基于条件函数依赖的隐私保护模型*

陈伟鹤，陈　霖
（江苏大学 计算机科学与通信工程学院，江苏 镇江 212013）

摘　要：数据拥有者发布的数据中如果包含条件函数依赖会导致数据的隐私受到攻击，由条件函数依赖产生的属性间的关联会带来潜在的隐私泄露问题。针对现有的隐私保护方法均无法保护包含条件函数依赖的数据的隐私，形式化地定义了基于条件函数依赖的隐私攻击，提出了隐私保护模型 *l*-deduction 来对包含条件函数依赖的数据进行隐私保护；并设计了相应的匿名算法来实现 *l*-deduction 模型。理论分析和实验结果表明，该方法既能保护包含条件函数依赖的数据的隐私，又具有较小的信息损失度。

关键词：隐私保护；数据发布；条件函数依赖；*l*-deduction；信息损失
中图分类号：TP309　　**文献标志码**：A　　**文章编号**：1001-3695(2012)10-3838-04
doi:10.3969/j.issn.1001-3695.2012.10.062

Privacy-preserving model on data with conditional functional dependency

CHEN Wei-he, CHEN Lin
(*School of Computer Science & Telecommunication Engineering*, *Jiangsu University*, *Zhenjiang Jiangsu* 212013, *China*)

Abstract: There exists privacy threat in the publishing data that contains CFDs. This paper showed that the correlations among attributes by CFDs could bring potential vulnerability to privacy. This paper formalized the CFD-based privacy attack, proposed the privacy model *l*-deduction and developed an algorithm that achieved *l*-deduction. Beyond a theoretical treatment of the problem, the experiments also demonstrate the effectiveness of the proposed technique.

Key words: privacy-preserving; data publishing; conditional functional dependency(CFDs); *l*-deduction; information loss

0　引言

随着网络、数据库技术的快速发展，大量的个人电子信息被存储、发布，并且用于数据挖掘、知识发现等领域，但同时也会引发隐私泄露[1]。许多数据收集者为了不同的目的发布数据，但是他们往往面临一个问题，即如何在不泄露用户隐私的情况下发布数据。传统的做法是将待发布的数据表中含有个人身份标志的属性(IDs)(如姓名、身份证号、家庭住址等)隐藏，但研究表明这种做法仍会造成隐私泄露。研究[2]表明，通过邮编、性别、出生日期等非标志符属性对选民登记表和隐匿了个体标志的医疗信息表进行连接操作，超过 87% 的美国公民的身份都可以被唯一标志。那些非标志符属性被称做准标志符属性(QI-属性)。

为了防止这种链接攻击，人们提出了很多隐私保护规则，例如，*k*-anonymity[2] 和 *l*-diversity[3]。然而如果将条件函数依赖应用在这些发布的数据上，便会引起隐私泄露，被攻击者所利用。

例 1　假设表 1 中包含条件函数依赖 Φ_0：[CC = 0 *, PN]→ZIP，即在满足“CC = 0 *”这个条件下，若任意两条记录中的 PN 值相同，则这两条记录中的 ZIP 也相同。其中：CC 指的是国家代码(country code)；PN 指的是电话号码(phone)；ZIP 指的是邮编(zip code)。假设攻击者知道这个背景知识(Φ_0：[CC = 0 *, PN]→ZIP)，然后将 Φ_0 应用到 3-多样性表(即每个拥有相同 QI-值的组中包含至少三个不同的敏感值)中。由于在第二组中 3333333 所对应的 ZIP 值为“0120 *”，攻击者就可以将表 2 中 t_3[ZIP] 由“012 * *”修改为“0120 *”。值得注意的是，表 2 中 t_8[PN] 虽然也为 3333333，但其所对应的 ZIP 值不发生变化，这是因为 t_8 中的 CC = 4 *，不符合 Φ_0 中的条件“CC = 0 *”，因而不存在 PN→ZIP 的函数依赖。在基于条件函数依赖 Φ_0 的推理下得到的匿名表如表 3 所示，其中由 t_3 所构成的组仅仅满足 1-多样性[3]。

表 1　原始数据表

ID	QI			sensitive
name	sex	CC	ZIP	PN
Mike	F	01	01221	1111111
Rick	M	01	01220	2222222
Joe	M	01	01202	3333333
Jim	F	01	01201	1000001
Ben	F	01	01202	3333333
Bob	F	01	01203	2000001
Alice	M	44	01204	2000001
Eve	M	44	01205	3333333
Lily	M	44	01206	1000002

例 1 表明，若条件函数依赖这个背景知识被攻击者所利用，可能会引起隐私泄露。因此，本文提出一种隐私保护模型 *l*-deduction 来解决基于条件函数依赖(CFDs)的隐私泄露问题。

收稿日期：2012-02-15；**修回日期**：2012-04-15　　**基金项目**：江苏省教育厅自然科学基金资助项目(09KJB520003)；江苏大学高级人才启动基金资助项目(07JDG031)

作者简介：陈伟鹤(1974-)，男，江苏常州人，副教授，博士，主要研究方向为数据库安全、模型检测(chenweiheujs@gmail.com)；陈霖(1988-)，女，硕士，主要研究方向为隐私保护.

表 2　3-多样性表

	sex	CC	ZIP	PN
t_1	*	0 *	012 * *	1111111
t_2	*	0 *	012 * *	2222222
t_3	*	0 *	012 * *	3333333
t_4	F	0 *	0120 *	1000001
t_5	F	0 *	0120 *	3333333
t_6	F	0 *	0120 *	2000001
t_7	M	4 *	0120 *	2000000
t_8	M	4 *	0120 *	3333333
t_9	M	4 *	0120 *	1000002

表 3　条件函数依赖推理后的表

	sex	CC	ZIP	PN
t_1	*	0 *	012 * *	1111111
t_2	*	0 *	012 * *	2222222
t_3	*	0 *	0120 *	3333333
t_4	F	0 *	0120 *	1000001
t_5	F	0 *	0120 *	3333333
t_6	F	0 *	0120 *	2000001
t_7	M	4 *	0120 *	2000000
t_8	M	4 *	0120 *	3333333
t_9	M	4 *	0120 *	1000002

1　相关工作

近年来,面向数据发布中的隐私保护越来越受到关注,保护个人隐私的一个有效手段是数据匿名化[4],而实现数据匿名化的隐私规则有许多。Samarati 等人最早提出 k-匿名[2]的概念,它要求公布后的数据中存在一定数量的不可区分的个体,使攻击者不能判别出敏感值所属的具体个体,从而防止了个人隐私的泄密。针对 k-匿名存在的不足,提出了 l-多样性[3]匿名策略,它要求一个等价组中至少有 l 个不同的敏感值。在后续的相关研究中,根据不同的隐私保护要求,提出了一些新的概念(如 t-closeness[5]、(α, k)-anonymity[6]、(c, k)-safety)[7])。大部分的匿名策略都是通过泛化 QI-值的方法来达到隐私保护的,但它们都无法解决上文所提到的基于 CFD 的隐私泄露问题。

Martin 和 Rastogi 等人[7,8]是研究攻击者知道任意的元组间相互关系的隐私保护问题的先驱者。随后,Kifer[9]指出攻击者可以根据“消毒”的数据集推算出数据间的相互关系,这些推算出的相互关系使得数据集具有潜在的易损性。然而文献[7~9]中均没有提出防止基于函数依赖的隐私泄露的方法。后来 Wang 等人[10]提出了一种隐私保护模型来保护包含全函数依赖(full functional dependencies,FFD)的数据的隐私,但它们依然无法保护包含条件函数依赖 CFD 的数据集的隐私。

2　背景和概念

定义 1　条件函数依赖[11]。关系 R 上成立了的条件函数依赖 Φ 形式为(R:$X\rightarrow Y$,T_p),其中:

$X,Y\in \mathrm{attr}(R)$, $\mathrm{attr}(R)$表示 R 的属性集合;

$X\rightarrow Y$ 为嵌入 Φ 的函数依赖,称 X 决定 Y,或 Y 依赖 X(X 叫做决定因素);

T_p 为定义在属性集 $X\cup Y$ 上的模式组(pattern tableau),由若干条件模式元组 t_p 构成。对于 $X\cup Y$ 中的每个属性 A,模式元组 t_p 定义了 A 的取值 $t_p[A]$为定义域 $\mathrm{Dom}(A)$中的某个常数 a,或者任意值,用‘_’表示。

由定义可见,T_p 通过绑定属性 A 及其取值 $t_p[A]$指定了嵌入式函数依赖成立的条件,如例 1 中 Φ_0 表示的条件函数依赖为(R:[CC,PN] →ZIP, (0 *,_ | _)),即 Φ_0 可以表示成:([CC = 0 *, PN]→ZIP),或([CC,PN]→ZIP,T_{p_0})。其中 T_{p_0} 如表 4 所示。

表 4　模式组 T_{p_0}

T_{p_0}		
CC	PN	ZIP
0 *	_	_

泛化是实现匿名化所使用的最早也是最广泛的技术[12],即将待发布的数据表中的元组划分成若干个组(QI-groups),并把准标志符的属性值(QI-值)作泛化处理,比如将生日从年/月/日泛化成年/月,将民族从满、回等泛化成少数民族,将数据 2、4、7 泛化成区间[0,10]等。本文采用泛化技术。

定义 2　匹配。对于属性 A 的取值 η_1、η_2,η_1 与 η_2 匹配是指 $\eta_1=\eta_2$,或者 η_1 可以泛化成 η_2,或者 η_2 可以泛化成 η_1;或者 η_1、η_2 中取值存在‘_’,用 $\eta_1\approx\eta_2$ 表示。

定义 3　条件函数依赖语义[11]。对于定义在关系 R 上的条件函数依赖 $\Phi=(R:X\rightarrow A,t_p)$,如果实例 I 中任意两个元组 t_1、t_2 都满足以下条件:$t_1[X]=t_2[X]\approx t_p[X]$,则 $t_1[A]=t_2[A]\approx t_p[A]$,那么称 I 满足 Φ,记做 $I\models\Phi$。

与传统的函数依赖不同,单个元组也有可能违背条件函数依赖约束,如表 2 中的 t_7 不满足 Φ_0。

定义 4　QI-group 的条件函数依赖语义。给定原始数据实例 T 和条件函数依赖 Φ,T 所对应的匿名数据集为 T^*,T^* 被划分[13]成 n 个 QI-groups $G_1,G_2,\cdots,G_n$,亦即 $T^*=\cup_{i=1}^{n}G_i$;$T\models\Phi$,若 $G_i(1\leqslant i\leqslant n)$中的每一条元组均满足 Φ,则称 G_i 满足 Φ,记做 $G_i\models\Phi$。

本文用计数查询结果的误差来衡量信息损失度。设 Q 是一个计数查询,$Q(T)$和 $Q(T^*)$分别为对原始表 T 和匿名后的表 T^* 进行查询得到的结果[11],则相对误差 $\mathrm{error}=\left|\frac{Q(T)-Q(T^*)}{\max\{Q(T).\delta\}}\right|$,设 δ 为数据集 T 的大小的 0.5%。

3　隐私保护模型

3.1　基于 CFD 的隐私攻击

定义 5　l-多样性规则[3]。如果一个 QI-group 至少包含 l 个“well-presented”敏感属性值,则该 QI-group 满足 l-diversity 规则。如果一个匿名数据 T^* 中的所有 QI-group 都满足 l-diversity 规则,则 T^* 满足 l-diversity 规则。

给定条件函数依赖 $\Phi=(R:A\rightarrow B,t_p)$,若匿名数据集 T^* 中不同的 QI-group 中至少存在两个元组 t_1 和 t_2,满足 $t_1[A]=t_2[A]=a$,且 $t_1\models\Phi$,$t_2\models\Phi$,那么攻击者可以推断出 $t_1[B]=t_2[B]$,此时便有可能存在基于条件函数依赖 Φ 的隐私泄露。假设 $t_1[B]=b_1^*$,$t_2[B]=b_2^*$,攻击者可以将 b_1^* 和 b_2^* 进行“相交”运算,本文用符号表示泛化值之间的相交运算。

定义 6　泛化值之间的相交运算 $\cap^*$。给定位于同一泛化树(图 1)中的两个泛化值 b_1^* 和 b_2^*,不妨设 b_1^* 所在的层数大于或等于 b_2^* 所在的层数,则 $b_1^*\cap^* b_2^*=b_2^*$。

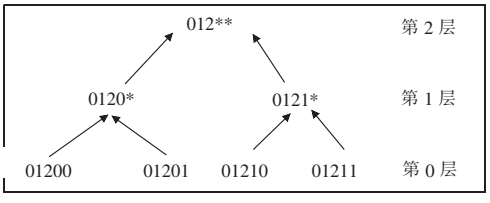

图 1　ZIP 泛化树

定义 7　基于 CFD 的 QI-groups 间的相交运算 $\cap_C$ 和减法运算 $-_C$。给定条件函数依赖 $\Phi=(R:A\rightarrow B,t_p)$和原始数据实例 T,T^* 为 T 的匿名数据,G_1、G_2 为 T^* 中的两个 QI-groups,$T\models\Phi$,$G_1\models\Phi$,$G_2\models\Phi$,令 $G_{12}=\{t_1\mid t_1\in G_1,\exists t_2\in G_2\ \text{s.t.}\ t_1[A]=t_2[A]\}$。那么 $G_1\cap_C G_2$ 为元组集 C,$\forall t\in G_{12}$,$\exists t'\in C$,

使得

$\forall A' \in A, t'[A'] = t[A]$;

$\forall B' \in B, t'[B'] = G_1[B] \cap^* G_2[B]$;

另外，$G_1 -_C G_2 = G_1 - G_{12}$。

定义 8 基于 CFD 的隐私攻击。给定原始数据集 T（T 包含 QI-属性 QI 和敏感属性 S）及条件函数依赖 $\Phi = (R: A \rightarrow B, t_p)(A, B \subseteq \text{QI} \cup S, A \subseteq \text{QI})$，$T^*$ 为 T 的匿名数据，T^* 满足 l-多样性。若 $\exists$ QI-groups $G_1, G_2 \in T^*$，使得：(a) $G_1 \cap_C G_2$；(b) $G_2 \cap_C G_1$；(c) $G_1 -_C G_2$；(d) $G_2 -_C G_1$ 中至少有一个非空或不满足 l-多样性，则 Φ 会造成基于 CFD 的隐私泄露。

3.2 基于 CFD 的隐私保护模型

定义 9 l-deduction。给定原始数据集 T，T 所对应的匿名数据集为 T^*，T^* 被划分成 n 个 QI-groups $G_1, G_2, \cdots, G_n$，亦即 $T^* = \cup_{i=1}^{n} G_i$。S_i 为 G_i 所对应的不重复的敏感值的集合。若

$\forall G_i, G_j \subseteq T^*$，如果 $|S_i \cap S_j| \neq 0$，那么 $|S_i \cap S_j| \geqslant l$；且

$\forall G_i, G_j \subseteq T^*, |S_i - S_j| \geqslant l$；

则称 T^* 满足 l-deduction。

4 算法

给定原始数据集 D（D 中的每条记录都满足条件函数依赖 Φ_D），本文采用相邻组相交机制来对敏感值进行分组，使得这些组都满足 l-deduction 中的条件。该机制要求所有的组 $G_1, G_2, \cdots, G_n$ 形成一条链，G_i 只与 G_{i+1} 和 G_{i-1} 相交，与其他 QI-groups 均不相交。

相邻组相交机制的实现步骤如算法 1 所示。

算法 1 实现相邻组相交机制的算法

输入：数据集 D，隐私参数 l；

输出：链集合 $\text{LS} = \{G_1, G_2, \cdots, G_n\}$。

步骤：

a) 根据敏感值将所有元组哈希到桶中，每个值对应一个桶；

b) 将所有桶按照其大小降序排列得到 $B_1, B_2, \cdots, B_m$，设桶 B_i 的大小为 b_i；

c) 将每个桶 B_i 都划分成两个部分 U_{B_i} 和 U_{B_i}'，U_{B_i} 由 B_i 中的 $b_i/2$ 个元素构成，$U_{B_i}' = B_i - U_{B_i}$；

d) $j = 1$；

e) $G_j = U_{B_1} \cup U_{B_2} \cup \cdots \cup U_{B_l}$；

f) 将前 l 个桶更新为 $U_{B_1}', \cdots, U_{B_l}'$；

g) 循环；

h) $G_{j+1} = U_{B_1}' \cup U_{B_2}' \cup \cdots \cup U_{B_l}' \cup U_{B_{l+1}} \cup U_{B_{l+2}} \cup \cdots \cup U_{B_{2l}}$；

i) 直到剩余的桶的个数 k 小于等于 l；

j) $G_{j+1} = U_{B_1}' \cup U_{B_2}' \cup \cdots \cup U_{B_l}' \cup U_{B_{l+1}}' \cup U_{B_{l+2}} \cup \cdots U_{B_k}$。

接下来，介绍构造 QI-groups 的算法（算法 2）。这些 QI-groups 要满足以下两个条件：(a) 满足 l-deduction 隐私规则；(b) 尽量减小由泛化带来的信息损失。

算法 2 构造匿名数据 T^*，使其满足 l-deduction

输入：数据集 T，隐私参数 l；

输出：数据集 T^*。

步骤：

a) $\text{GS} \leftarrow \{\}$，$i = 1$；{GS：存储 QI-groups 的集合}；

b) 将 T 根据敏感值划分成若干个子集 $T_1, \cdots, T_t$，使得每个子集中敏感值的振幅控制在一定范围之内；

c) 循环；

d) 对子集 T_i 根据相邻组相交机制进行划分得到链集合 TS_{T_i}；

e) $\text{GS} = \text{GS} \cup \text{TS}_{T_i}$；

f) i++；

g) 直到 $i > t$，即所有子集 $T_i (1 \leqslant i \leqslant t)$ 均根据步骤 d) 处理完毕；

h) 根据 GS 和泛化技术构造 T^*。

5 实验结果

本章通过实验来分析算法的性能，并将其与文献[14]中的多维 k-匿名泛化算法 Mondrian 进行比较。实验的硬件环境为 Intel Core2 Duo 2.9 GHz CPU，2 GB RAM，操作系统为 Microsoft Windows XP。本文提出的算法和 Mondrian 均采用 Java 实现。实验所采用的数据集为 Census 数据集，该数据集包含 500 000 个美国人口的个人信息；同时只保留 8 个属性：age，gender，education，marital，race，work class，country，salary-class。本文创建数据集 Sal，Sal 的敏感属性为 salary-class。本实验设计一个条件函数依赖为（[country = 0 *，salary-class] → work class）。

5.1 执行时间分析

图 2 给出了 l 的变化对执行时间的影响，从图 2 可知，程序的执行时间随 l 值的增大而缩短，这是因为 l 值越大，需要构造的 QI-groups 越少，程序便会更早终止。另外，本文算法的执行效率比 Mondrian 稍高，这是因为根据相邻组相交机制对数据集进行划分得到的链集合 $\text{LS} = \{G_1, G_2, \cdots, G_n\}$ 中 $G_i (1 < i < n)$ 与 G_{i+1} 和 G_{i-1} 相交，为了使匿名后的数据满足 l-deduction 规则的条件，G_i 必须满足 $2l$-多样性规则。而 Mondrian 要求匿名后得到的 QI-groups 均满足 l-多样性规则。由此可知，本文算法需要构造的 QI-groups 较少，因而执行时间较短。

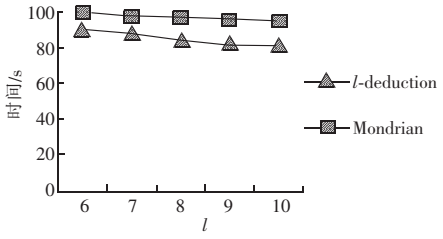

图 2 l 变化下的执行时间

5.2 信息损失分析

在进行信息损失分析时，采用查询语句：

SELECT count(*) FROM data WHERE pred(A_1^q) AND ⋯ AND pred(A_i^q) AND pred(A^s)

其中：A_i^q 表示 QI-属性；A^s 表示敏感属性；pred(A) 表示属性 A 上的谓词，谓词根据查询维度 qd 和查询选择率 s 这两个参数生成。特别地，给定 $qd \in [2,5]$，$s \in (0,1)$，创建集合 S_A，S_A 中包含 A^s 和随机选取的 $qd - 1$ 个 QI-属性；$\forall A \in S_A$，设 pred(A) 为 $A \in I$，I 为 A 上的一个区间，I 占 A 的 $s^{1/qd}$。最后，$\forall A' \in S_A$，pred(A) 为 $A' = *$。令 $qd \geqslant 2$ 且 $A^s \in S_A$，这样就可以检测出匿名数据的信息损失度大小。

图 3 给出查询维度 qd（查询属性的个数）对信息损失度的影响，当查询维度变大时算法具有更好的查询精确度。这是因为查询谓词包含一个属性域或是覆盖该域的 $s^{1/qd}$ 的区间，如果 s 不变，$s^{1/qd}$ 会随着 qd 的增大而变大。这表明了查询区间越大，则相对误差越小，如文献[13]所述。图 4 给出了查询选择率从 2% 增加到 10% 的相对误差，查询选择率越大，相对误差

越小。从实验中发现,信息损失度至多为 0.32,证明了算法的有效性。图 3 和 4 表明,*l*-deduction 和 Mondrian 的信息损失度相差不大,但 *l*-deduction 可以保护包含条件函数依赖的数据的隐私,而 Mondrian 尚未考虑到数据中隐藏的条件函数依赖。

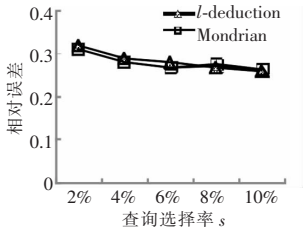

图 3　*s* 变化下的信息损失

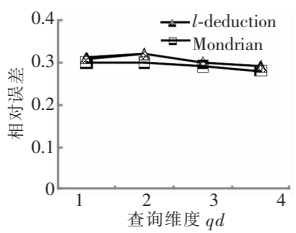

图 4　*qd* 变化下的信息损失

6　结束语

为了保护包含条件函数依赖的数据在发布过程的隐私,本文提出一种隐私保护模型 *l*-deduction,并设计了一个构造 *l*-deduction 模型的算法,执行该算法得到的匿名数据集不仅满足 *l*-deduction,而且有效地减少了泛化处理所带来的信息损失。今后的工作将考虑研究包含其他类型的函数依赖(如部分函数依赖(PFD))的数据的隐私保护问题。

参考文献:

[1] YUAN Yong-bin, YANG Jing, ZHANG Jian-pei, *et al*. Evolution of privacy-preserving data publishing[C]//Proc of International Conference on Anti-Counterfeiting, Security and Identification. 2011:34-37.

[2] SWEENEY L. *k*-anonymity:a model for protecting privacy [J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,2002,10(5):557-570.

[3] MACHANAVAJJHALA A, GEHRKE J, KIFER D, *et al*. *l*-diversity: privacy beyond *k*-anonymity[C]//Proc of the 22nd International Conference on Data Engineering. New York:ACM Press,2006:24-35.

[4] SHABTAI A, ELOVICI Y, ROKACH L. A survey of data leakage detection and prevention solution [M]. New York:Spinger,2012:47-68.

[5] LI Ning-hui, LI Tian-cheng. *t*-closeness: privacy beyond *k*-anonymity and *l*-diversity[C]//Proc of the 23rd International Conference on Data Engineering. 2007:106-115.

[6] WONG R C W, LI Jiu-yong, FU A W, *et al*. (α, k)-anonymity:an enhanced *k*-anonymity model for privacy-preserving data publishing [C]//Proc of the 12th ACM International Conference on Special Interest Group on Knowledge Discovery and Data Mining. New York: ACM Press,2006:754-759.

[7] MARTIN D J, KIFER D, MACHANAVAJJHALA A, *et al*. Worst-case background knowledge for privacy-preserving data publishing[C]//Proc of the 23rd International Conference on Data Engineering. New York:ACM Press,2007:126-135.

[8] RASTOGI V, SUCIU D, HONG S. The boundary between privacy and utility in data publishing[C]//Proc of the 33rd International Conference on Very Large Data Bases. [S. l.]:VLDB Endowment,2007: 531-542.

[9] KIFER D. Attacks on privacy and definetti's theorem[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York:ACM Press,2009:127-138.

[10] WANG Hui, LIU Rui-lin. Privacy-preserving publishing data with full functional dependencies[C]//Proc of the 15th International Conference on Database Systems for Advanced Applications. Berlin: Springer-Verlag,2010:176-183.

[11] FAN Wen-fei, GEERTS F, JIA Xi-bei, *et al*. Conditional functional dependencies for capturing data inconsistencies [J]. ACM Trans on Database Systems,2008, 33(2):1-48.

[12] 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.

[13] XIAO Xiao-kui, TAO Yu-fei. Anatomy: simple and effective privacy preservation[C]//Proc of the 32nd International Conference on Very Large Data Bases. [S. l.]:VLDB Endowment,2006:139-150.

[14] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Mondrian multidimensional *k*-anonymity[C]//Proc of the 22nd International Conference on Data Engineering. Washington DC:IEEE Computer Society, 2006:25-35.