doi:10.11959/j.issn.1000-436x.2016076

面向 DaaS 应用的数据集成隐私保护机制研究

周志刚，张宏莉，余翔湛，李攀攀
（哈尔滨工业大学计算机网络与信息安全技术研究中心，黑龙江 哈尔滨 150001）

摘 要：云计算的出现为多个数据拥有者进行数据集成发布及协同数据挖掘提供了更广阔的平台，在数据即服务模式（DaaS, data as a service）下，集成数据被部署在非完全可信的服务运营商平台上，数据隐私保护成为制约该模式应用和推广的挑战性问题。为防止数据集成时的隐私泄露，提出一种面向 DaaS 应用的两级隐私保护机制。该隐私保护机制独立于具体的应用，将数据属性切分到不同的数据分块中，并通过混淆数据确保数据在各个分块中均衡分布，实现对数据集成隐私保护。通过分析证明该隐私保护机制的合理性，并通过实验验证该隐私保护机制具有较低的计算开销。

关键词：云安全；数据即服务；隐私保护；匿名
中图分类号：TP393 **文献标识码：**A

Research on data integration privacy preservation mechanism for DaaS

ZHOU Zhi-gang, ZHANG Hong-li, YU Xiang-zhan, LI Pan-pan
(Research Center of Computer Network and Information Security Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: The emergence of cloud computing provides a broader platform for multiple data owners to make integrated data publishing and collaborative data mining. In data-as-a-service (DaaS) model, integrated data was deployed in a certain cloud platform with an untrusted service provider ta privacy leakage has become the challenge hindering application and popularization of DaaS model. For protecting privacy in the data integration stage, a two-layer privacy protection mechanism for DaaS-oriented application was given, which was independent ith the specific applications, partitioning data attributes into different parts. In addition, the corres ding fake data set was used to assure the balanced distribution of data in each part, which realized privacy protection of data integration. The experimental results indicate that the proposed strategy is feasible, simultaneously has the low computing overhead.

Key words: cloud security, data as a service, privacy protection, anonymity

1 引言

在目前的商业环境下，企业内部各部门甚至同类企业之间的数据共享已经成为制定决策、为用户提供高质量服务的基本需求，多个数据拥有者需要相互协作集成彼此的数据以实现数据共享。在此过程中有 2 个问题急需解决：1）对集成后数据的存储、维护及统计分析操作可能超出现有设备的载荷；2）集成后的数据含有更为丰富的知识，攻击者可能据此推导出其中的隐私数据。因此，在数据集成时，各数据提供者要对数据进行匿名化处理。云计算作为一种新型的数据操作方式为数据共享提供了一个强有力的软/硬件平台。有别于传统的以大型服务器为核心的计算模式，云计算以互联网及内部专用网为核心，采用虚拟化技术构建大规模数据中心，为云租户提供泛在网络信息共享、按需资

收稿日期：2015-06-10；**修回日期：**2015-12-29
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目（No.2011CB302605）；国家自然科学基金资助项目（No.61173144, No.60903166, No.61100188）
Foundation Items: The National Basic Research Program of China (973 Program) (No.2011CB302605), The National Natural Science Foundation of China (No.61173144, No.60903166, No.61100188)

源租用以及实际使用计费的新型服务模式。对云租户而言，云计算缓解了其一次性购买软/硬件的开销及其对数据存储管理维护的压力。然而当云租户将数据上载到云端，享受云计算所带来的巨大便利的同时，由于云端掌握了对云租户数据的最终控制权，这就不可避免地引入一些新型的挑战。

首先，在数据即服务（DaaS, data as a service）模式下，各云租户将集成后的业务数据上载到非完全可信的云服务提供商（CSP, cloud service provider）的平台上，之后对数据的存储及操作都在云租户非完全控制的环境中进行，CSP 可以直接查看数据信息导致云租户数据隐私泄露。即使数据在被上载到云端之前进行了匿名化处理，CSP 依然能够通过数据的分布信息采用数据统计挖掘的方式窃取云租户的数据隐私。

其次，由于数据中含有商业机密或隐私信息（如用户信用卡消费记录、医疗数据等），即使是需要进行数据共享的多方数据提供者，也必须在一定的匿名度（如 k-匿名）下实施数据共享，然而数据提供者对自己拥有局部数据的匿名化处理并不能保证集成后的数据也满足同样的匿名度要求。

第三，虽然 CSP 相对于云租户而言拥有无限的资源以及更安全的设备，但置于云端的云租户业务数据依然面对数据一致性、计算结果正确性等安全问题，例如，2011 年 3 月，谷歌邮箱爆发大规模的用户数据泄露事件，大约 15 万 Gmail 用户的邮件和聊天记录被删除。

针对上述问题，本文提出了多轮加细的匿名策略，将数据集进行逐轮加细划分，每一个轮次都选取当次全局信息增益最大的属性对数据进行划分，直至集成数据集不可再分，即达到各数据提供者商定的集成数据集的匿名门限。该策略防止了在多方进行数据集成时，非己方无法学习到比最终集成数据更多的知识；其次，通过定义半可信信誉等级和完全非可信信誉等级，根据 CSP 所处的信誉等级提出了一种面向 DaaS 应用的两级隐私保护机制。针对处于半可信信誉等级的 CSP，提出与应用无关的属性集划分策略，防止其根据各属性值的映射关系泄露数据隐私。对于处于完全非可信信誉等级的 CSP，提出分类索引的数据结构对云端返回结果的一致性及完整性进行验证。

2　相关工作

文献[1]给出了面向数据库应用的隐私保护综述，详细介绍了数据挖掘和数据发布中所用到的隐私保护技术。文献[2]主要从数据的机密性、数据的完整性、数据的完备性、查询隐私保护及访问控制策略这 5 个方面介绍了数据库服务安全与隐私保护方面的研究进展，其中，数据的机密性主要从基于加密和基于数据分布展开分析。数据加密虽然可以有效地防止数据隐私泄露，但在 DaaS 模式下，对数据加/解密的处理效率相对较低。同态加密技术[3]虽然可以直接对密文数据进行操作，但却需要大量的计算需求。

针对数据加密隐私保护的不足，研究者提出在数据明文的情况下，通过对敏感数据匿名化的方式防止隐私泄露。Sweeney 等[4]提出的 k-匿名原则，要求所发布的数据表中的每一条记录不能区分于其他 $k-1$ 条记录。(a,k)-匿名[5]对此进行了改进，保证每一个等价类中的数据，与任一敏感属性值相关的记录百分比不高于a 。l-diversity 保证每一个等价类的敏感属性至少有 l 个不同的值，t-closeness 在 l-diversity 基础上，考虑了敏感属性的分布问题，要求所有等价类中敏感属性值的分布尽量接近该属性的全局分布。

针对安全的多方计算领域，Clifton 等[6]提出分布式 k-匿名算法（DkA, distributed k-anonymity），该算法假设在垂直划分的数据环境下同一条记录有唯一的全局标识，数据集成的各方都只拥有部分属性的数据，利用可交换加密在通信过程中隐藏原始信息，再构建完整的匿名表判断是否满足匿名门限来实现数据隐私保护。但该算法的时间开销很大，文献[6]中对 defacto benchmark adult 数据集匿名化需要 12 天。文献[7]开发了一个针对关系数据计数、并、交、笛卡儿积 4 种典型操作的安全数据多方数据集成工具。Mohammed 等[8,9]基于分类树结构使用数据泛化技术实现数据集成各方的数据隐私保护，但集成后数据的信息损失较高，具体的信息损失度与数据集相关。文献[10]提出一种可追责计算框架，该框架可以实现数据集成的各方相互验证。扩展研究[11~13]意在为不同的集成数据挖掘任务设计安全协议，然而这些方法的计算开销过于昂贵。

针对云数据隐私，文献[14,15]通过完备格设计了属性分块树形结构，该树形结构中每一个实线框表示属性被分割的一个合理状态。文献[16]通过定义机密限制和属性可见请求分割数据集并采用分组匿名的方式保护数据隐私，但需要应用领域专家

事先建立属性约束规则集。文献[17, 18]提出(k,a,b,g)隐私保护机制，通过定义属性集合的隐私约束对数据进行垂直分割，使每一个数据分块中的属性都不会导致数据组合隐私泄露，并引入(a,b,g)3 个层次的组合均衡化概念，确保每个数据分块物理存储中各种数据切片出现的概率尽可能的平均，从而保护 DaaS 数据隐私，但与文献[16]类似，属性隐私约束集的构建需要领域专家的指导，且伪数据的生成、识别和混淆数据的重构都需要在可信第三方的协作下完成。

上述研究中，基于安全的多方计算技术提供了多方数据集成的隐私保证，但已知的这些方法[6~13]的计算开销过大，以至于难以在实际场景中应用，本文提出多轮加细的匿名策略，数据集成各方在满足预设的匿名门限条件下，通过多轮信息交互逐步细分数据集。在云数据隐私保护方面，文献[17]与本文的研究最为相近，都是使用数据分割技术防止数据隐私泄露，但采用的方法完全不同。本文提出一种与应用无关的属性划分策略，通过构造属性辨识集对属性集划分，使各数据分块内的属性组合不会导致隐私泄露。在此基础之上，本文还提出分类索引树数据结构，使云租户有能力验证 CSP 返回结果集的正确性及完整性。

3 问题定义

3.1 多租户外包数据集成架构

数据集成通过较以前更完备的数据集更好地制定决策，为用户提供高质量的服务为目标，多个数据拥有者将各自的数据进行融合。为了方便讨论，首先形式化地定义云租户所拥有的数据集为一个四元组$T(U,A,F,Class)$。其中，U 为数据对象集，即$U=\{x_1,x_2,\text{L},x_n\}$，每个 x_i 称为一个对象；A 为属性集 $A=\{a_1,a_2,\text{L},a_m\}$；$F$ 为 U 和 A 之间的关系集 $F=\{f_k:U\to V_k\}$，V_k 为 a_k 的值域；$Class$ 为决策属性。为了简化模型，本文以$T_1(U_1,A_1,F_1,Class_1)$、$T_2(U_2,A_2,F_2,Class_2)$2 个云租户数据集成为例，假设 T_1、T_2 具有相同的记录集且记录的属性集无交集，即$U_1=U_2$，$Class_1=Class_2$，$A_1\text{ I }A_2=\varnothing$。

多租户外包数据集成架构如图 1 所示，该系统主要由 2 个云租户(T_1,T_2)、云服务提供商(CSP)2 类实体组成。

租户 T_1 和租户 T_2 通过记录集的 ID 将各自的数据集进行集成，然后将集成的数据集上载至云端，以方便在全局数据集上实施数据挖掘，更好地制定决策。但由于数据中通常含有隐私信息，数据集成后所形成的数据集必须确保数据隐私安全。方案 1)先实施数据集成再匿名化以保护数据隐私，但各个云租户会在数据集成的过程中获取到其他云租户所持数据集的隐私信息；方案 2)先对各自云租户的数据集匿名化再进行数据集成，然而在部分属性集上满足匿名原则的数据集在数据集成后可能形成新的准标识符（QID, quasi-identifier），而新生成的 QID 不一定满足匿名原则，从而泄露隐私信息。鉴于以上 2 个方案所暴露的安全问题，本文提出多轮加细的匿名策略，首先将数据集泛化为同一个等价类，然后对数据集进行加细，每轮细化需要从所有的候选属性中选举一个当前信息增益最大的属性对数据集进行细化，直至满足在不违背匿名原则的

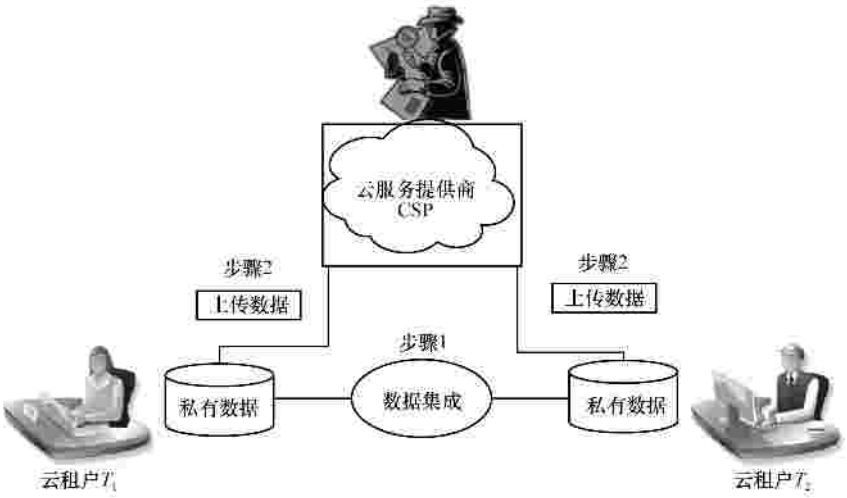

图 1 多租户外包数据集成架构

前提下达到数据集的最大细化，详细内容见第 4 节。这里假设所有的云租户都是理性的，即云租户在数据集成过程中不会隐藏自身高信息增益的属性来迫使其他租户共享更多的数据信息。至于云租户是恶意的系统模型，可以通过博弈理论来惩罚违规的云租户，使所有参与数据集成的云租户都遵守数据共享协议，从而将恶意系统模型降解为理性云租户系统模型。关于博弈理论文献[19]有详细的描述，再此不赘述。

3.2　威胁模型

由于将集成的数据集上载到云端，云租户把数据的最终控制权交由 CSP。租户数据的命运取决于 CSP 的信誉，本文把 CSP 的信誉分为 3 个等级：可信、半可信及完全非可信。可信 CSP 是指其严格遵守服务级别协议（SLA, service level agreement）为租户提供相应的服务，不会窃取租户数据隐私或篡改、破坏数据的完整性；半可信信誉等级是指 CSP 遵守 SLA 为租户提供服务，但 CSP 会通过查询请求、业务背景知识及结果集推测出数据中包含的业务信息，并将其泄露给数据拥有者的竞争对手；完全非可信信誉等级是指 CSP 会背离 SLA，篡改或返回租户查询结果集的子集。

首先，云租户数据中往往含有商业机密或隐私信息，在 DaaS 模式下租户丧失了对其数据的控制权，半可信的 CSP 推导出租户数据的隐私信息并将其泄露给云租户的竞争对手，租户将受到巨大损失；其次，CSP 推出“pay-for-use”的计价方式，而云租户的请求需要消耗大量的资源，出于经济利益的驱动，CSP 为节省计算量可能在租户的子数据集上完成租户的查询请求。综上，本文建立半可信的 CSP 威胁模型和完全非可信的 CSP 是合理的。

3.3　安全数据外包集成

定义 1　安全数据外包集成(secure data outsourcing integration)。P 为数据集成的云租户集合 $P=\{P_1,P_2,\text{L},P_n\}$，$T_i$ 为云租户 P_i 所拥有的数据表，A_i 为 T_i 表中所包含的属性集合 $A_i=\{a_1,a_2,\text{L},a_k\}$，且 $\forall A_i,A_j,A_i \text{ I } A_j=\varnothing$。$T$ 为 n 个云租户数据集成后所形成的数据表，其中，$T=\bigcup_{i=1}^{n}T_i$。安全数据外包集成必须满足以下 3 个条件：1) 满足数据匿名要求，即要求集成后的数据表中的每一条记录不能区分于其他 $k-1$ 条记录；2) 参与数据集成的任意云租户 P_i 无法从数据集成的交互过程中学习到比最终集成数据表 T 更多的知识；3) CSP 无法从集成数据表 T 中推导出隐私信息或统计知识。

为了安全有效地防止前面所提的威胁模型泄露数据隐私，数据隐私保护策略应该同时满足以下 3 个方面的要求。

1) 零知识性：CSP 无法通过数据统计、数据背景攻击等推导出比集成的数据集 T 更多的知识。

2) 数据正确性和完整性：隐私保护策略能够使云租户有能力验证 CSP 返回结果集的正确性及完整性。

3) 高效性：在数据隐私保护策略框架内，云端服务器应该在可比的时间复杂度下完成租户的查询请求。

4　多轮加细的匿名策略

在多方数据集成时，为了防止租户学习到比最终集成数据表 T 更多的知识，本文采用多轮加细的匿名策略。首先给出多轮加细的匿名的相关概念。

定义 2　k-匿名(k-anonymity)。准标识符为 m 个属性联合起来能唯一标识表中的一类敏感信息或隐私记录，且其任一子集都不能唯一标识。设 QID 为数据表 T 中准标识符集合，$num(QID_i)$ 表示在 T 中第 i 个标识符所含属性的对应属性值相同的记录的个数。k-匿名要求对于 $\forall QID_i \in QID$，使 $num(QID_i)\geq k$，其中，k 为租户商定的匿名门限值。

如表 1 所示，Shared 属性集为 S_1 和 S_2 为公共属性，其中，ID 为记录的标识符，$Class$ 为记录的决策/类属性；属性 a_1、a_2、a_3 和 a_4 分别表示年龄、眼镜处方、眼睛流泪度和散光，其中，a_1、a_2 是 S_1 的本地属性，a_3、a_4 是 S_2 的本地属性，并且 S_1、S_2 各自的数据集都满足 2-匿名。

表 1　　多方源数据

Shared		Party S_1		Party S_2	
ID	*Class*	a_1	a_2	a_3	a_4
1	hard	young	normal	myope	yes
2	hard	old	reduced	myope	no
3	hard	young	nomal	myope	yes
4	none	young	reduced	hyperope	no
5	none	young	reduced	hyperope	no
6	soft	young	more	hyperope	no
7	none	young	reduced	hyperope	no
8	soft	young	more	hyperope	no
9	hard	old	reduced	myope	no

定义 3 **等价类**(equivalence partitioning)。在 $T(U,A,F,Class)$ 上，对于 $\forall B \subseteq A$，记 $R_B=\{(x_i,x_j)\,|\,f_k(x_i)=f_k(x_j)(a_k\in B)\}$，$R_B$ 是 U 上的等价类。

定义 4 **加细**(refinement)。在 $T(U,A,F,Class)$ 上，$B,C\subseteq A$，设 R_B、R_C 是 U 上的等价类，若 $R_B\subseteq R_C$，即 R_B 对 U 的每一个划分都含于 R_C 的某个划分中，称 R_B 是 R_C 的加细。

多轮加细匿名算法的主要思想是：数据集成的各方就自己所拥有的本地数据计算各属性的信息熵并公布最大的熵值进行比较，各方选出本轮全局熵值最大的属性。该属性的所有者基于上一轮的数据划分结果对其进行加细划分，若划分结果不违背数据匿名约束，则公布划分结果，否则直接进行下一轮，直至没有属性能在满足匿名约束的前提下对数据加细划分产生贡献。多轮加细匿名策略如算法 1 所示，算法的时间复杂度为 $O(n)$，其中，n 与数据集的属性个数相关。下面以租户 S_1 为例描述多轮加细匿名算法，其他租户端的算法类似。

算法 1 多轮加细匿名算法

输入 各用户数据初始划分 P /*将所有的数据集划归为一个等价类*/

$A_1=\{a_1,a_2,\text{L},a_m\}$ /* A_1 为租户 S_1 的属性集*/

匿名门限 k

参与数据集成租户的所有属性个数 q

输出 数据加细结果集 P

步骤

```
Refinement(P, A1, k, q) {
WHILE ( q > 0 ) {
  IF( A1 ≠ ∅ ) {
      租户 S1 选取 max(IGain(ai)) ，并将其值广播;
    }
    接收其他租户广播的局部信息增益最大的属性值，并选出本轮全局信息增益最大的属性值 max(IGain(ax)) ；
    IF( ax ∈ A1 ) {
      A1 = A1 − {ax} ;
      扫描 P 中所有的等价类，若用属性 ax 对 Pi 进行加细所形成的所有子类 Pii 的元素个数 num(Pii) ≥ k ，则 ax 可以对 Pi 加细；否则扫描 Pi + 1 ；
      广播加细结果 P ；
    }ELSE{
      接收其他租户广播的本轮加细结果 P；
      }
    q = q − 1 ；
  }
RETURN P;
}
```

算法 1 中属性 a_i **的信息增益** $IGain(a_i)=I(Class)-E(a_i)$。**其中，** $I(Class)$ **为数据集所含信息量** $I(Class)=-\sum_{|R_{class}|}\left(\frac{|v(|R_{class}|_i)|}{|T|}\text{lb}\frac{|v(|R_{class}|_i)|}{|T|}\right)$，**式中** $|T|$ **为数据表** T **中记录个数，** $|v(|R_{class}|_i)|$ **为数据表根据** $Class$ **属性划分的第** i **类记录的个数；** $E(a_i)$ **为属性** a_i **的熵，** $E(a)=\sum_{|R_a|}\frac{|v(|R_a|_i)|}{|T|}I(a)$。

分析 数据的匿名性是由算法本身保证的。多轮加细匿名算法对数据实施自顶向下逐步细化，在交互过程中，每轮具有全局信息增益最大属性的租户严格遵照匿名门限细化数据集，第 q 轮细化的最终结果就是集成数据表 T，且易知前 $q-1$ 轮的细化结果集都比 T 粗糙，即租户在交互过程中不可能学到比集成数据表 T 更多的知识。

5 面向 DaaS 应用的隐私保护机制

5.1 针对半可信云隐私保护机制

对云租户上传的数据对象，半可信的 CSP 可根据数据间的关联关系泄露数据隐私。与文献[10~12]不同，本节提出一种与应用无关的数据分割方法，无需领域专家事先建立约束规则集，而是根据属性对信息决策的重要性不同，运用属性超图消解法分割数据集。首先，给出数据分割的相关概念。

定义 5 **准标识符**(quasi-identifier)。在 $T(U,A,F,Class)$ 上，对于属性集 $B\subseteq A$，使 $R_B=R_A$，且 B 的任何真子集都使等式不成立，称 B 为 T 的准标识符。

定义 6 **属性辨识集**(attribute discernibility set)。$T(U,A,F,Class)$ 为信息系统，记 $\frac{U}{R_A}=\{[x_i]_A\,|\,x_i\in U\}$，$D([x_i]_A,[x_j]_A)=\{a_k\in A\,|\,f_k(x_i)\neq f_k(x_j)\}$，称 $D([x_i]_A,[x_j]_A)$ 为 $[x_i]_A$ 与 $[x_j]_A$ 的属性辨识集。称

$\boldsymbol{D} = (D([x_i]_A,[x_j]_A) \mid [x_i]_A,[x_j]_A \in \frac{U}{R_A})$ 为属性辨识矩阵。辨识矩阵是辨识集的全体，辨识集中的元素用于区别不同等价类的各种属性。

定义 7　属性超图(attribute hypergraph)。属性超图可以定义为一个二元组 (V,HE)，其中，V 为集成数据表 T 中全体属性的集合，HE 是超边的集合，每一条超边表示属性辨识矩阵 $\boldsymbol{D}$ 的一项。

通过辨识矩阵查找准标识符 B，使 $R_B = R_A$，由定义 5 可知，在辨识矩阵中识别准标识符是一个 NP 问题，这里采用属性超图消解法。在提取准标识符时，每次选取超图中最大公共子边中的属性集作为候选集，并删除所有含有候选集属性的超边，如此迭代，直至超图中不含有超边为止，最终将所有候选集作笛卡儿积。其具体的准标识符提取算法如算法 2 所示。

算法 2　准标识符提取算法

输入　数据表 T 的属性集合 $V = \{a_1, a_k, \cdots, a_m\}$

属性辨识矩阵 $\boldsymbol{HE}\left\{\sum_{i=1}^{k}\sum_{j=i+1}^{k} D[i][j]\right\}$

输出　B

步骤

```
Finding_ B(V, HE) {
  WHILE ( HE ≠ ∅ ) {
    c = 0 ; count = 0 ; e = 0 ; Temp = ∅ ; Edge = ∅ ;
    /*扫描辨识矩阵 D ,找到最大公共子边*/
    FOR each HE_k {
      FOR( i = 0; i < D.size; i++ ) {
        IF ( HE_k ⊂ HE_i ) {
          count++ ;
          Temp = Temp ∪ HE_i ;
        }
      }
      IF ( c < count ) {
        c = count ; e = k ; Edge = Temp ;
      }
    }
  HE = HE − Edge ;
  B = B × HE_e ;
  }
RETURN B ;
}
```

定义 8　属性划分(attribute fragment)。$T(U, A, F, Class)$ 为信息系统，$B_k\ (k \le r)$ 为属性极小集（r 为极小集总数），记 $C = \bigcap_{k\le r} B_k$，$K = \bigcup_{k\le r} B_k - C$，$I = A - \bigcup_{k\le r} B_k$。其中，$C$ 为核心属性集，K 为重要属性集，I 为不必要属性集。

分析　本文的数据分割策略满足数据隐私保护需求。

证明　证明分 3 步进行。1) 证明命题“$\text{if } |B| \ge 2$，$\bigcap_{k\le |B|} B_k \ne \varnothing$”，根据算法 2，这是显然的。2) 证明命题“若 a 是核心属性，则 $\exists x_i, x_j \in U$，$D([x_i]_A,[x_j]_A) = \{a\}$”，通过反证法，假设 $\exists x_i, x_j \in U$，$D([x_i]_A,[x_j]_A) = \{a\}$，即对于 $a \in D([x_i]_A,[x_j]_A)$，$|D([x_i]_A,[x_j]_A)| \ge 2$，存在 $B = \cup\{D([x_i]_A,[x_j]_A) - \{a\} \mid [x_i]_A \cap [x_j]_A = \varnothing\}$。因此，对于 $[x_j] \cap [x_j]_A = \varnothing$，存在 $B \cap D([x_j] \cap [x_j]_A) \ne \varnothing$，使 $R_B = R_A$。所以 $\exists C \subseteq B$ 使 C 为准标识符，但 $a \notin C$，这与假设矛盾，原命题得证。3) 证明命题“若 a 为核心属性，则 $R_{B-\{a\}} \ne R_B$”。根据步骤 2)，$\exists x_i, x_j \in U$ 使 $f_a(x_i) \ne f_a(x_j)$ 并且 $f_b(x_i) = f_b(x_j)$，因此 $(x_i, x_j) \in R_{A-\{a\}}$，$(x_i, x_j) \notin R_A$，即 $R_{A-\{a\}} \ne R_A$。又由于 $R_B = R_A, B \subseteq A$，所以 $R_{B-\{a\}} \ne R_B$。综上，B 是准标识符，而 $B-\{a\}$ 和 a 不构成准标识符。

证毕。

例 1　如图 1 所示，

$\frac{U}{R_{\{a_1,a_2,a_3,a_4\}}} = \{\{x_1, x_3\}, \{x_2, x_9\}, \{x_4, x_5, x_7\}, \{x_6, x_8\}\}$，

$\frac{U}{R_{Class}} = \{\{x_1, x_2, x_3, x_9\}, \{x_4, x_5, x_7\}, \{x_6, x_8\}\}$。

根据定义 6，辨识矩阵如下

$$\boldsymbol{D}_{Class} = \begin{pmatrix} \varnothing & \varnothing & \{a_2,a_3,a_4\} & \{a_2,a_3,a_4\} \\ \varnothing & \varnothing & \{a_1,a_3\} & \{a_1,a_2,a_3\} \\ \{a_2,a_3,a_4\} & \{a_1,a_3\} & \varnothing & \{a_2\} \\ \{a_2,a_3,a_4\} & \{a_1,a_2,a_3\} & \{a_2\} & \varnothing \end{pmatrix}$$

根据辨识矩阵 $\boldsymbol{D}_{Class}$，建立属性超图如图 2(a)所示。由算法 2 中超图消解规则，首先消除包含超边 $HE(a_2)$的所有超边，结果如图 2(b)所示，然后消除超边 $HE(a_1, a_3)$，得空如图 2(c)所示。因此，$B = a_2(a_1, a_3)$，即 $B_1 = \{a_1, a_2\}$，$B_2 = \{a_2, a_3\}$。根据定义 8，核心属性集 C，重要属性集 K，不必要属性集为

$$C = B_1 \cap B_2 = \{a_2\}$$
$$R = B_1 \cup B_2 - C = \{a_1, a_3\}$$
$$L = A - (B_1 \cup B_2) = \{a_4\}$$

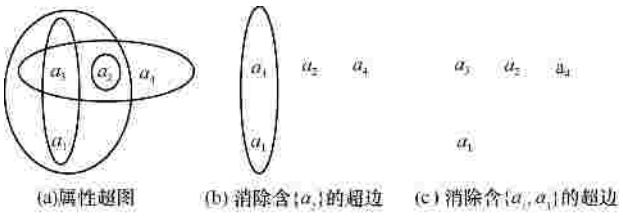

图 2 属性超图消解

数据分块策略割裂了数据间的关联关系，但 CSP 仍然可以通过统计不同数据分块中各属性值的分布关系泄露租户数据隐私。如表 1 所示，使用二元组（$v(a_i), n$）表示属性 a_i 值域中的每一个值在集成数据集 T 的数量。CSP 统计属性 a_3、a_4 及 $Class$ 的属性值分布情况(为了简化标记，式中用 d 替换 $Class$)。

a_3: {{normal, 2}, {reduced, 4}, {more, 2}}

a_4: {{yes, 2}, {no, 6}}

d: {{hard, 3}, {none, 3}, {soft, 2}}

根据最大覆盖原则，CSP 可以得到以下 3 条规则。

a_3: {reduced,4} → {d :{ hard,3} | d :{ none,3}}

a_3: {reduced,4} → a_4 :{ no,6}

d:{ none,3} → a_4 :{ no,6}

又因为属性 a_4 和 $Class$ 在一个数据分块中，CSP 可以得出租户数据中的商业机密，因此提出 (a,k)-组均衡化策略，使各属性值域在各数据分块中均衡分布，防止 CSP 泄露租户数据隐私。

定义 9 概率分布函数 (probability distribution function). $T(U, A, F, Class)$ 为信息系统，记 $R_B = \{(x_i, x_j) \mid f_k(x_i) = f_k(x_j)(a_k \in B)\}$ $(B \subseteq A)$，$U/R_B = \{[x_i]_B \mid x_i \in U\}$，$\frac{U}{R_d} = \{[x_i]_d \mid x_i \in U\}$，为表述方便 $\frac{U}{R_d} = \{d_1, d_2, \cdots, d_r\}$。设 $x_i \in U$，$D\left(\frac{d_k}{[x_i]_B}\right) = \frac{|d_k \cap [x_i]_B|}{[x_i]_B}$ $(k \le r)$，概率分布函数 $m_B(x_i) = \left(D\left(\frac{d_1}{[x_i]_B}\right), \cdots, D\left(\frac{d_r}{[x_i]_B}\right)\right)$。

定义 10 (a,k)-组均衡((a,k)-group balance)。设 $T(U, A, F, Class)$ 满足 k-匿名，属性集的所有非空子集构成 M 个组，$v(G_{A_i})$ 表示组 G_{A_i} 值域中一个可取的值。若对于任意的 G_{A_i} 有 $|v(G_{A_i})| \ge k$ 且 $m_{G_{A_i}}(v(G_{A_i})) \ge a$，则称 T 满足 (a,k)-组均衡。

为满足数据均衡化，使用插入伪造数据的方法，在满足各数据值符合 k-匿名的前提下，对使各数据分块中属性值的分布在基于决策属性的数据划分中满足预设的分布阈值 a。其具体的均衡化算法如算法 3 所示。

算法 3 (a,k)-组均衡算法

输入 基于属性集 A 的数据划分 $\frac{U}{Ra_i}(i = 1, \cdots, |A|)$，

基于 $Class$ 属性的数据划分 $\frac{U}{R_d}$，

属性值分布率 a，

匿名度 k

输出 U

步骤

```
Group_Balance(U/R_ai, U/R_d, a, k){
  FOR each a_i ∈ A {
    j = 1;
    WHILE( j ≤ |U/R_d| ){
      m = 1;
      WHILE ( m ≤ |U/R_ai| ){
        WHILE ( D(d_j / (U/R_ai[m])) < a ){
          insert(U/R_ai[m]) → d_j;
        }
        m++;
      }
      WHILE ( |d_j| < k ){
        insert(Min(U/R_ai[1], U/R_ai[m])) → d_j;
      }
    }
  }
```

RETURN U；

}

分析　本文数据分割策略对 CSP 满足零知识性。

证明　由算法 1 得集成数据 T 的数据匿名度为 $k(k>1)$，CSP 对集成数据满足零知识性，当且仅当执行数据分割策略后集成数据的数据匿名度小于 k。根据算法 2，集成数据 T 被分为 3 部分且每个部分都不含有 QID，CSP 重组一条记录的概率为 $\frac{1}{k^3}$，又根据算法 3，租户通过生成伪造数据对分块的数据集 T' 进行分组均衡化，使 CSP 无法通过数据分布统计攻击推导出更多的知识，同时 CSP 重组一条记录的概率小于 $\frac{1}{k^3}\left(\frac{1}{k^3}=\frac{1}{k}\right)$，所以本文数据分割策略对 CSP 满足零知识性。 证毕。

5.2　针对完全非可信云隐私保护机制

对于完全非可信的云，CSP 可能在经济利益的驱动下仅对云租户上传数据的子集进行操作，依靠属性分割割裂数据关联关系的策略已经不能满足需要。针对这种威胁模型，在 5.1 节的基础上，提出一个数据验证数据结构分类索引树（TIT, taxonomy index tree）。

定义 11　分类索引树。分类索引树是一个深度为 3 的数据验证树形结构(设 *root* 为第 0 层)，*root* 包含全总数据集，从 *root* 到叶节点依次根据数据集的 I、K、C 属性集作为分类条件逐层细化，每一个节点可以看作是一个三元组 $\left(B,\langle B_i,Index\rangle,Count\right)$，$B_i$ 为节点所在层的分类属性集，$\langle B_i,Index\rangle=\left\{\langle b_1,Index_1\rangle,\cdots,\langle b_n,Index_n\rangle\right\}$，其中，$b_i\in B_i$，$Index_i$ 为属性 b_i 指针指向同层中与本节点 b_i 属性值相同的节点。$B=\{A-B_i\ |A$ 为全总属性集$\}$，*Count* 为本节点所包含的数据个数。

算法 4　构建分类索引树算法

输入　数据集 S，记录个数 *Count*，
　　全总属性集 A，准标识符集 B

输出　*root*

步骤

Create_TIT $\left(S,Count,A,B\right)$ {

$root=\left(S,Count\right)$;

$C=\bigcup_{k\le r}B_k$，$K=B-C$，$I=A-B$；

/*创建第 1 层节点，其中，i 为层号，j 为本层节点的序列号*/

计算 R_C，$Node_{i=1,j=1,\cdots,|R_c|}-value=R_{I\,j=1,\cdots,|R_c|}$;

计算 $Node_{i=1,j}\_Count$；

/*同层 *Node* 索引化*/

FOR each $a_x\in I$ {

IF($Node_{i=1,n}(value(a_x))$

$=Node_{i=1,m}(value(a_x))$){

$Node_{i=1,n}\rightarrow next=Node_{i=1,m}$;

}

RETURN *root*;

}

第 2 层节点和第 3 层节点的创建同第 1 层节点（略）；

RETURN　*root*;

}

分析　构建分类索引树算法的时间复杂度为 $O(n)$。租户在上传集成数据前在本地构建分类索引树，由于 CSP 对云端的数据拥有绝对的控制权，租户无法阻止 CSP 违反 SLA 的行为。但通过分类索引树，租户可以验证 CSP 返回数据的正确性及完整性。分类索引树的 *root* 节点是全总记录的泛化，以下各层依次使用不必要属性集、重要属性集、核心属性集迭代地分割数据，形成由粗到细的分类树形结构。叶节点包含满足由 *root* 到叶节点路径限制的所有记录的 ID，租户通过分类索引树获得所查记录的个数及记录的 ID，从而验证 CSP 返回数据的正确性及完整性。图 3 为表 1 数据的分类索引树。

6　实验

6.1　实验环境

针对 DaaS 数据集成隐私保护，通过仿真模拟验证所提隐私保护模型的零知识性、数据正确性和完整性、高效性。

操作系统为 Linux CentOS 4.8，CPU 为 Intel(R) Core 2 Duo，主频 2.1 GHz，内存 2 GB，编程语言采用 Python，数据库使用 MySQL 5.5.23。

6.2　数据加细匿名开销实验

为有效验证数据多轮加细匿名算法的有效性，与文献[6]提出的经典分布式多方 k-匿名算法 DkA 进行比较，设计了 2 组对比实验(为了便于描述，实验以两方数据集成为例)。

实验 1　首先构造数据表 T，含有 22 个属性，1×10^5 条记录的数据集。其中，属性 a_1 为记录 ID，

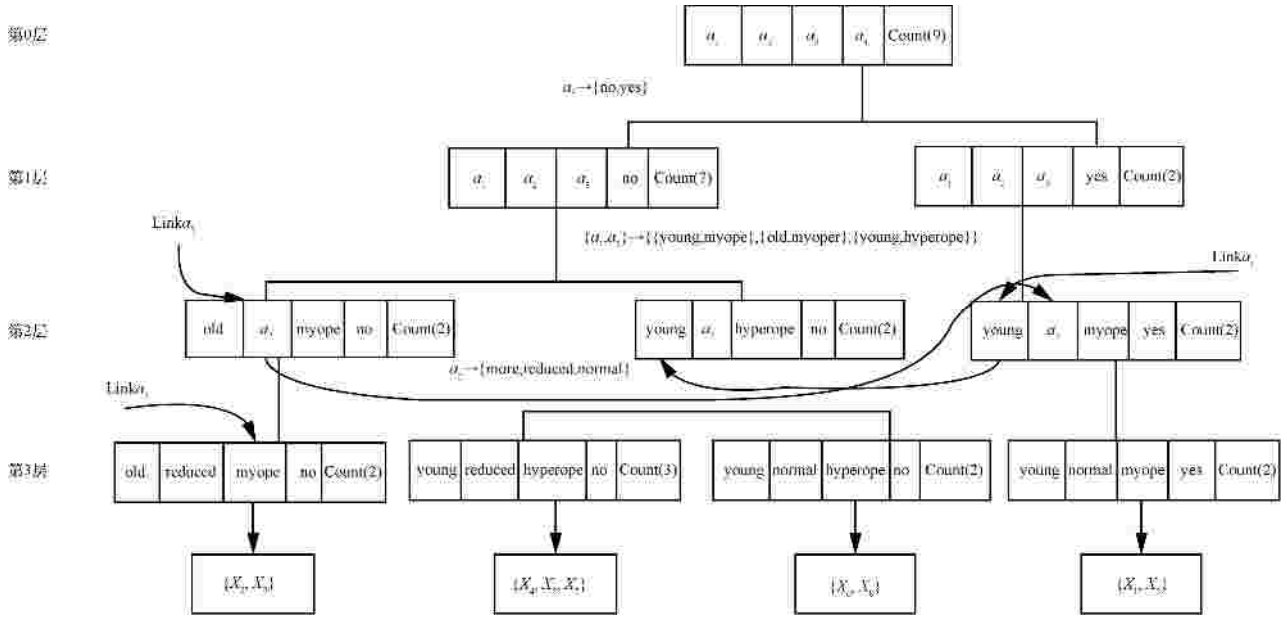

图 3 分类索引树

属性 a_2 为 *Class* 属性，使用伪随机器在各个属性的值域内产生 1×10^5 条记录，然后将数据集垂直分为 2 份：$\{a_1, a_2, a_3 \sim a_{12}\}$和$\{a_1, a_2, a_3 \sim a_{22}\}$。设定属性 $a_3 \sim a_{22}$ 的泛化距离集 $d\{d_1, d_2, \mathrm{L}, d_n\}$，其中，$d_i^{a_k}$ 表示属性 a_k 的第 i 层泛化距离，针对属性 a_k 第 i 层泛化所形成的同一个类中的所有记录值的距离不大于 $d_i^{a_k}$ 且 $d_1^{a_k} < \mathrm{L} < d_i^{a_k} < d_j^{a_k} < \mathrm{L} < d_n^{a_k}$ $(i < j)$。针对不同的匿名度 k={10, 20, 30, 40, 50}，对应的时间开销如图 4 所示，多轮加细算法对匿名度不明感(即保持恒定的时间开销)，而 DkA 算法的时间开销随着匿名度的增加呈线性增长。

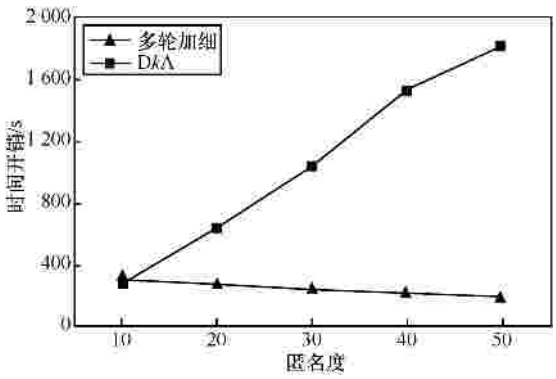

图 4 同数据不同匿名度时间开销

实验 2 设定匿名度 k=20，选取 5 个不同的数据集，数据集的生成方式同上，规模从 5×10^4~25×10^4，时间开销如图 5 所示。多轮加细算法对记录个数不明感（即保持恒定的时间开销），而 DkA 算法的时间开销随着待集成数据集规模的增加呈指数级增长。

实验 3 从数据集成后的知识损失来对 DkA 及本文提出的多轮加细算法进行比较。下面给出对知识损失的 2 个度量指标：知识损失度和相对知识损失度。

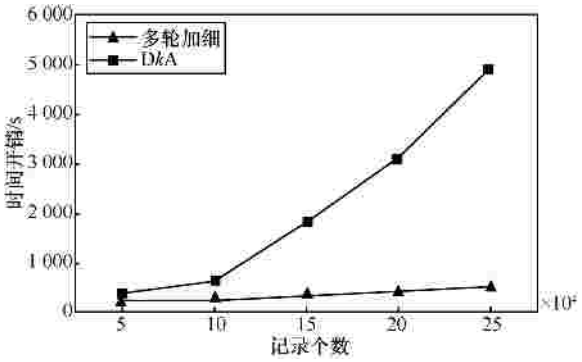

图 5 同匿名度不同数据量时间开销

知识损失度：知识损失度是用来衡量数据集在匿名后所含知识量减少程度的指标，知识损失度 $Info\_lose = 1 - \left(\dfrac{\text{匿名后数据集的信息量}}{\text{匿名前数据集的信息量}}\right)$，其量化式为 $Info\_lose = 1 - \dfrac{\left(\sum_{i=1}^{n} \dfrac{w_i}{d_i}\right)}{\left(\sum_{i=1}^{n} w_i\right)}$，其中，$w_i$ 为属性 i 的

权值且 $\sum_{i=1}^{n} w_i = 1$，d_i 为属性 i 匿名后的泛化距离。

相对知识损失度：相对知识损失度用来衡量实现数据匿名所不必要的知识量损失，相对知识损失度 $rela\_lose = 1 - \frac{\left(\sum_{i=1}^{n} \frac{w_i}{d_i}\right)}{\left(\sum_{i=1}^{n} \frac{w_i}{k}\right)}$，其中，$k$ 为匿名度。

实验 1 中多轮加细匿名算法和 DkA 算法的知识损失度及相对知识损失度如表 2 所示。

表 2　　知识损失度及相对知识损失度

$\frac{Info\_lose(\%)}{rela\_lose(\%)}$	多轮加细匿名算法	DkA 算法
10	3.7 / 0.3	5.2 / 1.8
20	6.8 / 0.5	9.4 / 3.1
30	9.1 / 0.6	13.6 / 5.1
40	11.3 / 0.8	18.7 / 8.2
50	12.7 / 0.5	22.6 / 10.4

实验结果表明：针对不同的匿名度 k，数据多轮加细匿名算法的时间开销较小，且数据匿名化后数据集的知识损失度较小。

6.3　数据分割开销实验

为了验证与应用无关的数据分割算法的有效性，针对不同属性个数，设计一组实验在满足隐私保护的同时，对数据集进行分割处理，对应的时间开销如图 6 所示。实验中记录的个数为 10 万，而属性集的个数由 1 万增长到 10 万。实验结果显示的时间开销随着属性个数的增加，呈线性增加。当属性个数为 10 万时数据分割的时间开销低于 12 s。然而，实际应用中，属性的数量不多于 10 万，且数据分割不是频繁操作，这样的时间开销是可以接受的。

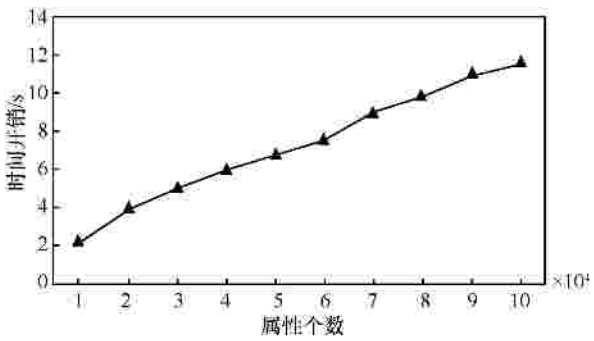

图 6　属性分割时间开销

7　结束语

本文就 DaaS 模式下多个云租户进行数据集成并将集成后的数据集部署在非可信的 CSP 端的情况，研究明文状态下的数据隐私保护机制。针对多租户分布式数据集成，提出多轮加细的匿名数据保护策略，在满足数据匿名的条件下通过租户间多轮协作每轮采用信息增益最大的属性加细数据集，使集成数据在完成数据隐私保护的前提下尽可能含有更多的信息；针对非可信的 CSP，根据其信誉等级不同，提出面向 DaaS 应用的两级隐私保护机制。若 CSP 处于半可信信誉等级，提出一种与应用无关的基于分块的隐私保护机制，隐藏数据之间的关联关系，并通过分组均衡化的方式，确保属性的值域均衡分布，防止 CSP 泄露租户数据隐私；若 CSP 处于完全非可信信誉等级，提出分类索引树数据结构，验证 CSP 返回数据的正确性及完整性。

未来工作主要包括：在假定多方云租户为非理性的前提下，研究在不存在可信第三方仲裁机制下的数据集成策略，防止恶意租户共享虚假数据信息迫使其他租户共享更多的信息；进一步研究面向 DaaS 应用的数据拆分分布均衡化策略，在保护数据隐私的同时提高数据处理效率，减少隐私保护机制给 DaaS 应用带来的数据精度降低问题，提高租户体验。

参考文献：

[1] 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.
ZHOU S G, LI F, TAO Y F, et al. Privacy preservation in database applications: a survey[J]. Chinese Journal of Computers, 2009, 32(5): 847-861.

[2] 田秀霞, 王晓玲, 高明, 等. 数据库服务—安全与隐私保护[J]. 软件学报, 2010, 21(5): 991-1006.
TIAN X X, WANG X L, GAO M, et al. Database as a service—security and privacy preserving[J]. Journal of Software, 2010, 21(5):991-1006.

[3] CRAIG G. Fully homomorphic encryption using ideal lattices[C]//The 41st Annual ACM Symposium on Theory of Computing (STOC). Bethesda, MD, USA, c2009:169-178.

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

[5] WONG R C, LI J, FU A W, et al. (a, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data

Mining (SIGKDD). Philadelphia, PA, USA, c2006: 754-759.

[6] JIANG W, CLIFTON C. A secure distributed framework for achieving anonymity[J]. The International Journal on Very Large Data Bases, 2006, 15(4): 316-333.

[7] CLIFTON C, KANTARCIOGLU, VAIDYA J. Tools for privacy preserving distributed data mining[J]. ACM SIGKDD Explorations, 2003, 4(2): 1-7.

[8] MOHAMMED N, FUNG B C M, DEBBABI M. Anonymity meets game theory: secure data integration with malicious participants[J]. Very Large Data Bases Journal (VLDBJ), 2011, 20(4):567-588.

[9] MOHAMMED N, FUNG B C M, et al. Centralized and distrib anonymization for high-dimensional healthcare data [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2010, 4(4):18: 1-18:33.

[10] JIANG W, CLIFTON C, KANTARCIOGLU M. Transforming semi-honest protocols to ensure accountability[J]. Data Knowl Eng, 2008,65(1): 57-74.

[11] DU W., HAN Y S, CHEN S. Privacy-preserving multivariate statistic cal analysis: linear regression and classification[C]//SIAM International Conference on Data Mining. Florida, c2004: 222-233.

[12] PINKAS B. Cryptographic techniques for privacy-preserving data mining [J]. ACM SIGKDD Explor News, 2002, 4(2): 12-19.

[13] VAIDYA J, CLIFTON C. Privacy preserving *k*-means clustering over vertically partitioned data[C]//The ACM SIGKDD, c2003: 206-215.

[14] CIRIANI V, VIMERCATI S, FORESTI S. Fragmentation design for efficient query execution over sensitive distributed databases[C]//The 29th ICDCS, Canada, c2009: 32-39.

[15] CIRIANI V, VIMERCATI S, FORESTI S. Selective data outs for enforcing privacy [J]. Journal of Computer Security 2011, 19: 531-566.

[16] VIMERCATI S, FORESTI S, JAJODIA S. Fragments and loose associations: respecting privacy in data publishing[J]. Proceeding of the VLDB Endowment, 2010, 3(12): 1370-1381.

[17] **张坤, 李庆忠, 史玉良. 面向** SaaS **应用的数据组合隐私保护机制研究**[J]. **计算机学报**, 2010, 33(11): 2044-2054.
ZHANG K, LI Q Z, SHI Y L. Research on data combination privacy preservation mechanism for SaaS[J]. Chinese Journal of Computers, 2010, 33(11): 2044-2054.

[18] ZHANG K, SHI Y, LI Q. Data privacy preserving mechanism based on tenant customization for SaaS[C]//The International Conference on Multimedia Information Networking and Security (MINES). Wuhan, China, 2009. 599-603.

[19] OSBORNE M J, RUBINSTEIN A. A course in game theory[M]. The MIT Press, Cambridge, UK , 1994.

作者简介：

周志刚（1986-），男，山西太原人，哈尔滨工业大学博士生，主要研究方向为云安全和信息安全。。

张宏莉（1973-），女，吉林榆树人，哈尔滨工业大学教授、博士生导师，主要研究方向为网络与信息安全、网络测量与建模、网络计算、并行处理等。

余翔湛（1973-），男，黑龙江哈尔滨人，哈尔滨工业大学教授，主要研究方向为网络容灾和信息安全。

李攀攀（1983-），男，山东曲阜人，哈尔滨工业大学博士生，主要研究方向为云安全和信息安全。