无可信第三方的数据匿名化收集协议

周治平*①② 李智聪①

①(江南大学物联网工程学院 无锡 214122)

②(江南大学物联网技术应用教育部工程研究中心 无锡 214122)

摘 要：针对半诚信的数据收集者对包含敏感属性(SA)数据收集和使用过程中可能造成隐私泄露问题，该文在传统模型中增加实时的数据领导者，并基于改进模型提出一个隐私保护的数据收集协议，确保无可信第三方假设前提下，数据收集者最大化数据效用只能建立在K匿名处理过的数据基础上。数据拥有者分布协作的方式参与协议流程，实现了准标识(QI)匿名化后SA的传输，降低了数据收集者通过QI关联准确SA值的概率，减弱内部标识揭露造成隐私泄露风险；通过树形编码结构将SA的编码值分为随机锚点和补偿距离两份份额，由K匿名形成的等价类成员选举获取两个数据领导者，分别对两份份额进行聚集和转发，解除唯一性的网络标识和SA值的关联，有效防止外部标识揭露造成的隐私泄露；建立符合该协议特性的形式化规则并对协议进行安全分析，证明了协议满足隐私保护需求。

关键词：数据隐私；隐私保护；K匿名；敏感属性；匿名化

中图分类号：TP309.2 **文献标识码**：A **文章编号**：1009-5896(2019)06-1442-08

DOI: 10.11999/JEIT180595

Data Anonymous Collection Protocol without Trusted Third Party

ZHOU Zhiping①② LI Zhicong①

①(*School of Internet of Things Engineering*, *Jiangnan University*, *Wuxi* 214122, *China*)

②(*Engineering Research Center of Internet of Things Technology Applications of Ministry of Education*, *Jiangnan University*, *Wuxi* 214122, *China*)

Abstract: Semi-honest data collectors may cause privacy leaks during the collection and use of Sensitive Attribute (SA) data. In view of the problem, real-time data leaders are added in the traditional model and a privacy-protected data collection protocol based on the improved model is proposed. Without the assumption of trusted third party, the protocol ensures that data collectors maximization data utility can only be established on the basis of K-anonymized data. Data owners participates in the protocol flow in a distributed and collaborative manner to achieve the transmission of SA after the Quasi-Identifier (QI) is anonymized. This reduces the probability that the data collector uses the QI to associate SA values and weakens the risk of privacy leakage caused by internal identity disclosure. It divides the coded value of the SA into two shares of a random anchor point and a compensation distance through the tree coding structure and the members of the equivalent class formed by K-anonymity elect two data leaders to aggregate and forward the two shares respectively, which releases the association between unique network identification and SA values and prevents leakage of privacy caused by external identification effectively. Formal rules are established that meet the characteristics of the protocol and analyze the protocol to prove that the protocol meets privacy protection requirements.

Key words: Data privacy; Privacy protection; K-anonymity; Sensitive Attribute (SA); Anonymization

1 引言

近年来，随着数据的战略重要性与日剧增，大数据产生的商业价值逐渐得到重视，许多应用平台以提供高质量服务为由，获取关联用户隐私的敏感性数据[1,2]。部分平台为实现某种商业目的，在用户未授权的前提下将数据提供给第三方进行处理和分析，间接造成用户隐私泄露。例如：Facebook因授权一款个性分析测试APP获取用户自己和朋友的个人相关信息，间接造成接近5×10^7个用户的信息泄露，面临罚款和信任危机[3]。

目前针对敏感数据隐私保护的研究，多数聚焦

收稿日期：2018-06-19；改回日期：2019-03-04；网络出版：2019-03-25

*通信作者： 周治平 zzp@jiangnan.edu.cn

在平台发布数据时的匿名化处理方案上，主要研究中心化匿名处理技术。文献[4]最早提出K匿名的算法，对聚集的数据进行匿名化处理，保证数据发布时每条数据记录与至少$k-1$条记录的准标识符(Quasi-Identifier, QI)一致。针对K匿名形成的等价类内部某一敏感属性(Sensitive Attribute, SA)值出现频率过高，易造成同质攻击的现象，文献[5]提出满足l多样性的K匿名算法。文献[6]为防止攻击者依据敏感语义实施偏斜性攻击，提出T近邻(T-Closenss)的K匿名算法，保证等价类内的敏感属性分布近视等于其在整个数据列表中的全局分布。文献[7,8]提出了差分隐私保护模型，基于此模型可以通过数学理论获取数据效用和隐私保护之间的均衡。这些研究均需要数据收集者聚集数据拥有者的原始数据，然而在商业动机的驱使下，难以保证数据收集者理想可信。

在无完全可信的数据收集者的前提下，文献[9]提出在数据收集者和数据拥有者之间设置匿名层，采用洋葱路由和混合网络的匿名通信技术，防止外部标识揭露造成隐私泄露，然而仅通过本地设备对个人准标识进行泛化处理，使其泛化程度不低于设定阈值的方案，不涉及任何隐私策略，泛化处理后的准标识仍可能存在唯一性，不能有效地防止内部标识的揭露。在无标识信道假设的前提下，文献[10]采用安全多方加密的方法防止内部标识揭露，并通过抑制聚集数据的准标识实现K匿名，防止联合准标识属性关联用户隐私；文献[11]基于局部保持映射的思想，利用同态加密技术加密敏感数据，在密文域上实现敏感数据的匿名化处理，防止内部标识的揭露。上述研究，在无可信数据收集者假设的前提下，实现对敏感数据的收集，某种程度上防止了内部或外部标识揭露造成的隐私泄露。然而，在用户节点资源受限的现实情况下，依赖复杂的加密技术进行匿名层以及无标识信道的假设是不切实际的。文献[12]基于模型结构的方式进行数据匿名化，采用分布架构的思想，实现局部社区数据分级匿名化处理，此模型未能考虑客户端之间数据传输造成的隐私泄露问题。

针对以上研究中可信第三方假设和无标识信道假设的不合理现象，通过改进模型和设计协议的方式，确保无可信第三方假设前提下敏感数据的匿名化收集。由数据拥有者分布协作的方式参与协议流程，每个数据拥有者被组织进一个等价类，该等价类包含至少k个共享泛化准标识的成员，从每个等价类随机选举代表聚集并转发敏感数据给数据收集者。基于唯一性和敏感性关联会造成隐私泄露的理论，设计符合该协议特性的形式化分析方法。

2　相关分析

假定数据拥有者为获取个人健康相关的服务，使用健康管理应用将本地存储的个人健康记录发送给数据服务器，以便数据挖掘者和专家对数据进行分析并对个人健康行为提供决策支持。传统的模型通常有3部分组成，分别为数据分析者、数据收集者和数据拥有者[13,14]，数据的收集建立在可信第三方数据收集者假设的基础上，数据收集者聚集所有数据拥有者的健康数据进行中心化的匿名处理后，发布数据给数据分析者或提供数据查询接口。基于此模型结构去实现数据匿名化，数据收集者必须拥有对数据拥有者原始数据的支配权，无疑增加了数据拥有者隐私泄露的风险。

2.1　数据模型及安全威胁

潜在隐私泄露威胁的数据模型主要包含3个属性，分别为唯一性标识(User IDentification, UID)、QI以及包含用户敏感信息的SA[15]。假设网间传输的实时数据流由多个元组组成，一个元组可以被描述为(UID, QI, SA)，其中UID为该元组的唯一标识，能够用他区分其他元组，通常在进行数据交换时要将其移除，避免攻击者利用此唯一性链接到该数据产生者的真实身份。QI由个人基本信息相关的准标识属性组成，例如：年龄、性别、邮编等，单个准标识属性可能不具备唯一性，联合多个准标识属性可能链接到具体的目标(例如，通过性别、生日、邮编等属性几乎可以确定唯一的目标身份)。SA具有敏感性，主要为元组数据产生者的敏感信息，例如：工资、诊疗信息等。忽略UID且具有单一敏感属性的数据流，可由一系列(QI, SA)组成。针对满足此数据模型的数据进行收集，要考虑两种情况的隐私泄露，内部标识揭露和外部标识揭露。其中，内部标识揭露指攻击者通过UID或QI区分用户的数据记录，从而关联个人的敏感信息，外部标识揭露指攻击者通过网间添加到载荷的网络头(网络ID)区分用户的数据记录，关联个人敏感信息[16]。攻击者通过背景知识了解到某个用户的年龄、性别、邮编等准标识属性信息，并且获取到无UID的数据列表，显然通过了解的准标识信息建立唯一性，就能关联出数据列表中该用户的敏感属性信息。

2.2　敏感属性编码

通过树形编码结构可将敏感属性分为两份份额，分别为随机锚点和补偿距离。树形结构的叶子

节点代表敏感属性的编码值W，目标用户在树形结构上随机选择一个节点，称为随机锚点R，定义补偿距离为$D=R\oplus W$，则已知目标用户的补偿距离D和随机锚点R，通过$W=R\oplus D$的方式可得到目标用户的敏感属性编码值W。如图1假设目标用户的敏感属性编码值为11011，选择的随机锚点为10110，由$D=R\oplus W$可得他的补偿距离为01101。补偿距离的前两位“01”表示随机锚点到目标节点的垂直距离为1，后3位为“101”表示锚点与目标结点水平距离为5，图1中随机锚点的位置下移1右移5可到达目标用户敏感属性编码值的位置。同时拥有敏感属性的两份份额，可以推导出敏感属性的编码值。已知目标用户的补偿距离为01101和随机锚点为10110，则通过$W=R\oplus D$可得到目标用户的敏感属性编码值为11011。

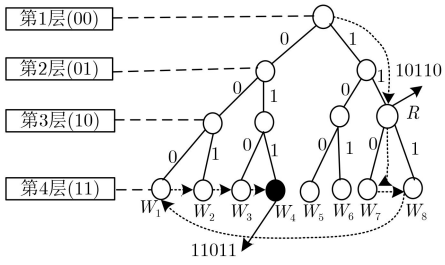


图 1 树形编码结构图

3 改进模型及协议描述

3.1 改进模型

区别于传统模型中心式的匿名化处理，从削弱数据收集者对原始数据的支配权出发，在原有系统模型基础上增加实时的数据领导者，参与数据收集的过程。数据领导者由匿名化QI构建的等价类中成员通过选举算法[17]获取，每个等价类需要推荐两个领导者。依据模型的动态性，将数据收集过程分为阶段1和阶段2，不同的阶段系统模型参与通信的实体不同，阶段1通信的实体为数据拥有者和数据收集者，阶段2增加两个数据领导者。图2表示参与实体的模型结构，阶段1主要包括初始化、QI匿名化、泛化准标识(Generalized Quasi-Identifier, GQI)的有效性验证，阶段2主要进行数据领导者选取、敏感属性数据的编码和传输。

假设数据收集者为半诚信的，即它完全遵守协议的执行过程，但试图通过执行协议中获取到的数据分析出特定数据拥有者关联的敏感信息[18]。此外，等价类成员选举的领导者也是半诚信的，它们对等价类其它成员的敏感信息感兴趣。隐私泄露的威胁除了来自半诚信的数据收集者和数据领导者外，还有恶意的窃听者。不失一般性，假定存在的

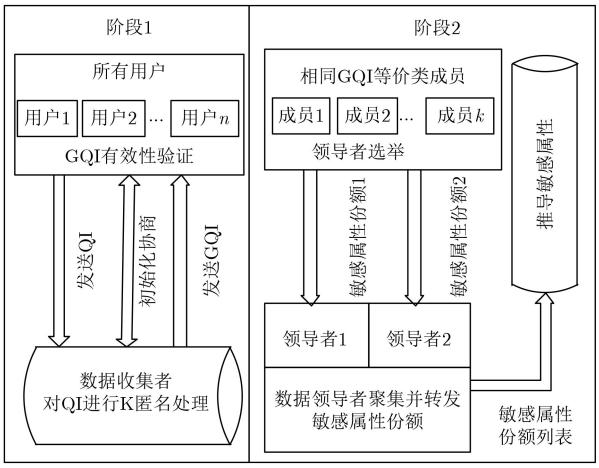


图 2 参与实体的模型结构

半诚信攻击者以及窃听者都已知背景知识，能够通过获取到的外部标识或内部标识建立到真实身份的链接。协议要在安全模型假设前提下，确保敏感数据的收集过程不会造成隐私泄露，即在整个协议执行过程中，除数据拥有者自身外任何实体不能确定某一数据拥有者对应的敏感信息。

3.2 协议描述

3.2.1 协议流程

阶段1 首先，通过初始化实现数据拥有者和数据收集者之间树形编码结构的协商，可以通过有效的秘钥协商协议传递树形结构的参数，实现数据拥有者和数据收集者之间编码结构信息的同步。然后，进行QI的匿名化处理以及GQI的有效性验证。所有的数据拥有者发送他们的QI给数据收集者，数据收集者通过K匿名算法对聚集到的QI列表进行泛化处理，生成由K个元组组成的若干个等价类，其中K值的大小取决于匿名化算法的设置以及匿名化的QI数据，每个等价类对应的K值不同，但都大于等于K匿名算法设定的最小值k。每个等价类具有统一的泛化准标识GQI。随后数据收集者将GQI发送给相应的共享GQI的数据拥有者，数据拥有者要对GQI进行有效性验证，确保自身QI中的每个准标识属性值都被包含在对应的GQI范围内，假设某一数据拥有者自身的年龄属性值为24，接收到的GQI中的年龄属性范围为25～27，此时年龄属性值不被包含在对应的GQI范围内，数据拥有者将认为泛化得到的GQI无效的，选择终止协议。

阶段2 首先，进行领导者的选举。所有数据拥有者基于相同的GQI构建分组，同组之间随机选择数据领导者，数据领导者的选举为共识问题，同组参与者位于同一网络中，因此，可采用选举算法从具有相同GQI等价类中随机选取两个领导者，两个领导者不为同一个用户。然后，进行敏感属性数据

的编码和传输。每个等价类中的所有成员，从树形结构选择一个随机锚点，并计算随机锚点与自身敏感属性编码值之间的补偿距离。等价类所有成员发送GQI和随机锚点给数据领导者1，发送GQI和补偿距离给数据领导者2。数据领导者1将聚集的整个等价类的GQI和随机锚点列表转发给数据收集者，数据领导者2将聚集的整个等价类补偿距离列表转发给数据收集者。最后，数据收集者在收到数据列表后，对整个列表进行遍历操作，通过随机锚点和补偿距离推断出敏感属性的编码值，并根据初始协商的编码结构信息获取编码值映射的敏感属性值，最终可以得到一个匿名化的数据列表。

3.2.2 协议具体步骤

假设共有N个数据拥有者$\boldsymbol{U}=(U_1,U_2,\cdots,U_N)$，其中$i\in[1,N]$。$U_i$拥有的数据为$(Q_i,S_i)$，其中，$Q_i$表示$U_i$的准标识属性数据，$S_i$表示$U_i$的敏感属性数据。DC为数据收集者，$Q_i$组成的列表$\boldsymbol{Q}$经DC匿名化处理，生成由若干等价类组成的列表$\boldsymbol{G}$，每个等价类将由$K$个共享泛化准标识$G_j$的元组组成，设共有$M$个等价类，则$\boldsymbol{G}=(G_1,G_2,\cdots,G_M)$，其中$j\in[1,M]$。等价类中元组对应的成员(数据拥有者)为$U_k$，其中$k\in[1,K]$。$L_1^j$和$L_2^j$为泛化准标识为$G_j$的等价类中，所有成员随机选择的领导者1和领导者2。某个等价类中成员U_k的敏感属性真实值为S_k，其对应的敏感属性编码值为W_k，U_k在树形编码结构上选择的随机锚点R_k，R_k到V_k的补偿距离为记为D_k，由R_k和D_k推导出敏感属性编码值S_k的过程为$W_k{=}R_k\oplus D_k$。如表1和表2分别表示阶段1和阶段2的协议步骤。

表 1　阶段1协议步骤

(1) for $U_i\in\boldsymbol{U}$, $1\le i\le N$ do
　　U_i发送Q_i给DC.
(2) DC通过K匿名将Q泛化为G
　　for $G_j\in\boldsymbol{G}$, $1\le j\le M$ do
(3) for G_j中元组对应的U_k, $1\le k\le K$ do
　　DC向U_k发送G_j
　　if 每个U_k验证G_j是有效，进入阶段2
　　else 终止协议

表 2　阶段2协议步骤

(1) for $G_j\in\boldsymbol{G}$, $1\le j\le M$ do
　　随机选取领导者L_1^j和L_2^j
　　for G_j中元组对应的U_k, $1\le k\le K$ do
　　发送(G_j,R_k)和(G_j,D_k)分别给L_1^j和L_2^j
(2) L_1^j和L_2^j分别聚集(G_j,R_k)和(G_j,D_k)列表的给DC
(3) for $1\le i\le N$ do
　　DC计算$W_i{=}R_i\oplus D_i$
(4) 搜索W_i映射的S_i得到数据列表$(\boldsymbol{G},\boldsymbol{S})$

4　协议安全分析

4.1　安全目标和相关理论

协议要实现的目标有以下几点：在协议的所有参与实体为半诚信的假设下，任何参与实体不能准确的关联用户的隐私；恶意的窃听者也不能关联用户的隐私；半诚信的数据收集者DC最终获取到的数据列表满足K匿名的隐私需求。便于协议安全分析，基本定义和协议数据交换的有向图展示如下。

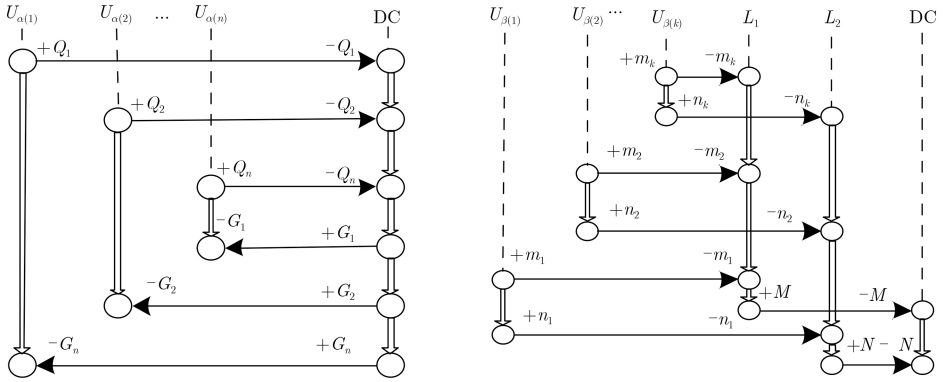
定义1　数据记录具有唯一性：假设$(d_1,d_2,\cdots,d_n)$是一个数据列表，$\exists i\in(1,n)$满足$\text{view}_\text{O}(d_{\pi(i)})\not\equiv\text{view}_\text{O}(d_i)$，则称$d_i$具有唯一性。其中O为观测者，$\pi$为对标注$(1,2,\cdots,n)$以任意的方式重新排列，“$\equiv$”表示多项式时间内解析等价。即在观测者O的角度，在数据列表中找不到与d_i解析等价的数据记录。

定义2　数据列表满足K匿名：假设$(d_1,d_2,\cdots,d_n)$是一个具有敏感属性的数据列表，将每条数据记录d_i分为d_i^+和d_i^-两部分组成，用d_i^-表示具有敏感属性的数据，d_i^+为d_i的其他属性。如果$\exists I\subseteq\{1,2,\cdots,n\}$且$|I|\ge k$, ∂为满足属于I的标注的任意排序，满足$\text{view}_\text{O}((d_1^+,d_1^-),(d_2^+,d_2^-),\cdots,(d_n^+,d_n^-))\equiv\text{view}_\text{O}((d_1^+,d_{\partial(1)}^-),(d_2^+,d_{\partial(2)}^-),\cdots,(d_n^+,d_{\partial(n)}^-))$。其中$i\in(1,n)$，$i\notin I$时$\partial(i)=i$。称数据列表$(d_1,d_2,\cdots,d_n)$是满足K匿名的数据列表。

根据协议的基本流程阶段1和阶段2协议的有向图如图3。其中，阶段2展示的为其中一个等价类的数据交换过程，其中$i\in(1,K)$，$m_i:(G_i,R_i)$，$n_i:(G_i,D_i)$，$M:(m_1,m_2,\cdots,m_k)$，$N:(n_1,n_2,\cdots,n_k)$，且所有在同一等价类中所有G_i的值是相等的。其中，图3(a)和图3(b)的i不代表同一个实体，在一个标准标注下，假设阶段1对标准标注进行了α排序简称$\alpha(i)$，阶段2对标准标注进行了β排序简称$\beta(i)$。

定理1　在该协议中，DC能找到两种排序方式α,β使得$\alpha(i)=\beta(i)$的概率小于等于$1/k$。

证明　阶段1和阶段2在一个标准标注下，假设$(Q_{\alpha(i)},G_{\alpha(i)})=(Q_j,G_j)$，如果$\alpha(i)=\beta(i)$则有$G_{\beta(i)}=G_j$。与$G_j$值相等的记录至少有$k$个，DC能找到至少$k$种排列方式满足$G_{\beta(i)}=G_j$，但仅有一种排列方式使得$\alpha(i)=\beta(i)$，因此，DC找到两种排序方式$\alpha,\beta$使得$\alpha(i)=\beta(i)$的概率小于等于$1/k$。



(a) 阶段1 (b) 阶段2

图 3 协议有向图

定理2 攻击者通过随机猜测的方式获取R_i和D_i的获胜概率均为$1/2^l$，其中l为编码长度。

证明 编码长度为l，依据编码规则，可以产生2^l个不同编码值，每个编码值都是唯一的，因此攻击者随机猜测一个编码值等于R_i或者D_i的概率为$1/2^l$。

4.2 协议有效性分析及证明

4.2.1 攻击手段和安全策略

(1) 内部标识揭露：协议将敏感属性与具有唯一性的准标识分开传输，将准标识通过K匿名处理后建立等价类，基于等价类建立分组，以组为单位传输敏感属性，降低了准标识与敏感属性的关联程度。在阶段1，U_i仅发送其Q_i给DC, Q_i不具备敏感属性，DC仅获取到U_i的Q_i，不能从中推导出有价值的敏感信息。在阶段2, DC能够获取到U_i泛化准标识G_i以及通过树形编码推测出的敏感属性值S_i，但是G_i是有K个用户共享的不具备唯一性，DC无法通过具有唯一性的准标识关联S_i；

(2) 外部标识揭露：该协议通过领导者聚集和转发敏感数据的方式，传递敏感性信息。在阶段2，敏感属性被聚集到数据领导者，由数据领导者转发整个等价类的敏感属性，仅揭露数据领导者ID情况下，完成敏感属性数据的收集，攻击者是无法通过数据领导者的ID关联其他用户的敏感信息；

(3) 内部攻击：为了防止身为等价类成员的数据领导者实施内部攻击，协议采用树形编码的方式，将敏感属性分为两个份额，分别由两个领导者聚集并转发敏感属性的份额。联合两个领导者，等价类成员U_i的S_i将被获取，U_i的隐私将被揭露。将敏感属性分为两份份额传输后领导者L_1仅获取到R_i，而D_i在领导者L_2中，L_1, L_2无法单独造成用户U_i的隐私泄露。此外，L_1, L_2的选择是随机的，每个组成员都有机会成为领导者，所有的等价类成员拥有的信息具有对称性，没有可观的利益驱使去产生恶意行为，且领导者之间具有匿名性，L_1与L_2无法相互确认身份，从而难以共谋；

(4) 无效GQI攻击：在初始化阶段，数据收集者要进行QI匿名化处理，构建至少k个成员的等价类，假设半诚信的收集者欺骗用户构建的等价类成员个数少于k，甚至GQI仍具有唯一性，造成无效GQI情况下的隐私泄露。泛化准标识为G_j的等价类中，每个成员在阶段1，要求验证其中每个的准标识属性是否包含在其相应的泛化准标识G_j范围内；

4.2.2 形式化规则

协议主要面临两种安全威胁分别为内部标识揭露造成的隐私泄露以及外部标识揭露造成的隐私泄露，两种原理都是在已知背景知识的前提下，通过外部标识或者内部标识的唯一性关联敏感属性拥有者的真实身份，从而造成隐私泄露。利用了BAN逻辑中的一些数学符号，并对部分符号重新定义，设计了符合本类协议的形式化语言。$\#_S(d_i)$：数据d_i具有敏感属性；$\#_U(d_i)$：数据d_i具有唯一属性；$U_i \ni d_i$：数据d_i的产生者身份为U; $P|\equiv\pi(i)$:半诚信的P已知了背景知识，可以建立所有唯一性到真实身份的链接，也即P知道标注i对应的真实身份。

规则1 $\{\text{view}_\text{O}(d_{\pi(i)}) \neq \text{view}_\text{O}(d_i)\} \vdash \#_U(d_i)$，如果$d_i$区别于其它数据记录，$d_i$具有唯一性。

规则2 $\{\#_U(a_i), a_i \in d_i\} \vdash \#_U(d_i)$，如果$a_i$具有唯一性，且$a_i$为$d_i$的一部分，则$d_i$具有唯一性。

规则3 $\{\#_S(b_i), b_i \in d_i\} \vdash \#_S(d_i)$，如果$b_i$具有敏感性，且$b_i$为$d_i$的一部分，则$d_i$具有敏感性。

规则4 $\{P \triangleleft d_i, \#_U(d_i), P|\equiv\pi(i)\} \vdash P|\equiv U_i \ni d_i$，如果$P$知道$d_i$且$d_i$具有唯一性，此外$P$已知背景知识能够建立所有唯一性与真实身份的链接，则P能推测出d_i产生者的真实身份U_i。

规则5 $\{P|\equiv U \ni d_i, \#_S(d_i)\} \vdash P|\Rightarrow U_i\text{privacy}$，如果$P$能推测出$d_i$产生者的真实身份$U_i$，并且$d_i$具

有敏感性，则P对U_i的隐私具有裁决权，P有能力揭露U_i的隐私。

规则 6　$\{P \triangleleft d_i, \#_S(d_i), \#_U(d_i), P| \equiv \pi(i)\} \vdash P| \Rightarrow U_i\text{privacy}$，假设$P$知道$d_i$并且$d_i$具有唯一性和敏感性，此外$P$已知背景知识能够建立所有唯一性与$U_i$的链接，则$P$对$U_i$的隐私具有裁决权，$P$有能力揭露$U_i$的隐私。

协议初始化阶段主要进行了树形编码的协商$\text{DC} \overset{\text{TreeCode}}{\longleftrightarrow} U$，也即$\text{DC} \triangleleft \text{TreeCode}$, $U \triangleleft \text{TreeCode}$，在整个协议中所有的实体为半诚信的，即$e| \equiv \pi(i)$, $\text{DC}| \equiv \pi(i)$, $U| \equiv \pi(i)$, $L_1| \equiv \pi(i)$, $L_2| \equiv \pi(i)$，其中e表示恶意的攻击者。证明协议能够实现以下目标，验证协议的有效性。

定理3　在协议中，如果实体$P \triangleleft \text{TreeCode}$且$P \triangleleft (R_i, D_i)$，则$P \triangleleft S_i$。

证明　由已知的随机锚点P_i和补偿距离D_i，通过$W_i{=}R_i \oplus D_i$可推测出敏感属性的编码值W_i，并依据树形编码结构TreeCode，找到与W_i映射的真实敏感属性值S_i。

4.2.3 有效性证明

引理1　协议满足，DC能够准确揭露用户U_i隐私的概率小于等于$1/k$。

证明　初始化阶段$\text{DC} \triangleleft \text{TreeCode}$, $\text{DC}| \equiv \pi(i)$。根据有向图在阶段1过程中，$\text{DC} \triangleleft (Q_1, Q_2, \cdots, Q_n)$, $\text{DC} \triangleleft (\text{ID}_1, \text{ID}_2, \cdots, \text{ID}_n)$，在阶段2过程中，$\text{DC} \triangleleft ((G_1, R_1), (G_2, R_2), \cdots, (G_k, R_k))$，$\text{DC} \triangleleft ((G_1, D_1), (G_2, D_2) \cdots, (G_k, D_k))$。阶段1中，对于$\forall i, j \in N$, $\text{ID}_i \neq \text{ID}_j$，所以对于$\forall i, j \in N$, $\text{view}_{\text{DC}}(\text{ID}_{\pi(i)}) \not\equiv \text{view}_{\text{DC}}(\text{ID}_i)$，由规则1得：$\#_U(\text{ID}_i)$，令$\text{ID}_i \in d_i$，则由规则2可得$\#_U(d_i)$，因此由规则4可得：$\text{DC}| \equiv U \ni d_i$。由已知和规则5可得：要揭露$U$的隐私必须满足$\#_S(d_i)$，而阶段1中DC没有获取到敏感信息。阶段2中，$\text{DC} \triangleleft (R_i, D_i)$又$\text{DC} \triangleleft \text{TreeCode}$，由定理3可得：$\text{DC} \triangleleft S_i$。在标准标注下，$\text{DC} \triangleleft S_{\beta(i)}$且$\text{DC}| \equiv U \ni d_{\alpha(i)}$，要满足规则5，需要保证$S_{\beta(i)} \in d_{\alpha(i)}$，当且仅当$\alpha(i) = \beta(i)$时满足条件。由定理1可得，$\alpha(i) = \beta(i)$的概率小于等于$1/k$，即DC获胜的概率小于等于$1/k$记为$\text{Adv}_{\text{DC}}^{U_i\text{privacy}}(A) \leq 1/k$。

引理2　协议满足，L_1, L_2能够准确揭露用户U_i隐私的概率小于等于$1/2^l$。

证明　$L_1| \equiv \pi(i)$, $L_1 \triangleleft \text{TreeCode}$, $L_1 \triangleleft (\text{ID}_1, \text{ID}_2, \cdots, \text{ID}_k)$, $L_1 \triangleleft ((G_1, R_1), (G_2, R_2), \cdots, (G_1, R_k))$。与引理1同理可以获得$\#_{UI}(d_i)$，其中$\text{ID}_i \in d_i$。要满足规则6，必须实现$L_1 \triangleleft (R_i, D_i)$，而$L_1 \triangleleft D_i$不成立，假设$L_1$通过随机猜测补偿距离的方式获取，由定理2可得获胜的概率小于等于$1/2^l$，即$\text{Adv}_{L_1}^{U_i\text{privacy}}(A) \leq 1/2^l$。同理对于$L_2$，无法实现$L_2 \triangleleft P_i$，随机猜测的概率也为$1/2^l$，因此$\text{Adv}_{L_2}^{U_i\text{privacy}}(A) \leq 1/2^l$，其中$l$为编码长度。

引理3　协议满足，恶意攻击者e无法正确揭露用户U_i隐私。

证明　恶意攻击者$e| \equiv \pi(i)$, $e \triangleleft (\text{ID}_1, \text{ID}_2, \cdots, \text{ID}_k)$, $e \triangleleft ((G_1, R_1), (G_2, R_2), \cdots, (G_1, R_k))$, $e \triangleleft ((G_1, D_1), (G_2, D_2) \cdots, (G_1, D_k))$。由规则6，必须实现$e \triangleleft S_i$。即使$e \triangleleft (R_i, D_i)$，而$e \triangleleft \text{TreeCode}$不成立，根据定理3可得$e \triangleleft S_i$不成立，因此$e$无法正确揭露用户$U$的隐私。

引理4　协议中，DC最终获取到的数据列表满足K匿名。

证明　由$\text{DC} \triangleleft \text{TreeCode}$，又阶段2过程中，$\text{DC} \triangleleft ((G_1, R_1), (G_2, R_2), \cdots, (G_k, R_k))$，$\text{DC} \triangleleft ((G_1, D_1), (G_2, D_2), \cdots, (G_k, D_k))$，根据定理3可得：$\text{DC} \triangleleft ((G_1, S_1), (G_2, S_2), \cdots, (G_k, S_k))$。由于在同一等价类中的$G_i$相等，则有$\text{view}_{\text{DC}}(G_1, G_2, \cdots, G_k) = \text{view}_{\text{DC}}(G_{\pi(1)}, G_{\pi(2)}, \cdots, G_{\pi(k)})$，又由$\text{view}_{\text{DC}}((G_{\pi(1)}, S_{\pi(1)}), (G_{\pi(2)}, S_{\pi(2)}), \cdots, (G_{\pi(k)}, S_{\pi(k)})) \equiv \text{view}_{\text{DC}}((G_1, S_1), (G_2, S_2), \cdots, (G_k, S_k))$，所以$\text{view}_{\text{DC}}((G_{\pi(1)}, S_{\pi(1)}), (G_{\pi(2)}, S_{\pi(2)}), \cdots, (G_{\pi(k)}, S_{\pi(k)})) \equiv \text{view}_{\text{DC}}((G_1, S_1), (G_2, S_2) \cdots, (G_k, S_k)) \equiv \text{view}_{\text{DC}}((G_1, S_{\pi(1)}), (G_2, S_{\pi(2)}), \cdots, (G_k, S_{\pi(k)}))$，其他等价类同理。$S_i$具有敏感性，令$S_i{=}d_i^-$, $G_i{=}d_i^+$，则能构造数据$((d_1^+, d_1^-), (d_2^+, d_2^-), \cdots, (d_n^+, d_n^-))$，且满足$\text{view}_e((d_1^+, d_1^-), (d_2^+, d_2^-), \cdots, (d_n^+, d_n^-)) \equiv \text{view}_e((d_1^+, d_{\partial(1)}^-), (d_2^+, d_{\partial(2)}^-), \cdots, (d_n^+, d_{\partial(n)}^-))$，其中用$\partial$表示将$I$条记录进行任意排序，由于等价类的成员至少为$k$，因此可以满足$|I| \geq k$。

5 协议性能分析

5.1 计算复杂度

数据拥有者主要有3个任务：QI发布、GQI有效性验证、领导者选举，分别用C_{s}, C_{v}和C_{e}代表，其中C_{s}和C_{v}为简单任务复杂度为$O(1)$。采用文献[17]的领导者选举算法需要花费的代价为$O(u)$, u参与者的数量，因此在一个等价类中选举一个领导者，成员需要花费的代价为$O(k)$, k为等价类成员的数量。协议要分别进行两次领导者的选举，则领导者选举花费的代价为$O(2k)$，每个数据拥有者的C_{e}为$O(2k)$。在数据收集端，有3个主要的任务分别为匿名化、GQI分发、敏感属性值获取，分别表示为C_{a}, C_{d}和C_{f}。C_{a}依赖匿名化技术在$O(n)$到$O(n^2)$之

间变化，最坏的情况为$O(n^2)$，由于要将所有的GQI发送给相应的数据拥有者，所以C_{d}的代价应为$O(n)$，其中n为参与的数据收集者的数量。此外，敏感属性值搜索时间等同于树形结构的搜索复杂度为$O(\log_2 m)$，m为编码的敏感属性域的尺寸，共需要搜索n个用户的敏感属性值，则C_{f}为$O(n\log_2 m)$。数据拥有者需要花费的代价C_{e}为$O(2k)$，而数据收集者需要花费的时间代价为$O(n\log_2 m+n^2+n)$。显然协议花费的代价，主要被具有高存储计算和存储能力的数据收集端承担。

5.2 隐私保护

假设编码的敏感属性域尺寸为T，则需要的编码长度为至少为l，$l=\log_2\log_a T+\log_2 T$，树的分支为$a$。数据收集者获取敏感属性的位置信息所需的时间和编码长度为正相关的。编码长度关系隐私保护程度，随着编码长度增加，攻击者获取准确编码值需要花费的攻击次数也将增加，用户的敏感属性值的隐私泄露风险将降低。此外，编码长度为l，则攻击者随机猜测一个值等于目标的随机锚点或补偿距离的概率为$1/2^l$，而攻击者随机猜测一个敏感属性值为目标真实敏感属性值的概率为$1/T$。图4展示了a为2，3，4时随着属性域尺寸增加，编码长度的变化情况。由图4可以看出，随着属性域尺寸的增加，编码长度会不断增加，攻击者获取目标隐私数据花费的代价将会增加，因此，增加属性域尺寸能够增加隐私保护等级。此外，树形结构的分支越少，编码长度也越长，因此，适当减少树形分支个数，能够增加隐私保护等级，但要确保分支数的减少不会过多增加DC对目标敏感属性的搜索时间。由图5可以看出，随着属性域尺寸的增大，攻击者通过直接猜测敏感属性值和生成随机锚点或补偿距离获胜概率将下降，增加了攻击者通过树形结构编码方式获取目标隐私的困难程度，再次说明属性域尺寸的增加能够增加隐私保护程度，此外，攻击者猜测敏感属性值的获胜概率要大于猜测随机锚点或补偿距离的获胜概率，攻击者获取随机锚点

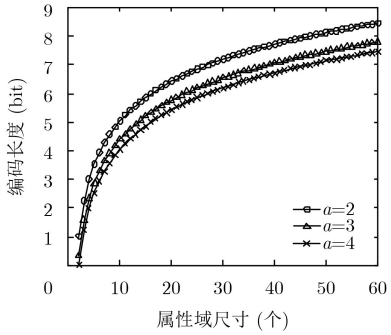


图 4 属性域尺寸和编码长度的关系

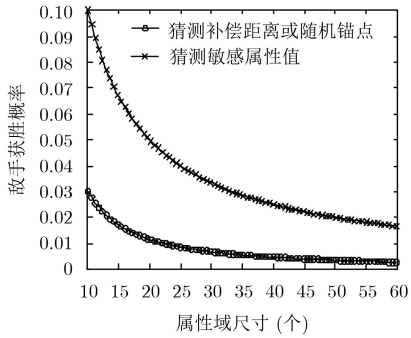


图 5 属性域尺寸和敌手获胜概率

或者补偿距离花费的代价，要高于直接猜测敏感属性值的代价，因此，树形结构的隐私保护程度不低于随机猜测的隐私保护程度，符合隐私安全需求。在公开电子医疗记录NPS数据集中，敏感属性域的尺寸为4848，基于此数量级假设攻击者想要猜测某用户的敏感属性值，其获胜的概率非常小。针对树形编码结构，数据领导者已经掌握了数据拥有者的敏感属性的部分份额，通过随机生成另一部分份额，获取完整的敏感属性，其获胜的概率为$\Pr(S^*=S)=1/2^l$，由图5可以看出，当属性域尺寸为60时，$\Pr(S^*=S)<0.02$，在实际数据中属性域尺寸足够大，$\Pr(S^*=S)<\xi,\xi$几乎可以忽略不计，因此方案的隐私保护程度能够满足实际的需求。

6 结束语

本文针对敏感数据收集过程中数据收集者半诚信可能间接造成隐私泄露问题，设计隐私保护的匿名化数据收集协议，协议实现半诚信的数据收集者最大化数据效用只能建立在K匿名的基础上，有效地降低了数据拥有者隐私泄露的风险。未来工作应该聚焦于个性化的隐私保护数据收集协议的设计，根据个人隐私需求的差异化设计不同匿名化等级，实现隐私保护力度自适应调节，避免过保护造成的数据效用低的现象。

参 考 文 献

[1] 曹珍富，董晓蕾，周俊，等. 大数据安全与隐私保护研究进展[J]. 计算机研究与发展，2016，53(10)：2137–2151. doi: 10.7544/issn1000-1239.2016.20160684.
CAO Zhenfu, DONG Xiaolei, ZHOU Jun, *et al*. Research advances on big data security and privacy preserving[J]. *Journal of Computer Research and Development*, 2016, 53(10): 2137–2151. doi: 10.7544/issn1000-1239.2016.20160684.

[2] 包国华，王生玉，李运发. 云计算中基于隐私感知的数据安全保护方法研究[J]. 信息网络安全，2017(1)：84–89. doi: 10.3969/j.issn.1671-1122.2017.01.013.
BAO Guohua, WANG Shengyu, and LI Yunfa. Research on

data security protection method based on privacy awareness in cloud computing[J]. *Netinfo Security*, 2017(1): 84–89. doi: 10.3969/j.issn.1671-1122.2017.01.013.

[3] IMRUL K and ADRIANA I. Privacy and security in online social networks: A survey[J]. *Online Social Networks and Media*, 2017, 4(3): 1–21. doi: 10.1109/ICME.2011.6012166.

[4] SWEENEY L. *k*-Anonymity: A model for protecting privacy[J]. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002, 10(5): 557–570. doi: 10.1142/S0218488502001648.

[5] MACHANAVAJJHALA A, GEHRKE J, KIFER D, *et al*. *l*-Diversity: Privacy beyond *k*-anonymity[C]. Proceedings of the 22nd International Conference on Data Engineering, Atlanta, USA, 2006: 24. doi: 10.1109/ICDE.2006.1.

[6] LI Ninghui, LI Tiancheng, and VENKATASUBRAMANIAN S. *t*-Closeness: Privacy beyond k-anonymity and *l*-diversity[C]. Proceedings of the 23rd International Conference on Data Engineering, Istanbul, Turkey, 2007: 106–115. doi: 10.1109/ICDE.2007.367856.

[7] DWORK C, KENTHAPADI K, MCSHERRY F, *et al*. Our data, ourselves: Privacy via distributed noise generation[C]. Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Petersburg, Russia, 2006: 486–503.

[8] DWORK C, NAOR M, PITASSI T, *et al*. Differential privacy under continual observation[C]. Proceedings of the 42nd ACM symposium on Theory of Computing, Cambridge, Massachusetts, USA, 2010: 715–724. doi: 10.1145/1806689.1806787.

[9] CLARKE A and STEELE R. A smartphone-based system for population-scale anonymized public health data collection and Intervention[C]. Proceedings of the 47th Hawaii International Conference on System Sciences, Waikoloa, USA, 2014: 2908–2917. doi: 10.1109/HICSS.2014.363.

[10] ZHONG Sheng, YANG Zhiqiang, and CHEN Tingting. *k*-anonymous data collection[J]. *Information Sciences*, 2009, 179(17): 2948–2963. doi: 10.1016/j.ins.2009.05.004.

[11] XUE Mingqiang, PAPADIMITRIOU P, RAÏSSI C, *et al*. Distributed privacy preserving data collection[C]. Proceedings of the 16th International Conference on Database Systems for Advanced Applications, Hongkong, China, 2011: 93–107.

[12] LI Hongtao, GUO Feng, ZHANG Wenyin, *et al*. (a, k)-Anonymous scheme for privacy-preserving data collection in IoT-based healthcare services systems[J]. *Journal of Medical Systems*, 2018, 42(3): 56. doi: 10.1007/s10916-018-0896-7.

[13] 刘琴, 刘旭辉, 胡柏霜, 等. 个人健康记录云管理系统中支持用户撤销的细粒度访问控制[J]. 电子与信息学报, 2017, 39(5): 1206–1212. doi: 10.11999/JEIT160621.

LIU Qin, LIU Xuhui, HU Baishuang, *et al*. Fine-grained access control with user revocation in cloud-based personal health record system[J]. *Journal of Electronics & Information Technology*, 2017, 39(5): 1206–1212. doi: 10.11999/JEIT160621.

[14] LUO Entao, BHUIYAN M Z A, WANG Guojun, *et al*. Privacy protector: Privacy-protected patient data collection in IoT-based healthcare systems[J]. *IEEE Communications Magazine*, 2018, 56(2): 163–168. doi: 10.1109/MCOM.2018.1700364.

[15] 龚奇源, 杨明, 罗军舟. 面向关系-事务数据的数据匿名方法[J]. 软件学报, 2016, 27(11): 2828–2842. doi: 10.13328/j.cnki.jos.005099.

GONG Qiyuan, YANG Ming, and LUO Junzhou. Data anonymization approach for microdata with relational and transaction attributes[J]. *Journal of Software*, 2016, 27(11): 2828–2842. doi: 10.13328/j.cnki.jos.005099.

[16] KIM S and CHUNG Y D. An anonymization protocol for continuous and dynamic privacy-preserving data collection[J]. *Future Generation Computer Systems*, 2019, 93: 1065–1073. doi: 10.1016/j.future.2017.09.009.

[17] VILLADANGOS J, CORDOBA A, FARINA F, *et al*. Efficient leader election in complete networks[C]. Proceedings of the 13th Euromicro Conference on Parallel, Distributed and Network-Based Processing, Lugano, Switzerland, 2005: 136–143. doi: 10.1109/EMPDP.2005.21.

[18] 罗恩韬, 王国军. 移动社交网络中一种朋友发现的隐私安全保护策略[J]. 电子与信息学报, 2016, 38(9): 2165–2172. doi: 10.11999/JEIT151479.

LUO Entao and WANG Guojun. A novel friends matching privacy preserving strategy in mobile social networks[J]. *Journal of Electronics & Information Technology*, 2016, 38(9): 2165–2172. doi: 10.11999/JEIT151479.

周治平：男，1962 年生，博士，教授，研究方向为检测技术与自动化装置、信息安全等.

李智聪：男，1992 年生，硕士生，研究方向为物联网安全认证.