, for example  $k$ -anonymity. They depend on background knowledge from the attack, and the published data cannot meet the needs of the analysis tasks. In contrast, differential privacy is a strong privacy model which provides strong privacy guarantees independent of an adversary's background knowledge and also maintains high utility for the published data. Because most existing methods and privacy models cannot accommodate both utility and privacy security of the data, in this paper, an application-oriented TDPS(transaction data publish strategy) is proposed, which is based on differential privacy and compressive sensing. Firstly, an entire Trie tree is constructed for a transaction database. Secondly, based on compressive sensing, we get a noisy Trie tree by adding the differential privacy noisy to the Trie tree. Finally, the frequent itemset mining task is performed on the noisy Trie tree. Theoretical analysis and experimental results demonstrate that the TDPS can preserve privacy of the sensitive data well, meanwhile maintain better data utility.

**Key words** privacy preserving; differential privacy; transaction data; Trie tree; compressive sensing

**摘 要** 近年来,隐私保护事务数据发布得到了研究者的广泛关注.事务数据的稀疏性导致个体隐私保护与数据效用性之间很难达到平衡.目前已有的方法大多是基于分组的匿名模型,但该类模型依赖于攻击者背景知识,且发布的数据无法满足事务数据分析任务的需要.针对事务数据隐私保护发布的数据安全性与效用性不足,基于差分隐私与压缩感知理论,提出一种有效的面向应用的事务数据发布策略(transaction data publish strategy, TDPS).首先构建事务数据库的完整Trie项集树,然后基于压缩感知技术对项集树添加满足差分隐私约束的噪音得到含噪Trie项集树,最后在含噪树上进行频繁项集挖掘任务.实验结果表明,TDPS不仅能很好地保护隐私,而且能有效保持数据效用性,满足事务数据分析任务对数据质量的要求.

**关键词** 隐私保护;差分隐私;事务数据;Trie树;压缩感知

**中图法分类号** TP311.13

收稿日期:2013-06-18;修回日期:2013-07-08  
基金项目:国家自然科学基金项目(61033010,61272065,61472453);广东省自然科学基金项目(S2011020001182);广东省科技计划基金项目(2009B090300450,2010A040303004,2011B040200007)  
通信作者:印 鉴(issjiyin@mail.sysu.edu.cn)

在信息时代互联网成为人们不可缺少的部分. 人们通过互联网购物会留下大量历史记录, 如购物记录、Web 点击流等事务信息. 这些信息记录被公司或机构广泛收集. 共享这些事务数据利于挖掘与利用数据中的知识, 如进行市场分析、客户行为分析等. 但事务数据中往往包含个体的敏感信息, 发布这些数据会泄露个体的隐私信息给研究者或恶意用户(称为攻击者)<sup>[1]</sup>. 随着云计算与电子商务的发展, 这种隐私泄露威胁日益严重.

隐私保护事务数据发布的关键问题是在保护个体隐私的同时为数据分析提供足够多的信息<sup>[2]</sup>. 一方面, 事务数据在发布之前必须进行匿名处理, 以保护个体隐私; 另一个方面, 数据发布的目的是为了有效分析数据, 发布的数据必须具有很高的效用性以满足分析的需要.

然而, 保证数据隐私与效用性之间的平衡是一个挑战. 首先, 由于事务数据的准标识符(quasi-identifiers)与敏感信息无明显区别, 攻击者易获得个体部分信息, 通过已知项组合查询数据集, 攻击者能唯一识别个体对应的事务, 导致受害者购买记录泄露. 显然, 预先知道攻击者所有可能已知项的组合是不可能的. 上述原因导致已有的关系数据库匿名模型如  $k$ -匿名模型<sup>[3-4]</sup>、 $l$ -diversity<sup>[5]</sup>等, 不能直接应用于事务数据<sup>[6]</sup>. 其次, 因为事务数据的高维与稀疏性, 匿名处理后的数据效用性不足<sup>[7]</sup>. 尽管已提出众多基于划分的事务数据匿名模型及对应的匿名算法<sup>[6,8-16]</sup>, 如  $k$ -anonymity<sup>[8]</sup>、 $k^m$ -anonymity<sup>[12-13]</sup>等, 但由于攻击者背景知识的复杂性及匿名算法过程的不确定性, 导致新的攻击方式出现<sup>[17-18]</sup>, 泄露个人隐私. 差分隐私(differential privacy)<sup>[19-21]</sup>是一种完全独立于攻击者背景知识的强隐私概念, 近年来已成为研究热点. 它假定攻击者拥有任意的背景知识, 无论特定的个体记录是否在数据集中, 对该数据集的分析或查询的结果在形式上不可区分, 即结果不强依赖于单个记录. 因此, 满足差分隐私约束的发布策略能对数据提供强有力的隐私保护.

数据的效用性是隐私保护数据发布的难点, 事务数据由于其稀疏特性导致传统的基于分组的匿名方法的效用性不足. 针对数据稀疏性的特点, 文献[22]基于压缩感知理论提出一种感知机制(compressive mechanism, CM). CM 将每条事务看作原始信号, 由于事务是稀疏的, 可以无损恢复原事务, 为防止泄露用户隐私, 对观测信号添加噪音, 使其满足差分隐私约束. CM 方法能有效保护用户隐私, 保证每条事务数据的效用性. 然而, 由于单独处理每条

事务数据, 打破了事务与事务之间项的联系, 即项集, 而这种联系经常用于数据分析, 如频繁项集挖掘. 因而这种感知机制 CM 不能直接应用于事务数据发布.

Trie 树<sup>[23-24]</sup>是一种常用于频繁项集挖掘<sup>[25-27]</sup>的高效数据结构, 它用来表示事务的项集, 合理地保持事务项与项之间的联系. 本文针对事务数据发布的隐私保护问题, 结合压缩感知理论与 Trie 树结构, 提出一种有效的满足差分隐私约束的事务数据发布策略, 该策略能很好地保护个人隐私并提供很高的数据效用性.

本文主要贡献如下:

1) 提出一种有效的满足差分隐私约束的事务数据发布策略(transaction data publish strategy, TDPS). 首先, 基于项集  $I$ , 构建事务数据库  $D$  的完整 Trie 树  $T_b^I$ ; 然后基于压缩感知技术对  $T_b^I$  添加满足差分隐私约束的噪音得到含噪 Trie 树  $NT_b^I$ ; 最后在  $NT_b^I$  上进行频繁项集挖掘任务.

2) 实验结果表明, TDPS 不仅能很好地保护隐私, 而且能有效保持数据效用性, 满足事务数据分析任务对数据质量的要求.

## 1 相关工作

本节简要介绍隐私保护事务数据发布的最新进展与差分隐私的应用.

### 1.1 隐私保护事务数据发布

近年来, 因为数据挖掘的广泛应用, 事务数据发布的隐私保护问题已成为研究热点<sup>[8-16,18,28]</sup>. 根据项的敏感性可划分为区分敏感项和不区分敏感项 2 类.

文献[9,14,16,28]将项分为敏感的(sensitive)或非敏感的(non-sensitive), 并假设攻击者的背景知识只包含非敏感项. Ghinita 等人<sup>[28]</sup>提出一种基于桶(bucketization)的方法, 限定推断敏感项的概率不能超过某个阈值, 同时为频繁模式挖掘保留项之间的关系; Xu 等人<sup>[9]</sup>假设攻击者最多拥有  $p$  个非敏感项, 采用全局消除的方式保留更多的项; 文献[14]通过保留频繁项集以及对边界的表示对文献[9]的方法进行改进; Cao 等人<sup>[16]</sup>假设攻击者的背景知识同时包含敏感项和非敏感项, 提出  $\rho$ -uncertainty 隐私概念, 要求包含敏感项的项集的置信度不能超过  $\rho$ . 但该类方法的效用性不足.

文献[8,12-13]没有区分项的敏感性, 适应性更强. Terrovitis 等人<sup>[13]</sup>假设攻击者的背景知识最多包含  $k$  个项, 提出一种新的隐私模型  $k$ -anonymity,

通过自底向上的泛化方式对数据进行匿名;为增强数据效用性, Terrovitis 等人<sup>[12]</sup>采用局部编码的方式满足  $k^m$ -anonymity 要求;He 等人<sup>[8]</sup>指出  $k^m$ -anonymity 的隐私保护力度低于  $k$ -anonymity, 并基于  $k$ -anonymity 提出一种自顶向下的局部泛化方法. 该类方法的效用性有所提高,但匿名过程均是确定性的,存在新的隐私泄露问题<sup>[17-18]</sup>.

1.2 差分隐私

差分隐私(differential privacy)<sup>[7,20-21]</sup>是一种完全独立于攻击者背景知识的强隐私概念,近年来已成为研究热点. 它假定攻击者拥有任意的背景知识,无论特定个体记录是否在数据集中,对该数据集的任意计算分析或查询的结果在形式上不可区分. 差分隐私随机算法对任意 2 个邻近数据集进行操作,得到的结果几乎是一致的. 目前,差分隐私已被应用于各种不同数据结构的隐私发布<sup>[7,29-32]</sup>. 特别地, Xiao 等人<sup>[30]</sup>基于小波分析提出一种  $\epsilon$ -差分隐私发布策略,为区间查询提供准确的结果;Hay 等人<sup>[29]</sup>针对图的度分布估计问题,提出一种有效的差分隐私算法;Mcsberry 等人<sup>[31]</sup>为用户行为提供推荐的同时满足差分隐私约束;Chen 等人<sup>[32]</sup>首次提出事务数据的差分隐私发布机制. 总之,差分隐私是一种有效的隐私保护机制,它在保护数据隐私的同时,为数据分析保留足够的有用信息.

2 预备知识

本节首先介绍事务数据库及它的 Trie 树表示;其次描述差分隐私模型;然后介绍压缩感知理论;最后给出效用性定义.

2.1 事务数据库

令  $I=\{1,2,\cdots,m\}$  是项的集合. 事务数据库  $D$  是事务的集合,其中每个事务  $T$  是项的集合,使得  $T\subseteq I$ . 每一个事务有一个标识符,本文称之为 TID. 表 1 描述了一个事务数据库.

Table 1 A Sample Transaction Dataset  
表 1 事务数据库示例

TID	Itemset	Count
T1—T10	1,2	10
T11—T20	1,3,4	10
T21—T30	1,2,4	10
T31—T40	2,4	10
T41—T50	1,3,4	10
Total		50

2.2 Trie 树<sup>[23]</sup>

Trie 树又称单词查找树,事务数据库能表示为 Trie 树,可在 Trie 树上直接进行事务数据分析任务,如频繁项集挖掘<sup>[26-27]</sup>. 定义事务数据库  $D$  的 Trie 树为  $T_D$ ,如图 1 所示:

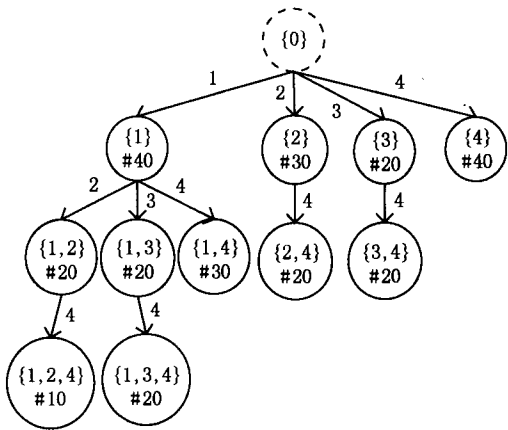


Fig. 1 The Trie tree for the sample transaction dataset.  
图 1 事务数据库示例的 Trie 树表示

图 1 表明,  $T_D$  覆盖了  $D$  的所有项集. 节点的标签(label)对应一个项集,节点中的数目(count)表示项集的支持度计数,标签为  $\{0\}$  表示空节点, nodes 表示子节点集合. Trie 树节点的定义如表 2 所示:

Table 2 Trie Tree Node  
表 2 Trie 树节点

Attribute	Description	Sample
label	The label of the node(itemset)	{1,2}
count	The support count of the itemset	20
nodes	Collection of child nodes	{1,2,3},{1,2,4}

令空节点为第 0 层,则第  $l$  层表示所有的  $l$ -项集,且第  $l$  层的节点数为  $C_l^I$ ,  $T_D$  总的节点数为  $\sum_{l=1}^{|I|} C_l^I = 2^{|I|} - 1$ . 如第 1 层的节点表示所有的 1-项集,第 1 层的节点数为  $C_1^I = 4$ ,总的节点数为 15. 第 1,2 层的项集支持度序列为 [40,30,20,40,20,20,30,20,20].

2.3 差分隐私

差分隐私(differential privacy)<sup>[19-21]</sup>是一种较新的隐私模型,能保证数据库中任意一行的改变(添加或删除)都不会使分析结果发生大的变化. 非交互式差分隐私定义如下:

1)  $\epsilon$ -差分隐私( $\epsilon$ -differential privacy)<sup>[19]</sup>. 随机算法  $A$  满足差分隐私约束,如果任意的 2 个事务数据集  $D$  和  $D'$ ,  $|D\Delta D'|=1$ ,对于所有输出数据集  $O$ ,

式(1)成立:

$$Pr[A(D)=O] \leq e^{\epsilon} Pr[A(D')=O], \quad (1)$$

其中,  $|D \Delta D'|=1$  表示事务数据集  $D$  和  $D'$  只有一条记录不同, 本文称为邻近事务数据集.

拉普拉斯机制<sup>[19]</sup>与指数机制<sup>[33]</sup>是满足差分隐私约束的 2 种标准方法, 它们均依赖于函数的全局敏感度.

2) 全局敏感度(global sensitivity). 任意函数  $f: D \rightarrow \mathbb{R}^d$ , 其敏感度为

$$\Delta f: \max_{D_1, D_2} \|f(D_1) - f(D_2)\|. \quad (2)$$

$D_1$  与  $D_2$  最多相差一条数据记录. 函数的敏感度越低, 差分隐私机制能保留越多的信息, 精确度越高. 根据输出结果是否为实数可将函数  $f$  分为实数型函数与非实数型函数, 分别对应拉普拉斯机制与指数机制.

3) 拉普拉斯机制. 对于实数型函数, Dwork 等人<sup>[19]</sup>提出的拉普拉斯机制有 3 个输入: 数据库  $D$ 、函数  $f$ 、隐私参数  $\epsilon$ . 添加的噪音服从概率密度函数为

$$p(x | \lambda) = \frac{1}{2\lambda} e^{-|x|/\lambda} \quad (3)$$

的拉普拉斯分布. 其中  $\lambda = \Delta f / \epsilon$ , 即噪音的量与数据库无关, 只与函数的敏感度及隐私参数有关. 文献<sup>[19]</sup>给出如下定理:

**定理 1.** 对于函数  $f: D \rightarrow \mathbb{R}^d$ , 随机算法

$$A(D) = f(D) + \text{Laplace}(\Delta f / \epsilon) \quad (4)$$

满足  $\epsilon$ -差分隐私要求.

4) 指数机制. 对于非实数型函数, Mcsherry 与 Talwar<sup>[33]</sup>提出指数机制. 以概率  $p$  从结果集  $R$  中选择结果  $r$  作为输出, 其中  $q(D, r)$  为结果  $r$  在数据集  $D$  上的效用性值.  $q$  的敏感度为

$$\Delta q = \max_{r, D_1, D_2} |q(D_1, r) - q(D_2, r)|, \quad (5)$$

其中,  $p \propto \exp(\epsilon q(D, r) / 2 \Delta q)$ .

**定理 2**<sup>[33]</sup>. 给定效用性函数  $q: (D \times R) \rightarrow \mathbb{R}$ , 对于数据库  $D$ , 随机算法  $A(D, q)$  以概率  $p$  返回  $r$ , 则  $A(D, q)$  满足  $\epsilon \times \Delta q$ -差分隐私要求.

差分隐私具有序列组合 (sequential composition) 性质<sup>[34]</sup>. 对于一个计算序列, 其中每个计算都满足差分隐私约束, 独立运行每个计算后将结果组合同样满足差分隐私约束.

**定理 3**<sup>[34]</sup>.  $A_i$  满足  $\epsilon_i$ -差分隐私, 对于数据集  $D$ , 计算序列  $A_i(D)$  满足  $\sum_i \epsilon_i$ -差分隐私.

## 2.4 压缩感知<sup>[35-38]</sup>

对于一维信号  $x \in \mathbb{R}^{N \times 1}$ , 其稀疏度为  $k$  (即含有  $k$  个非零值), 可以找到它的一维观测值:  $y = \Phi x$ , 其

中,  $\Phi \in \mathbb{R}^{M \times N}$ .  $\Phi$  的每一行可以看作是一个传感器, 它与信号的乘积保留了信号的部分信息, 这部分信息足以代表信号, 并能找到一个算法高概率恢复原信号  $x$ . 然而一般的信号  $x$  本身不是稀疏的, 需要找到某个稀疏基矩阵  $\Psi$  对  $x$  进行稀疏表示,  $x = \Psi s$ , 其中  $s$  为稀疏系数 ( $s$  只有  $k$  ( $k \ll N$ ) 个非零值). 则感知方程为

$$y = \Phi x = \Phi \Psi s = \Theta s, \quad (6)$$

其中,  $\Theta = \Phi \Psi$  为传感矩阵, 解出  $s$  的逼近值  $s'$ , 则原信号  $x' = \Psi s'$ .

上述分析表明, 压缩感知主要包括 3 部分:

1) 信号的稀疏表示

信号的稀疏表示是压缩感知的重要前提和理论基础, 可简单理解为信号中非零数目较少. 经典的稀疏化方法有离散余弦变换 (discrete cosine transform, DCT)、傅立叶变换 (fast Fourier transform, FFT)、离散小波变换 (discrete wavelet transform, DWT) 等.

2) 信号的观测矩阵

观测矩阵  $\Phi$  ( $n \times N, n \ll N$ ) 用来对  $N$  维原始信号  $x$  观测得到  $n$  维观测信号  $y$ . 为保证能够从观测信号  $y$  准确恢复原信号  $x$ , 观测矩阵需要满足一定的限制: 观测基矩阵  $\Phi$  与稀疏基矩阵  $\Psi$  的乘积满足 RIP 条件 (约束等距性)<sup>[36]</sup>. 其等价条件是测量矩阵  $\Phi$  和稀疏基矩阵  $\Psi$  不相关. Candes 和 Tao 等人<sup>[35-36]</sup>证明: 独立同分布的高斯随机矩阵可以成为通用的压缩感知测量矩阵.

3) 信号恢复算法

压缩感知理论能够基于式(6)求解稀疏系数  $s$ , 然后将稀疏度为  $k$  的信号从  $n$  维观测信号  $y$  中无损失地恢复出来. 解码最直接的方法是基于  $l_0$  范数 ( $0$  范数) 求解如下最优化问题:

$$\min_s \|s\|_{l_0} \quad y = \Phi \Psi s. \quad (7)$$

当观测矩阵  $\Phi$  满足约束等距性 (restricted isometry property, RIP) 条件时,  $l_0$  最小范数与  $l_1$  最小范数可得到相同的解, 则式(7)可转化为  $l_1$  最小范数下的最优化问题:

$$\min_s \|s\|_{l_1} \quad y = \Phi \Psi s. \quad (8)$$

得到稀疏系数  $s$  的逼近值  $s'$ , 则原信号  $x' = \Psi s'$ .

压缩感知的恢复算法主要分两大类: 1) 贪婪算法, 包括匹配追踪算法<sup>[39]</sup>、正交匹配追踪 (orthogonal matching pursuit, OMP) 算法<sup>[40]</sup>等; 2) 凸优化算法, 包括最小角度回归法<sup>[41]</sup>等.

## 2.5 效用性

为评估含噪 Trie 树  $NT_b'$  的效用性, 针对频繁

项集挖掘任务,采用真正(true positive,  $TP$ )、假正(false positive,  $FP$ )、准确率( $Accuracy$ )度量  $NT_b^I$  的效用性. 给定正整数  $k$ , 定义  $F_k(D)$  为原数据  $T_b^I$  的 Top  $K$  频繁项集的集合,  $F_k(\tilde{D})$  为结果数据  $NT_b^I$  的 Top  $K$  频繁项集的集合.

真正: 既在  $F_k(D)$  中又在  $F_k(\tilde{D})$  中的频繁项集的数目:  $TP = |F_k(D) \cap F_k(\tilde{D})|$ ;

假正: 不在  $F_k(D)$  中, 在  $F_k(\tilde{D})$  中的项集数目:  $FP = |F_k(\tilde{D}) - F_k(D) \cap F_k(\tilde{D})|$ .

准确率:  $Accuracy = |F_k(D) \cap F_k(\tilde{D})| / |F_k(\tilde{D})|$ .

3 事务数据发布策略 TDPS

隐私保护事务数据发布的核心问题是在保护隐私的同时保留足够多的数据效用性. 本文提出的发布策略 TDPS 结合压缩感知理论与 Trie 树结构, 满足差分隐私约束.

本节在算法 1 中首先给出 TDPS 的整体框架. TDPS 的输入为: 事务数据库  $D$ 、隐私参数  $\epsilon$ . 返回满足差分隐私约束的 Trie 树  $NT_b^I$ .

算法 1.  $TDPS(D, \epsilon)$ .

输入: 事务数据集  $D$ 、差分隐私参数  $\epsilon$ ;

输出: 含噪音的 Trie 树  $NT_b^I$ .

① 基于  $I$ , 构建  $D$  的完整 Trie 树:  $T_b^I \leftarrow BuildTrie(root, I, D)$ ; /\* 3.1 节描述了构建的详细过程,  $root$  为根节点 \*/

② 按层遍历获得所有项集的支持度序列:  $S_b^I$ ;

③ 基于压缩感知理论, 对  $S_b^I$  添加满足差分隐私约束的噪音得到:  $NS_b^I \leftarrow AddNoise(S_b^I)$ ; /\* 3.2 节详细说明了添加噪音 3 种方法 \*/

④ 根据  $NS_b^I$  更新  $T_b^I$ , 得到  $NT_b^I$ ;

⑤ 返回满足差分隐私约束的 Trie 树  $NT_b^I$ .

算法 1 中包含 2 个子过程, 分别为算法 2 与算法 3, 下面分 2 节详细介绍每个子过程.

3.1 构建 Trie 树

Trie 树的构建过程中, 如果只考虑构建事务数据库  $D$  的 Trie 树  $T_D$ , 则会泄露个体隐私, 因为  $T_D$  仅覆盖了  $D$  已有的项集, 无法保证项集合  $I$  中所有可能的项集均能以非零概率出现在  $T_D$  中, 不满足差分隐私要求. 例如, 如果表 1 新增加一条事务  $T_{51} = \{2, 3\}$  得到  $D'$ , 尽管可以对其支持度计数添加噪音, 但是无法从结构上满足差分隐私要求, 因为  $T_{D'}$  比  $T_D$  多出节点  $\{2, 3\}$ .

本文基于  $I$ , 构建关于  $D$  的完整 Trie 树  $T_b^I, T_b^I$

覆盖了  $I$  的所有可能项集, 从结构上满足差分隐私要求. 对于表 1 所示的示例, 构建的完整 Trie 树如图 2 所示:

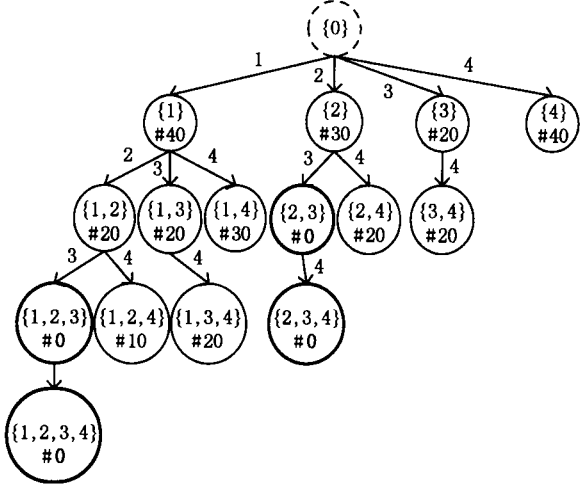


Fig. 2 The complete Trie tree about  $D$  based  $I$ .

图 2 基于  $I$  关于  $D$  的完整 Trie 树  $T_b^I$

图 2 表明, 不存在  $D$  中的项集  $\{2, 3\}, \{2, 3, 4\}, \{1, 2, 3\}, \{1, 2, 3, 4\}$  也包含在  $T_b^I$  中. 算法 2 给出了 Trie 树的详细构建过程,  $BuildTrie$  是一个递归过程, 其输入为: 当前处理节点、项集合  $I$ 、事务数据库  $D$ .

3.2 噪音添加策略

通过算法 2, 得到基于  $I$  关于  $D$  的完整 Trie 树  $T_b^I$ , 本节讨论如何生成满足差分隐私约束的  $NT_b^I$ .

算法 2.  $BuildTrie(root, I, D)$ .

输入: 当前节点  $root$ 、项集  $I$ 、事务数据库  $D$ ;

输出: Trie 树.

/\*  $max$  函数返回对应项集的项的最大编号, 空节点返回 0 \*/

①  $i = max(root.label) + 1$ ;

② for all  $i \leq max(I)$ ;

③ 创建新节点  $node$ , 标签:  $root.label \cup i$ ;

④ 支持计数:  $Count(root.label \cup i, D)$ ;

⑤  $root.Add(node)$ ;

⑥  $i++$ ;

⑦ for each  $subNode \in root.nodes$ ;

⑧  $BuildTrie(subNode, I, D)$ ;

⑨ end for

⑩ end for

添加噪音之前, 需要对  $T_b^I$  进行如下预处理, 得到项集支持度序列  $S_b^I$  的观测值  $y$ .

首先, 按层遍历  $T_b^I$ , 得到关于  $I$  的所有项集在  $D$  的支持度计数序列  $S_b^I$ , 序列长度为  $N = 2^{|I|} - 1$ .

其次, 基于压缩感知理论, 通过某个观测基  $\Phi$

与  $S_b^l$  的乘积得到观测值  $y$ . 即

$$y = \Phi S_b^l. \tag{9}$$

其中,  $\Phi \in \mathbb{R}^{n \times N} (n \ll N), y \in \mathbb{R}^n$ . 如果  $S_b^l$  非稀疏, 则需要引入稀疏基矩阵  $\Psi \in \mathbb{R}^{N \times N}$ :

$$y = \Phi \Psi S_b^l. \tag{10}$$

通过恢复算法, 如正交匹配追踪算法 OMP, 能从观测信号  $y$  恢复原始信号, 即原项集支持度计算  $S_b^l$ . 然而, 如果不对  $y$  添加噪音或对恢复过程进行随机化处理, 恢复得到的  $S_b^l$  会泄露个体隐私, 不满足差分隐私约束. 因此, 基于压缩感知理论, 本文提出两种添加噪音的方法: 第 1 种是基于拉普拉斯机制, 即 TDPS\_LP, 对观测信号  $y$  添加噪音; 第 2 种基于指数机制, 即 TDPS\_EP, 对正交匹配追踪算法进行随机化处理, 使之满足差分隐私约束, 算法 3 对 TDPS\_EP 进行了描述. 2 种方法的流程图分别如图 3 与图 4 所示:

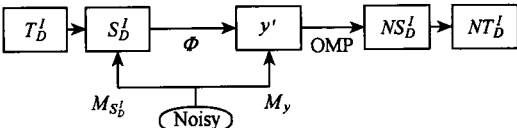


Fig. 3 TDPS\_LP flow graph.  
图 3 TDPS\_LP 流程图

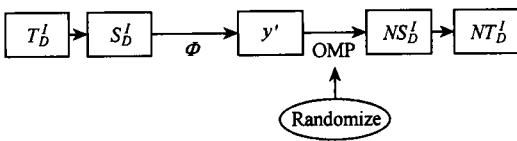


Fig. 4 TDPS\_EP flow graph.  
图 4 TDPS\_EP 流程图

**算法 3.**  $TDPS\_EP(y, \epsilon, s)$ .

输入: 信号  $y$ 、隐私参数  $\epsilon$ 、稀疏系数  $s$ ;  
输出: 含噪 Trie 树  $NS_b^l$ .

- ① 初始化:  
 $r^0 = y, c^0 = 0, \Lambda^0 = \emptyset, l = 0, \epsilon' = \epsilon/s$ ;
- ② 循环作步骤③~⑤, 直到  $l \leq s$ ;
- ③ 计算:  $h^l = \Phi^T r^l$ ;
- ④ 选择:

$$\Lambda^{l+1} = \Lambda^l \cup \left\{ \text{return } j \text{ with } p' \propto \exp\left(\frac{\epsilon'}{2\Delta_q} q(h^l, |h^l(j)|)\right) \right\};$$

- ⑤ 更新:  
 $c^{l+1} = \min_z \{ \|y - \Phi z\|_2 : \text{supp}(z) \subseteq \Lambda^{l+1} \};$   
 $r^{l+1} = y - \Phi c^{l+1};$

$l = l + 1$ ;

⑥ 返回:  $NS_b^l = c^{s+1}$ .

下面分别讨论 TDPS\_LP 与 TDPS\_EP 的详细过程.

3.2.1 TDPS\_LP

基于拉普拉斯机制, 有如下基本方法对观测信号  $y$  添加噪音:

1) 对  $S_b^l$  添加噪音

称该方法为  $M_{S_b^l}$ , 对原始信号  $S_b^l, M_{S_b^l}$  使用拉普拉斯机制直接对其添加噪音, 即

$$LS_b^l = S_b^l + Lap\left(\frac{\Delta}{\epsilon}\right)^N, \tag{11}$$

其中,  $\Delta = \max_{D_1, D_2} \|S_{D_1}^l, S_{D_2}^l\|_1$  为  $S_{D_1}^l$  与  $S_{D_2}^l$  的最大 1 范式距离. 则  $y' = \Phi \times LS_b^l$ . 这种方法可以描述为式(12):

$$M_{S_b^l}(\Phi, LS_b^l) = \Phi \times LS_b^l = \Phi \times \left( S_b^l + Lap\left(\frac{\Delta}{\epsilon}\right)^N \right). \tag{12}$$

**引理 1.**  $M_{S_b^l}$  对观测信号  $y$  的期望平方误差 (expected squared error, ESE) 为

$$ESE_{S_b^l} = \frac{2\Delta^2}{\epsilon^2} \sum_{i=1}^n \sum_{j=1}^N \Phi_{i,j}^2. \tag{13}$$

证明. 令随机变量  $X$  服从分布  $Lap\left(\frac{\Delta}{\epsilon}\right)$ , 均值为 0.

$$ESE_{S_b^l} = E\left[\frac{1}{N} \sum_i (M_b(\Phi, LS_b^l)_i - (\Phi S_b^l)_i)^2\right] = \frac{1}{N} E\left[\sum_i (\Phi X)^2\right] = E\left[\sum_{i=1}^n \sum_{j=1}^N (\Phi_{i,j} X_j)^2\right]. \tag{14}$$

基于期望的线性性, 则:

$$ESE_{S_b^l} = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^N \Phi_{i,j}^2 \sum_{j=1}^N E[X_j^2]. \tag{15}$$

因为服从均值为 0 的拉普拉斯分布, 则方差:

$$D(X) = E(X^2) - [E(X)]^2 = E(X^2) = \frac{2\Delta^2}{\epsilon^2}. \tag{16}$$

则:

$$ESE_{S_b^l} = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^N \Phi_{i,j}^2 \sum_{j=1}^N E[X_j^2] = \frac{2\Delta^2}{\epsilon^2} \sum_{i=1}^n \sum_{j=1}^N \Phi_{i,j}^2. \tag{17}$$

证毕.

可见, 期望平方误差与观测基矩阵中元素的平方和成正比.

2) 对观测信号  $y$  添加噪音

称该方法为  $M_y$ , 与  $M_{S_b^l}$  方法不同,  $M_y$  直接对

观测信号  $y$  添加噪音, 观测基矩阵  $\Phi$  的敏感度为  $\Delta' = \max_j \sum_i |\Phi_{i,j}| \Delta$ , 即  $\Phi$  绝对值和最大的列.  $M_y$  可表述如下:

$$y' = M_y(\Phi, S_b^l) = \Phi \times S_b^l + Lap\left(\frac{\Delta'}{\epsilon}\right)^n. \quad (18)$$

引理 2.  $M_y$  对观测信号  $y$  的期望平方误差为

$$ESE_y = \frac{2n\Delta'^2}{\epsilon^2} = \frac{2n\Delta^2}{\epsilon^2} \max_j \sum_i \Phi_{i,j}^2. \quad (19)$$

证明. 同引理 1.

通过  $M_y$  或  $M_{S_b^l}$  得到满足差分隐私约束的观测信号  $y'$ , 可通过 OMP 求解原信号的近似  $NS_b^l$ . 文献 [42] 对于含噪信号有如下定理:

定理 4. 精确和稳定恢复的 RIP 条件<sup>[42]</sup>考虑模型:

$$y = \Phi S_b^l + w, \quad (20)$$

其中,  $w$  是满足差分隐私约束的噪音. 如果  $\Phi$  满足  $\delta_{2k} \leq 0.4931$  的 RIP 条件, 则优化问题式 (9) 或式 (10) 的解  $NS_b^l$  满足:

$$\|NS_b^l - S_b^l\|_1 \leq O(\|w\|_2). \quad (21)$$

定理 4 表明, 只要  $\Phi$  满足 RIP 条件, 则恢复对于含有噪音的观测的误差是稳定的. 当  $w=0$  时 (即没有添加任何噪音) 且  $S_b^l$  是  $k$  稀疏的, 则得到的解是精确解, 但这种解会泄露个体的隐私. 当  $w \neq 0$  时 (即对观测信号添加噪音), 则误差的上界为  $O(\|w\|_2)$ , 因此噪音越小, 数据的效用性越高, 也容易泄露用户隐私.

3.2.2 TDPS\_EP

基于拉普拉斯加噪方法 TDPS\_LP 对观测信号  $y$  添加噪音后得到含噪观测信号  $y'$ , 然后通过确定性的恢复算法 OMP 得到含噪原始信号.

基于指数机制的加噪方法与之不同, TDPS\_EP 处理干净的观测信号, 通过随机化 OMP 恢复算法, 使恢复的信号满足差分隐私约束.

噪音添加策略 TDPS\_EP 与 OMP 只在第 4 步 (选择) 不同, OMP 选择内积最大的列, 而 TDPS\_EP 以概率  $p' \propto \exp\left(\frac{\epsilon'}{2\Delta_q} q(h^l, |h^l(j)|)\right)$  选择  $\Phi$  的第  $j$  列, 其中  $q(\cdot, \cdot)$  为效用函数,  $q(h^l, |h^l(j)|) = \alpha |h^l(j)|$ , 即为内积的绝对值与缩放因子  $\alpha$  的乘积.  $q(\cdot, \cdot)$  的敏感度为

$$\Delta_q = \alpha \cdot \max_{j, y \oplus \bar{y} = 1} |(\Phi^T y)_j - (\Phi^T \bar{y})_j|. \quad (22)$$

定理 5. 加噪方法 TDPS\_EP 满足  $\epsilon \times \Delta_q$ -差分隐私.

证明. 基于指数机制, 令  $\epsilon' = \epsilon/s$ , 每次循环均

满足  $\epsilon' \times \Delta_q$ -差分隐私. 共循环  $s$  次, 根据定理 3,

TDPS\_EP 满足  $\sum_{i=1}^s (\epsilon' \times \Delta_q) = \epsilon \times \Delta_q$ -差分隐私.

证毕.

4 实验分析

本节通过实验评估事务数据发布策略 TDPS 的效用性. 实验对比了 3 种加噪方法对数据集效用性的影响, 首先通过 TDPS 得到  $NT_b^l$ , 然后在  $NT_b^l$  进行频繁项集挖掘任务, 评价指标为: 真正、假正和准确率.

通过误差  $Error = \|S_b^l - NS_b^l\|_2 / \|S_b^l\|_2$  评估加噪数据与原始数据的差异, 并通过误差  $Error$  的对比验证了文献 [22] 提出的 CM 方法不适用于事务数据发布.

实验中本文提出的 3 种加噪方法分别表示如下: 1) 对原始信号  $S_b^l$  加噪音 (TDPS\_LP\_Singal); 2) 对观测信号 (结果)  $y$  加噪音 (TDPS\_LP\_Result); 3) 对恢复过程加噪音 (TDPS\_EP). 其中前 2 种方法基于拉普拉斯机制, 第 3 种方法基于指数机制.

CM<sup>[22]</sup> 基于拉普拉斯机制, 添加噪音的量为  $e \propto Lap(\sqrt{s}/\epsilon)$ , 其中  $s$  为稀疏度, 即事务中包含项的个数;  $\epsilon$  为隐私参数.

实验数据集采用真实数据集 MSNBC, 项集  $I$  从中随机抽取 10 个, 即  $|I|=10$ , 实验数据详细信息如表 3 所示. 其中  $|D|$  表示事务个数,  $|I|$  表示总的项集数,  $\max |T|$  和  $\text{avg} |t|$  分别表示事务的最大长度和平均长度. MSNBC 数据集按时间顺序记录了用户访问的 URL 类别, 本文将其转变为事务数据集, 每个事务包含用户所有访问过的 URL 类型, 忽略了时间先后信息.

Table 3 Experiment Dataset

表 3 实验数据集

Dataset	$ D $	$ I $	$\max  T $	$\text{avg}  t $
MSNBC	989 818	10	10	1.304 4

实验中参数设置如表 4 所示.

通过发布策略 TDPS 的第 1 步与第 2 步获得所有项集的支持度序列:  $S_b^l$ , 因为  $S_b^l$  是非稀疏的, 本文通过小波基变换得到稀疏信号, 稀疏度为 20. 基于拉普拉斯机制与指数机制本文提出了 3 种不同的加噪方法.

Table 4 Parameter Setting for Experiments  
表 4 实验参数设置

Methods	Parameter	Value
Top K Frequent Itemset	$k$	20,40,60,80,100,120,140,160,180,200
Laplace Mechanism	$\epsilon$	0.001,0.005,0.008,0.01,0.05,0.1,0.5,1.1,1.5,2,5,7,9
Exponential Mechanism	$\epsilon$	0.1,0.3,0.5,0.7,0.9,1,3,5,7,9,11,13,15
	$\alpha$	0.000 01
Compressive Sensing	Sparsity	20
	Rows of $\Phi$	800

4.1 基于拉普拉斯机制

实验对比了基于拉普拉斯机制的加噪方法:对原始信号加噪音;对观测结果加噪音;CM<sup>[22]</sup>.

图 5 给出了这 3 种方法的加噪数据与原始数据的误差比较. 图 5 表明,随着隐私参数  $\epsilon$  的增大,误差 *Error* 越小. 特别地,当  $\epsilon \geq 0.05$  时,本文提出的 2 种加噪方法的误差趋于稳定,且都非常小,说明隐私参数对误差的影响较小. 满足  $\epsilon$  差分隐私保护时,  $\epsilon$  越小,加入的噪声越多,隐私保护的级别越高. 在实际应用中可设置较小的隐私参数值,保持数据效用性同时能有效保护数据隐私. 同时从图 5 可以看出, CM 的误差高于本文提出的方法,验证了 CM 不能直接应用于事务数据的发布. 主要原因是对每条事务数据处理时加入了过多的噪音,并严重破坏了事务与事务之间项集的联系. 尽管 CM 方法不适用于事务数据的发布,但它对于实时统计流数据的发布能取得非常好的效果<sup>[22]</sup>. 由于 CM 方法破坏了事务之间项的联系,不适用于事务数据的发布,因此下面的实验中不再对比 Top K 频繁项集的效用性.

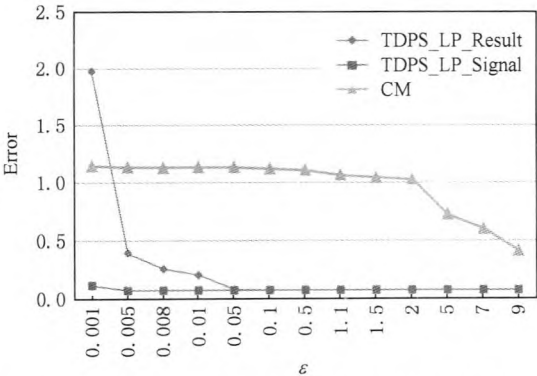


Fig. 5 Error comparison on different nosiyy methods based on laplace mechnism.

图 5 基于拉普拉斯机制加噪方法的误差比较

表 5 显示了 Top K 频繁项对 TDPS\_LP\_Singal 与 TDPS\_LP\_Result 的效用性的影响. 对生成的结果数据集进行 Top K 频繁项集挖掘任务,设置隐私参数  $\epsilon=1.1$ ,其中,  $R$  表示对观测信号(结果)加噪,

$S$  表示对原始信号加噪.  $TP, FP, Accuracy$  分别表示真正、假正、准确率,详见 2.5 节. 表 5 的结果表明,这 2 种加噪方法能够为 Top K 频繁项集挖掘任务提供足够多的信息. Top100 频繁项集准确率能保持在 80%左右,但随着  $k$  的增加,由于加入噪音的原因,准确率呈下降趋势,但 Top200 也能保持在 63%左右.

Table 5 Utility for Top K Frequent Itemset Mining with TDPS\_LP( $\epsilon=1.1$ )

表 5 TDPS\_LP 的 Top K 频繁项集挖掘效用性( $\epsilon=1.1$ )

$k$	TP		FP		Accuracy	
	R	S	R	S	R	S
20	19	20	1	0	0.95	1
40	38	38	2	2	0.95	0.95
60	54	57	6	3	0.9	0.95
80	67	66	13	14	0.84	0.83
100	80	78	20	22	0.80	0.78
120	90	92	30	28	0.75	0.77
140	94	101	46	39	0.67	0.72
160	108	109	52	51	0.68	0.68
180	117	117	63	63	0.65	0.65
200	126	125	74	75	0.63	0.63

表 6 显示了隐私参数  $\epsilon$  对 TDPS\_LP\_Singal 与 TDPS\_LP\_Result 的效用性的影响. 设置  $k=60$ ,即

Table 6 Utility for Top K Frequent Itemset Mining with TDPS\_LP( $k=60$ )

表 6 TDPS\_LP 的 Top K 频繁项集挖掘效用性( $k=60$ )

$\epsilon$	TP		FP		Accuracy	
	R	S	R	S	R	S
0.01	40	56	20	4	0.67	0.93
0.05	54	57	6	3	0.90	0.95
0.1	52	56	8	4	0.87	0.93
0.5	54	55	6	5	0.90	0.97
1.1	55	57	5	3	0.92	0.95
1.5	55	57	5	3	0.92	0.95



考察 Top60 频繁项集的效用性. 随着  $\epsilon$  的增加, 效用性越好, 然而其隐私等级越低.

总的来说, 基于拉普拉斯机制的加噪方法, 不仅能提供强的隐私保护, 还能为数据分析提供足够的信息.

4.2 基于指数机制

基于指数机制的加噪方法 TDPS\_EP 通过随机化 OMP 使之满足差分隐私要求.

图 6 显示了隐私参数  $\epsilon$  对 TDPS\_EP 误差的影响. 图 6 表明, 隐私参数  $\epsilon$  对误差的影响非常大, 随着  $\epsilon$  的增大, 其误差越小, 其隐私等级越低. 结合图 5 与图 6, 可以发现 TDPS\_LP 优于 TDPS\_EP, 其原因是随机化 OMP 过程中的随机选择过程添加了更多的噪音.

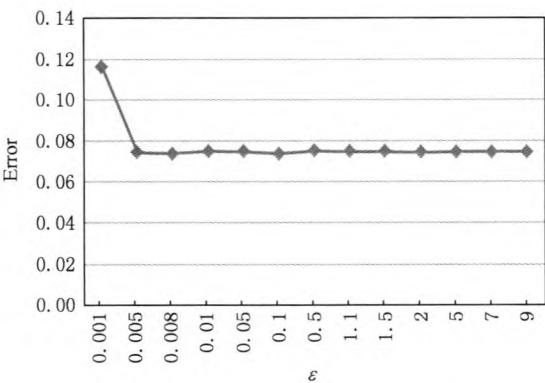


Fig. 6 the error of TDPS\_EP VS.  $\epsilon$  value  
图 6  $\epsilon$  对 TDPS\_EP 误差的影响

表 7 给出了 TDPS\_EP 的效用性受 Top K 频繁项集影响的情况, 设置隐私参数  $\epsilon=1$ . 可以发现, 随着  $k$  的不断增加, Top K 频繁项集准确率下降得很快.

Table 7 Utility for TopK Frequent Itemset Mining with TDPS\_EP( $\epsilon=1$ )

表 7 TDPS\_EP 的 TopK 频繁项集挖掘 ( $\epsilon=1$ )

K	TP	FP	Accuracy
20	17	3	0.85
40	22	18	0.55
60	30	30	0.5
80	34	46	0.425
100	39	61	0.39
120	46	74	0.38
140	52	88	0.371
160	60	100	0.375
180	65	115	0.36
200	70	130	0.35

表 8 给出了 TDPS\_EP 的效用性受隐私参数  $\epsilon$  影响的情况, 设置  $K=60$ . 表 8 表明, 随着  $\epsilon$  的增加, 效用性越好.

Table 8 Utility for TopK Frequent Itemset Mining with TDPS\_EP( $K=60$ )

表 8 TDPS\_EP 的 TopK 频繁项集挖掘 ( $K=60$ )

$\epsilon$	TP	FP	Accuracy
0.1	18	42	0.3
0.5	19	41	0.32
1	28	32	0.47
5	49	11	0.82
9	51	9	0.85
15	53	7	0.88

图 7 显示了 TDPS\_EP 效用性受隐私参数  $\epsilon$  与  $k$  影响的总体情况. 观察  $\epsilon=1$  与  $\epsilon=5$  的曲线, 发现准确率有一个较大的跳跃, 这是因为在 OMP 随机化过程中, 正确选择内积最大列的概率与  $\epsilon$  呈指数关系. 当  $\epsilon \geq 5$  时, 对准确率的影响已不明显.

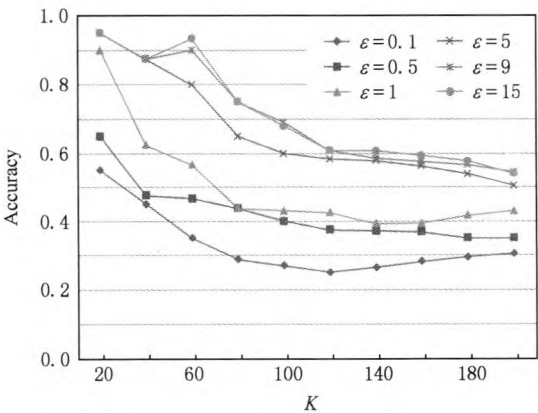


Fig. 7 Utility for TDPS\_EP VS.  $\epsilon$  value and K value  
图 7  $\epsilon$  与 K 对 TDPS\_EP 效用性的影响

5 结束语

本文通过分析发现目前已有的隐私保护事务数据发布方法大多是基于分组的匿名模型, 但该类模型安全性不强且结果数据效用性较差. 针对事务数据隐私保护发布的数据安全性与效用性不足, 提出一种有效的满足差分隐私约束事务数据发布策略 TDPS, 该策略不仅能保持很高的数据效用性, 而且能对数据库中的用户提供强的隐私保护. 基于拉普拉斯机制与指数机制, 我们分析了 3 种不同的加噪方法. 在实验环境中验证了 TDPS 的效用性.

下一步工作中, 1) 将讨论如何快速地构建完整 Trie 树, 以提升算法效率; 2) 针对 Trie 树中父节点

支持度计数大于等于子节点支持度计数的特点,优化  $NT_b^d$  的噪音支持度计数,使其满足一致性约束<sup>[29,32,43]</sup>将利于提高频繁项集挖掘的效用性。

## 参 考 文 献

- [1] Zhou Shuigeng, Li Feng, Tao Yufei, et al. Privacy preservation in database applications: A survey [J]. Chinese Journal of Computers, 2009, 32(5): 847-861 (in Chinese)  
(周水庚, 李丰, 陶宇飞, 等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报, 2009, 32(5): 847-861)
- [2] Xu Yong, Qin Xiaolin, Yang Yitao, et al. A QI weight-aware approach to privacy preserving publishing data set [J]. Journal of Computer Research and Development, 2012, 49(5): 913-924 (in Chinese)  
(徐勇, 秦小麟, 杨一涛, 等. 一种考虑属性权重的隐私保护数据发布方法[J]. 计算机研究与发展, 2012, 49(5): 913-924)
- [3] Sweeney L. K-anonymity: A model for protecting privacy [J]. International Journal of Uncertainty Fuzziness and Knowledge Based Systems, 2002, 10(5): 557-570
- [4] Liu Yubao, Huang Zhilan, Fu A W C, et al. A data privacy preservation method based on lossy decomposition [J]. Journal of Computer Research and Development, 2009, 46(7): 1217-1225 (in Chinese)  
(刘玉葆, 黄志兰, 傅慰慈, 等. 基于有损分解的数据隐私保护方法[J]. 计算机研究与发展, 2009, 46(7): 1217-1225)
- [5] Machanavajjhala A, Kifer D, Gehrke J, et al. L-diversity: Privacy beyond k-anonymity [J]. ACM Trans on Knowledge Discovery from Data (TKDD), 2007, 1(1): 1-3
- [6] Fung B C M, Wang K, Chen R, et al. Privacy preserving data publishing: A survey of recent developments [J]. Computing, 2010, 5(4): 1-53
- [7] Hong Y, Vaidya J, Lu H, et al. Differentially private search log sanitization with optimal output utility [C] //Proc of the 15th Int Conf on Extending Database Technology. New York: ACM, 2012: 50-61
- [8] He Y, Naughton J F. Anonymization of set-valued data via top-down, local generalization [J]. Proceedings of the VLDB Endowment, 2009, 2(1): 934-945
- [9] Xu Y, Wang K, Fu A W C, et al. Anonymizing transaction databases for publication [C] //Proc of the 14th Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 767-775
- [10] Ghinita G, Kalnis P, Tao Y. Anonymous publication of sensitive transactional data [J]. IEEE Trans on Knowledge and Data Engineering (TKDE), 2011, 23(2): 161-174
- [11] Loukides G, Gkoulalas D A, Malin B. COAT: Constraint-based anonymization of transactions [J]. Knowledge and Information Systems, 2011, 28(2): 251-282
- [12] Terrovitis M, Mamoulis N, Kalnis P. Local and global recoding methods for anonymizing set-valued data [J]. The VLDB Journal, 2011, 20(1): 83-106
- [13] Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data [J]. Proceedings of the VLDB Endowment, 2008, 1(1): 115-125
- [14] Xu Y, Fung B, Wang K, et al. Publishing sensitive transactions for itemset utility [C] //Proc of the 8th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2008: 1109-1114
- [15] Gkoulalas Divanis A, Loukides G. Utility-guided clustering-based transaction data anonymization [J]. Trans on Data Privacy, 2012, 5(1): 223-251
- [16] Cao J, Karras P, Raïssi C, et al.  $\rho$ -uncertainty: Inference-proof transaction anonymization [J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 1033-1044
- [17] Kifer D. Attacks on privacy and deFinetti's theorem [C] //Proc of the ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2009: 127-138
- [18] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets [C] //Proc of the IEEE Symp on Research in Security and Privacy. Piscataway, NJ: IEEE, 2008: 111-125
- [19] Dwork C, Mcsherry F, Nissim K. Calibrating noise to sensitivity in private data analysis [C] //Proc of the 3rd Int Conf on Theory of Cryptography. Berlin: Springer, 2006: 265-284
- [20] Dwork C. Differential privacy in new settings [C] //Proc of the 21st Annual ACM-SIAM Symp on Discrete Algorithms Society for Industrial and Applied Mathematics (SODA 2010). New York: ACM, 2010: 174-183
- [21] Dwork C. Differential privacy: A survey of results [C] //Proc of the 5th Int Conf on Theory and Applications of Models of Computation. Berlin: Springer, 2008: 1-19
- [22] Li Y D, Zhang Z, Winslett M, et al. Compressive mechanism: Utilizing sparse representation in differential privacy [C] //Proc of the 10th Annual ACM Workshop on Privacy in the Electronic Society. New York: ACM, 2011: 177-182
- [23] Aho A V, Ullman J D, Hopcroft J E. Data structures and algorithms [J]. Reading, MA: Addison-Wesley, 1983
- [24] Aoe J I, Morimoto K, Sato T. An efficient implementation of trie structures [J]. Software: Practice and Experience, 1992, 22(9): 695-721
- [25] Pietracaprina A. Mining frequent itemsets using patricia tries [C] //Proc of the 3rd IEEE ICDM Workshop on Frequent Itemset Mining Implementations. Piscataway, NJ: IEEE, 2003: 786-798
- [26] Bodon F. Surprising results of trie-based FIM algorithms [C] //Proc of the 4th IEEE ICDM Workshop on Frequent Itemset Mining Implementations. Piscataway, NJ: IEEE, 2004: 865-878
- [27] Bodon F. A fast apriori implementation [C] //Proc of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations. Piscataway, NJ: IEEE, 2010: 654-667

[28] Ghinita G, Tao Y, Kalnis P. On the anonymization of sparse high-dimensional data [C] //Proc of the 25th Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2008: 715-724

[29] Hay M, Li C, Miklau G, et al. Accurate estimation of the degree distribution of private networks [C] //Proc of the 9th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2009: 169-178

[30] Xiao X, Wang G, Gehrke J. Differential privacy via wavelet transforms [C] //Proc of the IEEE 26th Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2010: 225-236

[31] Mcsherry F, Mironov I. Differentially private recommender systems: building privacy into the net [C] //Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 627-636

[32] Chen R, Mohammed N, Fung B C M, et al. Publishing set-valued data via differential privacy [J]. Proceedings of the VLDB Endowment, 2011, 4(11): 1087-1098

[33] Mcsherry F, Talwar K. Mechanism design via differential privacy [C] // Proc of the 48th Annual IEEE Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2007: 94-103

[34] Mcsherry F D. Privacy integrated queries: An extensible platform for privacy-preserving data analysis [C] //Proc of the 35th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2009: 19-30

[35] Candes E J, Tao T. Decoding by linear programming [J]. IEEE Trans on Information Theory, 2005, 51(12): 4203-4215

[36] Candes E J. The restricted isometry property and its implications for compressed sensing [J]. Comptes Rendus Mathematique, 2008, 346(9): 589-592

[37] Candes E J, Tao T. Near-optimal signal recovery from random projections: Universal encoding strategies [J]. IEEE Trans on Information Theory, 2006, 52(12): 5406-5425

[38] Donoho D L. Compressed sensing [J]. IEEE Trans on Information Theory, 2006, 52(4): 1289-1306

[39] Mallat S G, Zhang Z. Matching pursuits with time-frequency dictionaries [J]. IEEE Trans on Signal Processing, 1993, 41(12): 3397-3415

[40] Tropp J A, Gilbert A C. Signal recovery from random measurements via orthogonal matching pursuit [J]. IEEE Trans on Information Theory, 2007, 53(12): 4655-4666

[41] Efron B, Hastie T, Johnstone I, et al. Least angle regression [J]. The Annals of Statistics, 2004, 32(2): 407-499

[42] Mo Q, Li S. New bounds on the restricted isometry constant  $\delta_{2k}$  [J]. Applied and Computational Harmonic Analysis, 2011, 31(3): 460-468

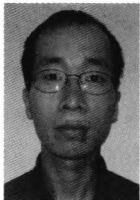
[43] Hay M, Rastogi V, Miklau G, et al. Boosting the accuracy of differentially private histograms through consistency [J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 1021-1032



**Ouyang Jia**, born in 1986. Received his PhD from Sun Yat-sen University. His major research interests include data mining and data privacy.



**Yin Jian**, born in 1968. Professor and PhD supervisor in Sun Yat-sen University. Senior member of China Computer Federation. His major research interests include data mining and machine learning.



**Liu Shaopeng**, born in 1984. Received his PhD from Sun Yat-sen University. His major research interests include text mining and topic model.

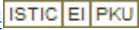


**Liu Yubao**, born in 1975. Received his PhD from Sun Yat-sen University. His major research interests include database system and data mining.

# 一种有效的差分隐私事务数据发布策略

作者：[欧阳佳](#)，[印鉴](#)，[刘少鹏](#)，[刘玉葆](#)，[Ouyang Jia](#)，[Yin Jian](#)，[Liu Shaopeng](#)，[Liu Yubao](#)

作者单位：[中山大学信息科学与技术学院 广州 510006](#)

刊名：[计算机研究与发展](#)

英文刊名：[Journal of Computer Research and Development](#)

年，卷(期)：2014, 51 (10)

本文链接：[http://d.g.wanfangdata.com.cn/Periodical\\_jsjyjfz201410007.aspx](http://d.g.wanfangdata.com.cn/Periodical_jsjyjfz201410007.aspx)