软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(7):1963−1980 [doi: 10.13328/j.cnki.jos.005363]

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

面向云数据的隐私度量研究进展*

熊金波[1,2,3], 王敏燊[1], 田有亮[2], 马 蓉[1], 姚志强[1,3], 林铭炜[1]

[1](福建师范大学 数学与信息学院,福建 福州 350117)

[2](贵州省公共大数据重点实验室(贵州大学),贵州 贵阳 550025)

[3](福建省网络安全与密码技术重点实验室,福建 福州 350007)

通讯作者: 姚志强, E-mail: yzq@fjnu.edu.cn

摘 要: 隐私保护技术是云计算环境中防止隐私信息泄露的重要保障,通过度量这种泄露风险可反映隐私保护技术的隐私保护强度,以便构建更好的隐私保护方案.因此,隐私度量对隐私保护具有重大意义.主要对现有面向云数据的隐私度量方法进行综述:首先,对隐私保护技术和隐私度量进行概述,给出攻击者背景知识的量化方法,提出云数据隐私保护技术的性能评价指标和一种综合评估框架;然后,提出一种云数据隐私度量抽象模型,从工作原理和具体实施的角度对基于匿名、信息熵、集对分析理论和差分隐私这 4 类隐私度量方法进行详细阐述;再从隐私度量指标和度量效果方面分析和总结这 4 类方法的优缺点及其适用范围;最后,从隐私度量的过程、效果和方法这 3 个方面指出云数据隐私度量技术的发展趋势及有待解决的问题.

关键词: 隐私泄露;隐私度量;数据隐私;隐私保护;差分隐私

中图法分类号: TP309

中文引用格式: 熊金波,王敏燊,田有亮,马蓉,姚志强,林铭炜.面向云数据的隐私度量研究进展.软件学报,2018,29(7):1963−1980. http://www.jos.org.cn/1000-9825/5363.htm

英文引用格式: Xiong JB, Wang MS, Tian YL, Ma R, Yao ZQ, Lin MW. Research progress on privacy measurement for cloud data. Ruan Jian Xue Bao/Journal of Software, 2018,29(7):1963−1980 (in Chinese). http://www.jos.org.cn/1000-9825/5363.htm

Research Progress on Privacy Measurement for Cloud Data

XIONG Jin-Bo[1,2,3], WANG Min-Shen[1], TIAN You-Liang[2], MA Rong[1], YAO Zhi-Qiang[1,3], LIN Ming-Wei[1]

[1](College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117, China)

[2](Guizhou Provincial Key Laboratory of Public Big Data (Guizhou University), Guiyang 550025, China)

[3](Fujian Provincial Key Laboratory of Network Security and Cryptology, Fujian Normal University, Fuzhou 350007, China)

Abstract: Privacy protection technology is an important guarantee to prevent the privacy disclosure of sensitive information in the cloud computing environment. In order to design better privacy protection schemes, a privacy measurement technique is required that can reflect the privacy protection intensity by measuring the disclosure risk of privacy information in the privacy protection schemes. Therefore, privacy measurement is of great significance for the privacy protection of the cloud data. This paper systematically reviews the existing methods of

*基金项目: 国家自然科学基金(61772008, 61502102, 61370078, 61363068); 福建省自然科学基金(2015J05120, 2016J05149, 2017J05099); 贵州省公共大数据重点实验室开放课题基金(2017BD KFJJ028); 福建省高校杰出青年科研人才培育计划(2015, 2017); 贵州省科技拔尖人才项目(黔教合 KY[2016]060)

Foundation item: National Natural Science Foundation of China (61772008, 61502102, 61370078, 61363068); Natural Science Foundation of Fujian Province, China (2015J05120, 2016J05149, 2017J05099); Guizhou Provincial Key Laboratory of Public Big Data Research Fund (2017BDKFJJ 028); Distinguished Young Scientific Research Talents Plan in Universities of Fujian Province (2015, 2017); Science and Technology Top- Notch Talent Support Project in Guizhou Province Department of Education (黔教合 KY[2016]060)

本文由“面向隐私保护的新型技术与密码算法”专题特约编辑刘吉强教授推荐.

收稿时间: 2017-05-30; 修改时间: 2017-07-13, 2017-08-22; 采用时间: 2017-09-05; jos 在线出版时间: 2017-10-17

CNKI 网络优先出版: 2017-10-17 13:42:47, http://kns.cnki.net/kcms/detail/11.2560.TP.20171017.1342.013.html

privacy measurement for the cloud data. Firstly, an overview of the privacy protection and privacy measurement is provided along with descriptions of some quantitative methods of the background knowledge for the attacks, some performance evaluation indexes and a comprehensive evaluation framework of the privacy protection schemes for the cloud data. Moreover, an abstract model of the privacy measurement for the cloud data is proposed, and the existing privacy measurement methods are elaborated based on anonymity, information entropy, set pair analysis theory and differential privacy respectively from the perspective of working principle and the specific implementation. Furthermore, the advantages and disadvantages and the application scopes of the above four types of privacy measurement methods are analyzed by the privacy measurement indexes and effectiveness. Finally, the development trends and the future problems of the privacy measurement for the cloud data are summarized in terms of the privacy measurement processes, effects and methods.

Key words: privacy disclosure; privacy measurement; data privacy; privacy protection; differential privacy

2016 年,美国 Code 大会上发布消息称“2016 年全球互联网用户已经超过 30 亿,互联网全球渗透率高达 42%”.随着互联网+与大数据时代的到来,以云计算、大数据为代表的信息技术深刻改变了信息服务模式以及人们的学习、工作和生活方式,这些模式和方式的变革同时促进了云服务的优化与发展,不断推动云应用的广泛普及.云服务为实现数据的计算与存储、发布和共享等提供了极大的方便,越来越多的用户愿意将个人数据提供给医疗机构、银行、科研机构和大型网络企业等服务提供商,这些服务提供商通常采用私有云、公有云或混合云的方式对用户数据实施存储与管理[1].为了科学研究或其他方面的应用,云服务提供商需要面向大众发布和共享数据,这种通过云服务实现外包数据的存储、计算、发布和共享,不仅可以节省成本,也可以提高效率.但是,用户的个人数据被上传到云端,其所有权与管理权发生分离,用户将失去对个人数据的物理控制权[1].一方面,用户可能不经意间将包含隐私信息的数据未经处理上传给云服务提供商,由于云服务提供商不一定可信,因此存在隐私泄露风险;另一方面,攻击者采用数据挖掘等技术窃取外包数据中的隐私信息.特别是涉及用户薪资、健康状况、银行卡号和身份证号等隐私信息的数据通过云服务进行数据存储、共享和发布时容易被攻击者(本文将恶意泄露用户隐私信息的相关人员统称为攻击者)非法收集、传播与利用,将导致用户隐私信息被滥用与用户利益受损等严重问题.针对隐私信息泄露的问题,目前主要的隐私保护技术有基于数据匿名的隐私保护技术、基于数据失真的隐私保护技术、基于密码学的隐私保护技术以及它们之间的组合技术[2].隐私信息泄露涉及的因素繁多,尤其是在云计算和大数据环境中,设计出一种完美的隐私保护技术仍极具挑战性.目前,隐私保护领域的研究工作主要是设计一种具有最优的隐私保护强度、数据可用性以及处理开销平衡的隐私保护技术,其中,隐私保护强度是评估隐私保护技术最重要的性能指标.如何评价和优化面向云数据的隐私保护技术,亟需相关的隐私度量方法量化云数据隐私保护技术的隐私保护强度.将隐私保护强度直观反馈给研究者,以便评价和优化面向云数据的隐私保护技术.因此,隐私度量在云数据隐私保护研究中具有重大意义.

在隐私保护领域的研究中,隐私度量通常通过度量指标或度量方法揭示隐私保护方法中存在的隐私信息泄露风险,从披露隐私信息的角度侧面反映出隐私保护方法的隐私保护强度.隐私度量最早源于隐私匿名技术,在隐私匿名保护技术的研究过程中,隐私度量问题一直备受研究者关注.文献[3]提出,当数据中的隐私信息泄露风险为 0 时,该数据达到了完美隐私保护,完美隐私保护能够对数据中的隐私信息实现最大程度的保护;而把不采取任何保护措施的数据视为隐私信息泄露风险最大的数据.之后,学者们相继提出了基于信息熵、集对理论和差分隐私的隐私度量方法.传统的隐私度量方法主要针对小规模、结构化、存储在传统关系数据库中的数据;而云数据通常是企业级、大规模、多源多维、非结构化的数据,也可能是普通用户的小规模数据,还可能是经过加密处理后的数据.因此,相对传统的隐私保护度量技术,面向云数据的隐私度量既要考虑小规模、结构化数据的隐私泄露度量,又要考虑大规模、非结构化数据的隐私泄露度量.

本文以云数据为研究对象,对现有的隐私度量方法进行归纳与述评.首先概述面向云数据的隐私保护技术,并给出其性能评价指标与一种综合评价框架;然后详细分析基于匿名、信息熵、集对理论和差分隐私这 4 类隐私度量方法的理论基础、优缺点以及应用范围;最后对面向云数据的隐私度量的未来研究趋势进行展望与预测,以期为科研人员准确把握该领域最新研究动态和未来发展方向提供借鉴.

本文第 1 节对隐私保护技术和隐私度量进行概述,详细阐述攻击者拥有的背景知识的表达与量化,并提出一种面向云数据的隐私保护技术的综合评价框架.第 2 节首先提出隐私度量的抽象模型,然后系统阐述典型的 4

类隐私度量方法的理论基础和实现细节.第 3 节分析总结上述 4 类隐私度量方法的度量指标、度量效果、优缺点以及应用范围.第 4 节指出该领域将来的研究方向.第 5 节总结全文.

1 云数据隐私保护方法

隐私信息通常指的是数据中用户不愿公布的个人身份和敏感属性相关联的信息[2].包含隐私信息的数据,如医疗数据、教育数据和用户注册数据等.这些数据包含多个元组,其中每一个元组对应一个用户,每一个元组包含多个属性.这些属性可分为 3 类:① 显示标识符(explicit identifier):能够唯一标识一个用户身份的属性,如姓名、身份证号码等;② 准标识符(quasi-identifier):不能唯一标识一个用户身份的属性,需多个属性组合才能唯一标识一个用户身份,如地址、性别和生日等;③ 敏感属性(sensitive attribute):涉及隐私信息的属性,如薪资、健康状况和财务信息等.

1.1 隐私保护技术概述

用户的个人数据发布到云端前,为保护数据中的隐私信息,需使用隐私保护技术对数据进行处理.常见的面向云数据的隐私保护技术可以分成 3 类[2,4,5]:基于数据匿名的隐私保护技术、基于数据失真的隐私保护技术和基于密码学的隐私保护技术.

(1) 基于数据匿名的隐私保护技术.通过对数据进行抑制与泛化等操作隐藏隐私信息,代表性的技术有 *k*-anonymity、*l*-diversity 和 *t*-closeness 等.与传统的数据匿名技术相比,现有的云数据匿名技术的模型、思想融合了大数据计算的相关技术.在云计算环境中,数据具有多维多源、大规模等特征,这使得匿名技术的效率变得至关重要.文献[6]采用分布式计算模型设计并实现了大数据的匿名系统,文献[7]使用多线程技术对匿名技术进行并行化处理,大幅度提高了大数据的匿名效率.

(2) 基于数据失真的隐私保护技术.通过对数据添加噪声等操作隐藏隐私信息,该技术能够保持在某些数据的总体特征或数据属性不变的情况下对数据进行干扰,当干扰越大时,数据失真越大,隐私保护强度越高,但数据可用性越低.代表性技术有差分隐私保护技术等.文献[8]提出了一种基于差分隐私的拓展技术,实现了对大规模多维数据的隐私保护.

(3) 基于密码学的隐私保护技术.通过对数据加密的方式保护隐私信息.该技术具有高隐私保护强度,但对云数据进行加、解密,需要进行大量复杂的运算,计算开销较大,且处理数据的效率低,代表性技术有收敛加密技术和同态加密技术等.文献[9]描述了同态加密技术在云数据中的应用,该加密方法可使密文数据在未解密的前提下直接进行运算,从而实现了在对密文数据进行计算的同时保护数据隐私.

1.2 隐私保护技术的性能评价指标

在云计算环境中,数据存储和计算、发布等均由云服务器处理.通过云服务可以提高数据的计算和发布的效率,但用户的隐私信息也存在被披露的风险.因此,在发布或共享数据前,需要使用隐私保护技术对数据的隐私信息进行处理,保护隐私信息的同时保证数据可用性和减少数据处理过程的开销.隐私保护技术的目标主要是防止数据中隐私信息的泄露,因此,隐私保护技术最主要的性能指标为隐私保护强度;其次,还需综合考虑数据可用性与处理开销.

(1) 隐私保护强度 $1/R(S)$.主要通过隐私信息泄露的风险 $R(S)$来加以反映[2],隐私信息泄露的风险越小,隐私保护强度越高.

(2) 数据可用性 I.主要通过隐私保护技术处理后数据的缺损来加以反映[10],数据缺损越高,数据可利用率则越低.数据可用性的度量方法有:分辨率度量 DM[11,12]、数据信息损失 ILoss[13,14]、分类度量 CM[15]以及比较重构数据与原始数据的差异度[16]等.

(3) 处理开销 C.数据计算开销、通信开销和存储开销是衡量隐私保护技术可行性的重要指标[17].计算开销指的是隐私保护技术处理数据时需占用的计算资源,通信开销指的是在云计算环境中用户与云端之间传输数据所需通信量,存储开销指的是云端存储经隐私保护处理后的数据空间大小.

将以上 3 个指标结合起来,采用线性加权法构建隐私保护技术性能综合评价表达式如下所示:

$$E=\alpha\frac{1}{R(S)}+\beta I+\gamma C,\ \alpha+\beta+\gamma=1 \tag{1}$$

其中,E 为隐私保护技术的评价值,α、β和γ为权重系数.通过 E 整体评估隐私保护技术的性能,权重α、β和γ可根据隐私保护需求设定.

1.3 攻击者背景知识的量化

隐私保护技术的主要目的是保护数据中隐私信息,防止攻击者窃取用户的隐私信息,导致用户利益受损,因此,攻击者的攻击能力对隐私信息泄露将产生直接影响.现有的方法研究隐私度量中攻击者模型主要从攻击者是否拥有背景知识与拥有背景知识的量的角度来分析[18,19].原始数据经过隐私保护技术处理得到观察数据,攻击者结合背景知识和观察数据,利用推理方法推测原始数据中的隐私信息.攻击者拥有的背景知识越多且背景知识的内容与隐私信息关联性越大,推测出隐私信息的可能性越大.因此,在隐私度量中引入攻击者拥有的背景知识可以更准确地描述隐私泄露的程度.由于很难预测攻击者拥有的背景知识的具体内容以及背景知识的量,导致攻击者的背景知识的准确量化极具挑战性,所以在隐私度量中,先对攻击者能够获取的知识做出假设,再利用关联规则等方法对背景知识进行量化.

1) 攻击者背景知识的构成

攻击者的背景知识一般由先验知识、观察知识和后验知识构成.先验知识包括公共信息和用户资料,如社交网络和购物网站提供的姓名、家庭地址和工作单位等信息,攻击者可通过网络搜索和调查获得.观察知识是指攻击者通过网络攻击和冒充合法用户请求数据服务等方式截获的观察数据内容,包括查询内容和敏感属性信息等.后验知识是指攻击者利用先验知识和观察知识,通过推理方法获得的新知识.

2) 攻击者背景知识的量化

背景知识的量化实质是将先验知识、观察知识和后验知识与隐私信息之间的关联性用条件概率描述[20].关联规则(association rule)[24]是一种通过计算关联属性之间的条件概率或组合条件概率的大小表示属性之间的关联程度.将关联规则应用到量化背景知识的过程如下.

(1) 假设用数据集合描述攻击者的先验知识和观察知识,每一个子数据集对应一项独立知识,如:用户身份集合 $U=\{u_1,u_2,...,u_n\}$, 数据集 $D=\{d_1,d_2,...,d_n\}$ 中的子数据集分别表示攻击者的一项知识.

(2) 攻击者将先验知识与观察知识进行关联,如:攻击者在时刻 t 观察到数据 d_t 中有用户 u_k,假设该随机事件为 $\{d_t\,|\,u_k\}$, 其发生的概率为 $P(d_t\,|\,u_k)$.

(3) 分析各随机事件之间的相互关联性,用条件概率表示任意两个事件之间的关联强度.如:攻击者观察到用户 u_k 在连续两个观测时刻 t 和 t+1 分别位于数据 d_t 和 d_{t+1} 中,即 $\{d_t\,|\,u_k\}$ 和 $\{d_{t+1}\,|\,u_k\}$. X_1 和 X_2 的关联强度表示为 $P(\{d_{t+1}\,|\,u_k\}\,|\,\{d_t\,|\,u_k\})$.

(4) 统计攻击者知道的所有事件间的关联,计算其关联的所有条件概率或组合条件概率来量化攻击者所拥有的背景知识.

Du 等人[21]提出了一种 TOP-(K_+,K_-)最强关联规则方法,该方法通过引用关联规则分析背景知识中非敏感属性值和敏感属性值之间的条件概率来量化背景知识,并证明了攻击者拥有的背景知识越多,隐私信息越容易被披露.王等人[22]在文献[21]的基础上不仅考虑了背景知识所包含的关联规则数目和强度,还考虑其概率分布,将所有关联规则根据规则强度分为 n 类,n 值越大,各关联规则之间区分得越细致,使背景知识的量化更加准确.Li 等人[23]分别引用了 Kernel 回归估计方法和大约推断法(Ω-estimate)以量化攻击者背景知识中的先验知识和后验知识.先验知识和观察知识的量化过程:假设原始数据集 $D=\{d_1,d_2,...,d_n\}$, 数据子集 d_i 中每个元组关联一个用户信息. d_i 中包含 n 个准标识符(QI)属性 $d[QI]=\{A_1,A_2,...,A_n\}$ 和敏感属性 A_s,定义 $T[A_i](1<i<n)$为准标识符属性 A_i 的属性值域;定义 $T[A_s]$为敏感属性 A_s 的属性值域.所有的准标识符属性值 $T[QI]$和所有敏感属性 $T[A_s]$ 的概率分布为$\Sigma=\{(P_1,P_2,...,P_m)\,|\,\Sigma_{1\leqslant i\leqslant m}P_i=1\}$.根据关联规则给出攻击者的先验知识的量化函数 $P_{pri}:T[QI]\rightarrow\Sigma$, 其中, P_{pri} 采用回归估计 Kernel 函数(kernel regression estimation)进行大约估计,分别对一维数据和多维数据进

行分析并给出计算方法.基于上述先验知识,提出了一种基于贝叶斯推理原理的大约推断法以估计攻击者拥有的后验知识.

1.4 隐私度量概述

隐私度量主要是度量隐私保护技术的隐私保护强度,该强度通常用攻击者披露隐私信息的风险大小来描述,即通过攻击者从被保护数据中披露隐私信息的风险量来侧面加以反映.假设具有相关背景知识 K 的攻击者 F_K 结合观察数据 D' 与背景知识 K 能够披露的隐私信息为 S,则隐私信息 S 被披露的风险可表示为

$$R(S)=Pr(F_K,D') \tag{2}$$

当 $R(S)$越大时,隐私信息被泄露的风险越大,隐私保护技术的隐私保护强度越小;反之,隐私保护技术的隐私保护强度就越大.

1.5 云数据隐私保护技术综合评价框架

通过对隐私保护技术的隐私保护强度、数据可用性和数据处理开销这 3 项指标进行分析,本文提出一种面向云数据的隐私保护技术综合评价框架,如图 1 所示.

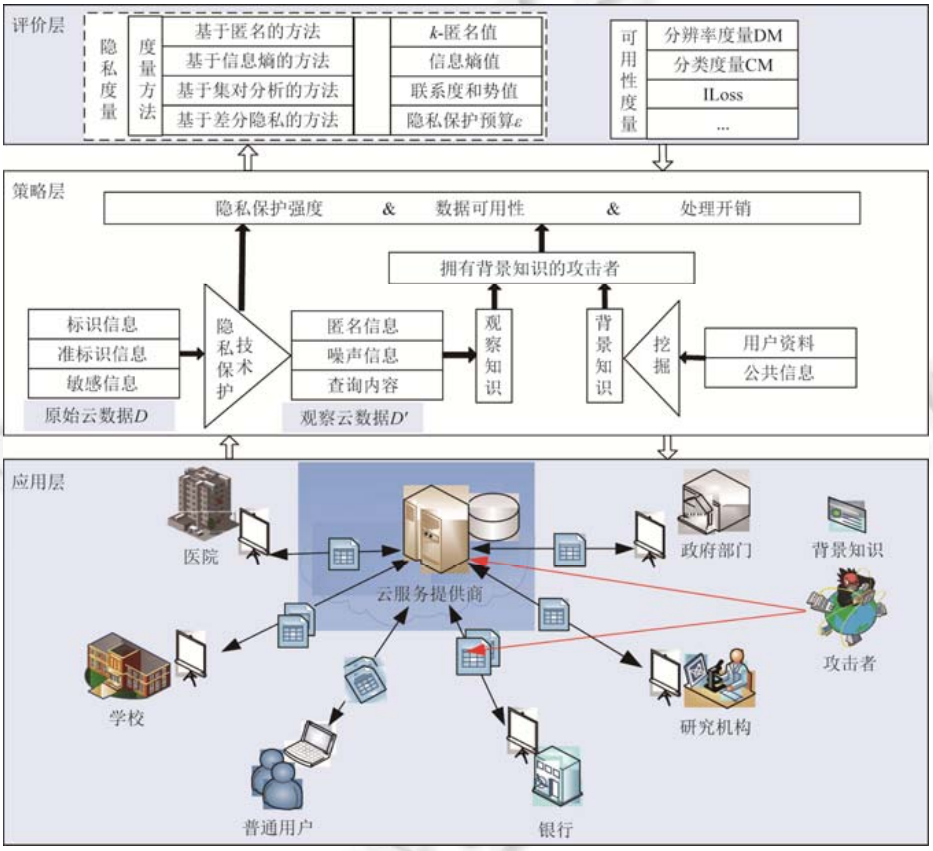

Fig.1 An evaluation framework of privacy protection scheme for cloud data

图 1 面向云数据的隐私保护技术综合评价框架

该框架涵盖如下 3 个层面:① 应用层,涉及数字银行、健康医疗和智慧教育等云应用与云服务提供商之间的数据存储、计算和发布等,包含用户、攻击者和云服务提供商 3 个参与实体;② 策略层,涉及应用层中使用的隐私保护技术和攻击者,用户根据隐私保护需求选择合适的隐私保护技术,而攻击者结合背景知识和观察数据

披露隐私信息;③ 评价层,主要针对策略层中的隐私保护技术采用相应的度量方法或度量指标对其隐私保护强度进行分析量化,然后分析数据可用性和隐私保护技术处理数据的开销,最后结合 3 方面因素综合分析评价隐私保护技术的性能.该评价框架的 3 个层面是一种交互模型.应用层中的用户对隐私保护需求会影响策略层中隐私保护技术的选择与构建;策略层中隐私保护技术的隐私保护强度、数据可用性和处理开销均由评价层的度量方法度量;评价层将隐私保护技术性能的评估结果反馈给研究者,研究者根据用户的实际需求对隐私保护技术进行重新选择与优化,从而形成闭环,为用户提供更好的隐私信息保护.

2 云数据隐私度量方法

不同的隐私保护技术在不同的隐私需求下,处理数据的方式和性能不同,目前还没有一种通用度量方法能够度量所有面向云数据的隐私保护技术的隐私保护强度.本文分析现有的隐私度量方法,抽象出一种面向云数据的隐私度量模型,如图 2 所示.该模型中结合云数据隐私保护技术和攻击者的背景知识,通过度量方法量化隐私保护技术的隐私保护强度.根据隐私度量方法的不同特点,可分成基于匿名的隐私度量方法、基于信息熵的隐私度量方法、基于集对分析理论的隐私度量方法和基于差分隐私的隐私度量方法.

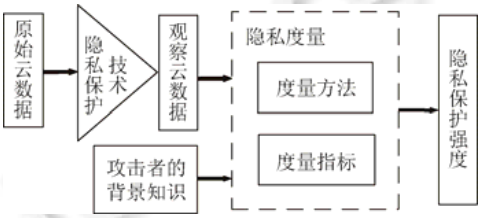

Fig.2 Model of privacy measurement for cloud data

图 2 云数据隐私度量模型

2.1 基于匿名的隐私度量方法

基于匿名的隐私度量方法根据匿名云数据的 *k*-匿名值和敏感属性值分布的概率分析数据的匿名程度;分析匿名程度和攻击者拥有的相关背景知识,使用贝叶斯推理等方法推测隐私信息;通过比较推测的信息与隐私信息之间的差异度反映隐私保护强度.基于匿名的云数据隐私度量框架,如图 3 所示.

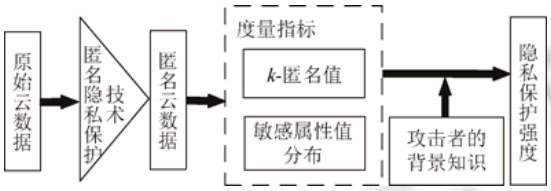

Fig.3 A privacy measurement framework based on anonymity for cloud data

图 3 基于匿名的云数据隐私度量框架

2.1.1 *k*-匿名与敏感属性值分布

k-匿名值是隐私度量中的经典指标之一.在 *k*-匿名(*k*-anonymity)中,*k*-匿名值表示数据集中的准标识符属性的匿名程度,每个匿名数据集中至少有 *k*–1 个不可识别的元组,并且 *k* 个匿名元组被识别的概率相等,没有相关背景知识的攻击者从这 *k* 个元组中识别出一个元组的隐私信息的概率为 $1/k$[25].匿名数据中的 *k*-匿名值越大,攻击者越难推测出隐私信息,则隐私保护强度越高.由于 *k*-匿名只对数据中准标识符属性进行匿名处理,没有对敏感属性作任何约束,攻击者利用与隐私信息相关的背景知识,根据匿名数据集中敏感属性值分布能够推测出用户与敏感属性值之间的对应关系[26].因此,仅依赖 *k*-匿名值作为匿名数据集的度量指标不够全面且度量结果不够准确.

满足 *k*-匿名的条件下,对数据集中各敏感属性值进行约束,其分布越均匀,数据集的匿名程度越高.*l*-多样性

(*l*-diversity)[26]对匿名数据集中出现频率高的敏感属性值进行约束,将每个等价类(数据集满足 *k*-匿名时,数据中在准标识符上具有相同值的元组的集合)包含至少 *k* 个匿名准标识符和至少有 *l* 个不同的敏感属性值作为匿名云数据的匿名程度的指标.Li 等人[27,28]基于 *k*-匿名和 *l*-多样性提出了一种基于计算敏感属性值分布的度量方法,在满足 *k*-匿名的条件下,引用 EMD(earth mover's distance)方法计算数据中敏感属性值的全局分布和任意等价类中同一敏感属性值分布的差异度.*k*-匿名值越大,差异度越小,匿名程度越高,隐私信息泄露风险越小.张等人[29]指出文献[27,28]采用的 EMD 方法没有考虑等价类与数据间敏感属性值分布的稳定性,针对该问题,基于 EMD 方法和 KL 散度提出了一种度量方法 EKD(EMD and KL divergence distance).其中,EMD 方法计算敏感属性值间的分布差异度,KL 散度衡量相邻的敏感属性值分布的稳定性差异.数据集中属性值的多样化程度影响了隐私保护的程度,匿名数据元组中的非敏感属性匿名程度越高,等价类中各敏感属性值的分布差异越小,攻击者能够推测出隐私信息的可能性越小.

2.1.2 基于贝叶斯推理的度量方法

根据前面分析的 *k*-匿名值和敏感属性值分布的概率,文献[30−32]提出了一种基于贝叶斯推理的度量隐私信息泄露的方法.首先根据攻击者拥有的背景知识与匿名数据的元组分别构建攻击者背景知识和匿名数据的二叉树图;再构建通过贝叶斯推理推测出信息关联的二叉树图;通过分析比较推测的信息与隐私信息之间的差异度来度量隐私信息泄露的风险,两者之间的差异度越小,隐私信息泄露的风险越大.度量过程如图 4 所示.

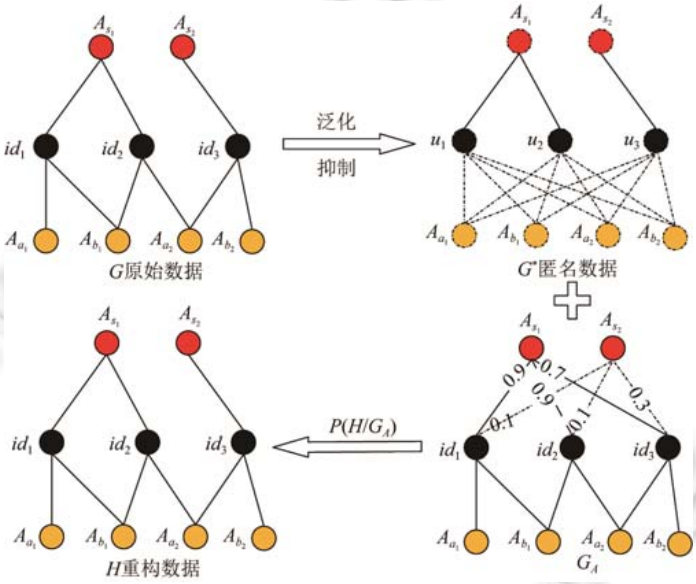

Fig.4 Anonymous privacy disclosure measurement based on Bayesian inference

图 4 基于贝叶斯推理的匿名隐私信息泄露度量

假设原始云数据中有 3 个用户的显示标识符为 id_1、id_2、id_3,有 A_a 和 A_b 两种准标识符属性和一种敏感属性 A_s,构建二叉树图 *G*.由匿名隐私保护技术分别对原始数据的显示标识符属性和准标识符属性进行处理,用编号 u_1、u_2、u_3 隐藏显示标识符.当攻击者缺少相关背景知识时,推测 id_1 对应敏感属性值 A_{s_1} 的概率为 0.33.当攻击者拥有相关背景知识时,构建基于背景知识推测用户与敏感属性值之间关联的二叉树图 G_A.图 G_A 中,显示标识符 id_1、id_2、id_3 和敏感属性值 A_{s_1} 和 A_{s_2} 之间的实线表示两者之间存在对应关系,虚线表示两者之间不存在对应关系,边权重值表示两者对应关系的概率.通过贝叶斯推理理论推导出拥有背景知识的攻击者重构用户与敏感属性值的关联概率为 $P(H|G_A)$.由图 4 中的例子可得,$P(H|G_A)$=(0.9×0.9×0.3)/((0.9×0.9×0.3)+(0.1×0.9×0.7)+(0.9×0.1×0.7))=0.66,拥有背景知识的攻击者推测 id_1 对应敏感属性值 s_1 的概率值由 0.33 增至 0.66.通过推测隐私信息的先验概率 *P* 和后验概率 $P'|G_A$ 的差异度来表示用户的隐私信息泄露风险,表示为

$$\delta(P, P' \mid G_A, G^*) \leqslant P(id_i, A_{s_j}) \to \delta(w_{ij}, w'_{ij}) \leqslant P(id_i, A_{s_j}) \tag{3}$$

其中,$\delta(x,y)=|x-y|$是对隐私信息泄露的度量,w_{ij}'表示攻击者推测用户显示标识符和敏感属性值对应关系的边权重值,$P(id_i, A_{s_j})$表示用户显示标识符和敏感属性值关联概率,$P(id_i, A_{s_j})$越小,隐私信息被泄露的风险越小.

2.2 基于信息熵的隐私度量方法

信息熵(information entropy)是一种量化信息不确定性的方法[33].隐私信息属于信息,因此可以通过信息熵来量化隐私信息的不确定性.熵值越大,表示隐私信息在云数据中的不确定性越高,越不易被披露,隐私信息泄露的风险越低.当存在攻击者且拥有背景知识时,引入条件熵可度量攻击者根据背景知识推测出隐私信息的不确定性.互信息用于度量原始云数据中隐私信息和观察数据中隐私信息之间相交互的平均信息量,描述隐私信息的整体泄露风险.

2.2.1 基于信息熵的度量方法

定义 1(信息熵($H(X)$)). 假设存在随机变量$X=\{x_1,x_2,\ldots,x_n\}$,随机变量X的概率为$P=\{P(x_1),P(x_2),...,P(x_n)\}$,定义各个信息的自信息量的数学期望为信息的平均自信息量,称为无条件熵[33].信息熵的单位由自信息的单位来决定,即取决于对数选取的底,为了方便,计算通常取 2.表达式如下:

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \tag{4}$$

Clauß 等人[34]引用信息熵描述数据集中隐私信息的不确定性,假设数据集中各隐私信息集合用 $S=\{s_1, s_2,\ldots,s_n\}$表示,$P(s_i)$表示推测隐私信息$s_i$的概率,通过信息熵公式(4)计算数据集中隐私信息的不确定性.信息熵值越大(当各用户的显示标识符与敏感属性值的关联性的概率为等概率时,熵值最大),隐私信息的不确定性程度越高,攻击者越难推测出隐私信息,隐私信息泄露的风险越低;反之,信息熵值越小,隐私信息泄露的风险越高.彭等人[35]为了使信息熵的度量更直观,将隐私保护系统描述成一种通信模型,如图 5 所示.攻击者无任何与数据信源中隐私信息关联的背景知识,仅通过分析数据信宿中的观察数据来披露隐私信息,用信息熵描述隐私信息在数据信宿中的平均隐私信息量,隐私信息的不确定性直接反映隐私信息泄露的风险.张等人[36]指出数据的动态操作(插入、删除和修改等)将会影响敏感属性值在数据中分布不均匀的问题.针对由敏感属性值分布不均匀导致的隐私信息泄露问题,提出了一种方法引用信息熵度量云数据操作过程中的隐私信息的泄露风险.首先假定一个数据集 $D=\{d_1,d_2,\ldots,d_n\}$,经过动态操作后对应的数据子集中敏感属性值出现的概率为$\{P(d_1),P(d_2),\ldots,P(d_n)\}$,当各子集取值出现的概率是等概率时,信息熵取值最大,隐私信息泄露的风险最低;反之,概率分布越不均匀,信息熵取值越小,隐私信息泄露的风险越高.

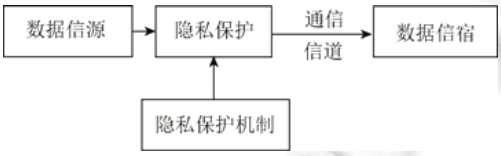

Fig.5 Communication model of privacy protection

图 5 隐私保护通信模型

2.2.2 基于条件熵的度量方法

定义 2(条件熵($H(X/Y)$)). 假设存在随机变量 $X=\{x_1,x_2,\ldots,x_n\}$,再给定一个随机变量 $Y=\{y_1,y_2,...,y_n\}$,如果已知 Y 的值,条件熵表示通过 Y 推测出 X 的平均不确定性[33].表达式如下:

$$H(X/Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log_2 p(x_i / y_j) \tag{5}$$

存储同一用户信息的各数据集之间不可避免地存在一定程度的关联性.假设数据 D 与 D' 之间的依赖概率为 $P(D/D')$,数据 D 中某一属性值 A_1 以概率 $P(A_1/A_2)$ 依赖于数据 D' 中某一属性值 A_2,则 $H(D/D')$ 表示在已知

数据 D' 的条件下,对数据 D 的不确定性.当 $H(D/D')$ 取得最小值 0 时,D 可由 D' 推测出来;反之,当 $H(D/D')$ 取得最大值时,D 和 D' 相互独立,攻击者已知 D' 很难推测出 D[37].基于云数据间的关联性问题,张等人[36]引用条件熵实现了一种云数据操作过程中的隐私信息泄露的风险的度量,数据进行增加、删除和修改等操作之后,利用条件熵描述攻击者经过长期发送请求服务获取数据 D' 后,对数据 D 中的隐私信息 S 仍然存在的不确定性.在对数据进行增加、修改和删除等操作之后,如果数据间的条件熵值不减少,则动态数据的隐私泄露风险不会增大.引用信息熵度量的不确定程度只考虑了隐私保护技术的保护程度在攻击者无相关背景知识条件下的隐私度量,对攻击者拥有的背景知识缺乏考虑.实际应用中,攻击者往往会结合背景知识对数据进行分析,与隐私信息相关的背景知识会增大隐私信息泄露的风险.针对这一问题,文献[35]基于条件熵提出一种拥有背景知识攻击者攻击的隐私度量方法,并构建了包含攻击者的隐私保护信息熵模型.假设攻击者已获得了背景知识 K,利用背景知识推测数据中隐私信息 S 的不确定程度为 $H(S/K)$.拥有背景知识 K 的攻击者结合已获得的数据中与隐私信息关联的数据 D' 推测隐私信息 S,用背景知识攻击条件熵 $H(S/D'K)$ 描述攻击者推测原始数据 D 中的隐私信息 S 存在的不确定性,$H(S/D'K)$ 值越大,隐私信息泄露的可能性越小,隐私保护强度越高.

2.2.3 基于互信息的度量方法

定义 3(互信息($I(X;Y)$)). 假设随机变量 $Y=\{y_1,y_2,...,y_n\}$,随机变量 $X=\{x_1,x_2,\ldots,x_n\}$.互信息表示从 Y 中获得关于 X 的平均信息量,即获得 Y 前、后,关于 X 的不确定性减少的量[33].表达式如下:

$$I(X;Y)=\sum_{j=1}^{m}\sum_{i=1}^{n}p(x_iy_j)\log_2\frac{p(x_i/y_j)}{p(x_i)}=H(X)-H(X/Y) \tag{6}$$

用互信息 $I(X;Y)$描述隐私信息的泄露风险[38−42],假设原始云数据 D 中隐私信息为 S,攻击者获取到隐私信息关联的数据 D' 后,原始云数据经过隐私保护技术处理后,通过互信息计算攻击者在已知信息 D' 之前和之后,关于隐私信息 S 在原始云数据中不确定性减少的量反映隐私泄露的风险.不确定性减少的量越多,隐私信息泄露的风险越大.文献[43,44]利用互信息量化攻击者拥有的数据 D' 和原始云数据 D 之间相互独立的程度.假设攻击者从 D' 中获得与隐私信息 S 具有关联的子集为 d.攻击者通过子集 d 推测原始云数据 D 中隐私信息 S 不确定性采用 $H(S/D'=d)$表示.用 $I(S;D'=d)=[H(S)-H(S/D'=d)]$表示已知 D' 中的子集 d 隐私信息的不确定性减少的量.通过互信息可以反映数据子集 d 和隐私信息 S 之间相互独立的程度,当互信息值越大,隐私信息 S 在数据中的不确定性减少的量越多,两者之间的关联性越强,隐私信息泄露的风险越高.

2.3 基于集对分析理论的隐私度量方法

集对分析理论(set pair analysis)是一种定性定量相结合并能解决确定与不确定性问题的分析方法[45].该理论的基本思想是:假定有两个数据集,分析它们的特性并得到两个集合所有的特性总数为 N,同一特性个数为 M,对立的特性个数为 Q,其余不相互对立且不相互同一的特性个数为 L,分别称其比值 M/N、L/N 和 Q/N 为特定问题背景下的两个集合的同一度、差异度和对立度.再利用集对的联系度对数据集间的特性作定量刻画.表达式如下:

$$\mu=\frac{M}{N}+\frac{L}{N}i+\frac{Q}{N}j,\ \text{其中},\frac{M}{N}+\frac{L}{N}+\frac{Q}{N}=1 \tag{7}$$

其中,μ 为两数据集之间的联系度,i 为差异度标记,j 为对立度标记,当 i 和 j 作为系数参加运算时,定义 $j=-1$,i 在区间[−1,1]依条件取值.为简便计算,令 $a=M/N$、$b=L/N$、$c=Q/N$,则式(7)可写成 $\mu=a+bi+cj$.联系度 μ 能够反映数据与所在范围、数据集与数据集之间的确定性与不确定性的相互联系,因此可用来描述两数据集之间的关联隐私信息的不确定性.

文献[46]基于集对分析理论提出了一种集对分析隐私度量方法,该方法是对数据集之间的联系关系进行度量分析.首先分析数据集的相邻数据子集对所具有的同一度、差异度和对立度特性并加以量化描述,得到联系度表达式.为防止数据集中起始值较大或较小的属性对权值的影响,通过数字或特殊符号对数据集中存在的不一致和噪声数据进行数据处理.对准标识符属性值的处理方式是取集合中的最小值 $\min B_i$、最大值 $\max B_i$ 和实际值 B_i 构造出一个三元区间,即:$[B_i]=[\min B_i,B_i,\max B_i]$. 然后利用集对分析理论对数据集中各属性值建立一个

集对联系度表达式,将三元数区间转化为

$$\mu(B_i)=a_{B_i}+b_{B_i}+c_{B_i} \tag{8}$$

其中,当 $B_i\neq 0$ 时,

$$a_{B_i}=\frac{B_i}{\max B_i+\min B_i},\ b_{B_i}=\frac{(\max B_i-B_i)(B_i-\min B_i)}{(\max B_i+\min B_i)B_i},\quad c_{B_i}=\frac{\max B_i\cdot\min B_i}{(\max B_i+\min B_i)B_i} \tag{9}$$

当 $B_i=0$ 时,

$$a_{B_i}=\frac{B_i}{\max B_i+\min B_i},\ b_{B_i}=\frac{\min B_i}{\max B_i+\min B_i},\ c_{B_i}=\frac{\max B_i}{\max B_i+\min B_i} \tag{10}$$

将各准标识符属性值通过数据集结方法进行集结,获得集结之后的联系度表达式为

$$\mu(B)=a_B+b_Bi+c_Bj \tag{11}$$

然后计算出表达式中同一度、差异度和对立度之间的势值或记分函数来反映隐私保护的泄露程度.定义联系度的势值 $Shi(\mu)=a_B/c_B$ 和记分函数 $S(\mu)=a_B-c_B$ 作为度量指标.根据计算的势值大小将隐私信息分为同势、均势和反势,从度量数据发布或共享过程中的隐私信息泄露风险可分析得到:同势表示隐私信息泄露的风险较低,反势表示隐私信息泄露的风险较高,均势介于同势与反势之间.

2.4 基于差分隐私的隐私度量方法

定义 4(差分隐私(differential privacy))[47,48]. 存在两个相邻数据集 D_1 和 D_2 且两数据集间有差别的元组最多为一条.给定随机算法 R,$range(R)$是随机算法 R 所有可能输出结果的集合.若随机算法 R 对于任何一对相邻数据集 D_1 和 D_2 的任意输出结果为 $O(O\subseteq range(R))$,且满足下列不等式,则随机算法 R 满足ε-差分隐私保护.

$$\Pr[A(D_1)\in O]\leqslant e^{\varepsilon}\cdot\Pr[A(D_2)\in O] \tag{12}$$

其中,$\Pr[\cdot]$表示隐私信息泄露的风险,由随机算法 R 随机控制;参数 ε 为差分隐私保护预算.ε取值越小,隐私保护强度越高.

差分隐私保护技术在实现隐私保护的同时,提供了一种可量化、评估隐私信息泄露风险的方法.该技术通常是通过设置不同的ε值实现隐私保护强度的划分,ε越小,加入的噪声越多,隐私信息泄露的风险越低,隐私保护强度越高.差分隐私保护技术常用的噪声添加机制有指数机制和拉普拉斯机制,数据添加噪声量的大小由全局敏感度Δf(添加或删除数据集中任一元组对算法输出值所造成的最大改变值)和ε共同确定[49].差分隐私保护预算ε常用的分配策略有线性分配、均匀分配、指数分配、自适应分配以及混合策略分配等[50].文献[51,52]结合差分隐私算法中ε可控的特点,根据用户对数据隐私保护强度的要求,通过调整噪声的分配策略生成符合 Lap$(\Delta f/\varepsilon)$分布的噪声;将ε值反映的隐私保护强度反馈给用户,用户根据不同隐私保护强度,将数据提供给具有不同数据权限的数据请求者,隐私保护预算ε提供了直观的隐私保护强度,进而实现数据隐私保护强度的可控性和数据与数据请求者的对应性.在数据中添加噪声,不可避免地会造成数据信息可用性的损失.当ε越小时,添加的噪声越多,隐私保护强度越高,数据信息损失越大.反之,隐私保护强度随之降低,数据信息损失也减少.

2.4.1 关联差分隐私的度量方法

由差分隐私定义可知,当数据集中的元组之间相互独立时,差分隐私保护和隐私度量无需考虑攻击者拥有的背景知识的内容以及背景知识的量.即便攻击者获取了除一个元组之外的元组信息,该元组的隐私信息也无法披露.但文献[53−56]研究发现,当数据中的元组之间存在关联性时,攻击者可以通过观察到数据中的一条或者多条元组推测出与之关联的元组的隐私信息.因此,差分隐私保护下,数据的隐私保护强度的度量不仅受限于自身的ε,还受与之关联数据集的ε'的影响.针对存在部分用户的信息在不同数据的元组中具有直接关联性或间接关联性,文献[57]提出了一种多数据集关联的差分隐私度量方法,用于度量某一数据集中的元组受其他数据集中元组影响时的隐私信息泄露的风险.假设 $D=\{d_1,d_2,\ldots,d_n\}$表示一系列数据集,数据集中包括一条或者多条用户元组或者与之关联用户的元组.数据集 d_i^1 和 d_i^2 $(1\leqslant i\leqslant n)$为数据集 d_i 中相邻的两个数据子集,并且两数据集中的元组之间至多有一条存在差异,其他所有的元组完全相同.一种关联隐私保护机制的随机算法 R,作用域范

围为 $range(R)$,则将关联隐私保护机制 R 的隐私信息泄露的程度表示为 $CDPL(R)$,表达式如下:

$$CDPL(R)=\log\frac{\Pr[R(d_i^1)\in range(R)\mid d_{-i}]}{\Pr[R(d_i^2)\in range(R)\mid d_{-i}]} \tag{13}$$

其中,$d_{-i}=D\backslash\{d_i\}$.当且仅当 $CDPL(R)\leqslant|\varepsilon|$时,隐私保护机制 R 满足 ε-关联差分隐私,数据中隐私信息的隐私保护强度满足用户对隐私保护的预期,$CDPL(R)$越小,隐私泄露的风险越小.

2.4.2 互信息差分隐私的度量方法

将互信息与差分隐私相结合来度量隐私信息泄露的风险,通常是将差分隐私中的ε作为互信息的约束值,此时的ε一般为用户根据差分隐私计算的较优的隐私保护强度[58−60].如果互信息计算出的隐私泄露风险值小于等于 ε值,则数据中隐私信息的隐私保护强度满足用户对隐私保护的预期.文献[61]中,假设原始数据 $D=\{d_1,d_2,\ldots,d_n\}$经过一个随机隐私保护机制 R 处理后,攻击者可从观察数据 D' 获取到信息,在随机机制 R 满足 ε-差分隐私保护的条件下,用互信息描述攻击者从观察数据 D' 中获取的信息后,原始数据中的隐私信息 S 的不确定性减少的量.当互信息的值小于等于ε时,表明该隐私保护机制的隐私保护满足ε-差分隐私保护,隐私信息泄漏的风险很小.假设具有背景知识 K 的攻击者结合从观察数据 D' 中获取的信息后,可用互信息 $I(S;D'/K)$描述原始数据中的隐私信息 S 的不确定性减少量.

3 分析与评价

3.1 隐私度量指标分析

文献[62,63]提出了不同场景下的隐私度量指标,本文针对上述 4 类隐私度量方法,对各自的隐私度量指标进行分析总结,主要从指标的取值范围,指标与隐私强度、数据可用性和处理开销之间的关系这 4 个方面进行分析,见表 1.

Table 1 Metrics analysis of privacy measurement

表 1 隐私度量指标分析

隐私度量方法	隐私度量指标	指标取值范围	指标与隐私强度	指标与数据可用性	指标与处理开销
基于匿名的隐私度量方法	k-匿名值	[0,∞]	单调递增	单调递减	单调递增
	敏感属性值分布概率	[0,1]	单调递减	单调递增	单调递减
基于信息熵的隐私度量方法	信息熵值	[0,∞]	单调递增	单调递减	–
	条件熵值	[0,∞]	单调递增	单调递减	–
	互信息熵值	[0,∞]	单调递减	单调递增	–
基于集对分析理论的隐私度量方法	势值	[0,∞]	单调递增	单调递减	–
	记分函数	[−1,1]	单调递增	单调递减	–
基于差分隐私的隐私度量方法	差分隐私保护预算	[0,∞]	单调递减	单调递增	单调递减

3.2 隐私度量方法的对比分析

隐私保护技术的选择和攻击者拥有的背景知识的量化分析等都与隐私度量方法的度量效果密切相关,因而隐私度量方法的性能评估应当从多个方面进行综合评估.表 2 根据各隐私度量方法实现的原理、度量效果和主要优缺点对现有研究成果进行了分析评估.

基于匿名的隐私度量方法度量隐私保护强度取决于数据匿名的程度.根据匿名云数据的 k-匿名值和敏感属性值分布的概率,再结合贝叶斯推理推测的信息和原始云数据中隐私信息的差异度分析隐私信息泄露的风险.该度量方法可以结合攻击者以及攻击者拥有的背景知识进行度量,使得度量结果更全面且准确性高.在云计算环境中,对度量小规模的数据容易实现且能获得准确的度量结果.在度量大规模多维数据时,由于构建二叉树图的过程比较复杂且处理效率低,因此,需要采用分布式计算或者多线程技术进行并行度量以提高效率.基于匿名的隐私度量方法应用范围有限,主要用于度量匿名云数据.

Table 2 Comparative analysis of privacy measurement methods

表 2 隐私度量方法的对比分析

方法类型	实现原理	度量效果	主要优点	主要缺点
基于匿名的隐私度量方法	*k*-匿名,敏感属性值分布概率,贝叶斯推理	准确性高,效率低	可结合攻击者以及背景知识进行度量	应用范围受限,构建二叉树图较为复杂
基于信息熵的隐私度量方法	信息熵,条件熵,互信息熵	准确性不高,效率高	通用性高,可结合攻击者以及背景知识进行度量	度量结果易受异常值、错误数据和残缺数据的影响
基于集对分析理论的隐私度量方法	集对分析理论,信息熵集对分析理论	准确性不高,效率低	原理简单,通用性高	建立数据间集对关系较困难,未结合攻击者背景知识度量
基于差分隐私的隐私度量方法	差分隐私,互信息差分隐私	准确性高,效率高	原理简单,数据元组独立时,无需量化攻击者的背景知识	应用范围受限,数据元组相互关联时,难以反映关联隐私的保护强度

基于信息熵的隐私度量方法可以针对攻击者以及攻击者拥有的背景知识等建立相对应的度量模型进行度量,使得度量比较全面,该方法原理简单、易于实现.基于信息熵的隐私度量方法使用大数据计算技术能够实现云计算环境中的大规模多维数据的度量.但文献[64,65]指出,信息熵作为隐私度量指标容易受异常值、错误数据和残缺数据的影响.即使攻击者能够高概率地识别隐私信息,数据中其余低概率的属性值仍然可以导致高的熵值;隐私度量时不同概率分布的构造,会产生同样的熵值,从而度量结果的准确性还受概率分布的影响.因此,基于信息熵的隐私度量方法需要对数据中异常值或错误数据进行处理,才能使度量结果更准确.

基于集对分析理论的隐私度量方法能够实现定性与定量相结合对隐私信息的确定性与不确定性的分析,原理较简单.该方法的关键是分析相邻数据集具有的同一度、差异度和对立度特性,并进行度量刻画.在云计算环境中,度量小规模的数据隐私泄露风险容易实现;度量大规模多维数据时,分析数据集对的同一度、差异度和对立度的效率较低,需要考虑采用分布式计算或者多线程技术进行并行分析以提高效率.该方法未结合攻击者以及攻击者拥有的背景知识进行分析度量,进而影响隐私度量结果的准确性.

基于差分隐私的隐私度量方法度量隐私保护强度主要取决于差分隐私中的ε值,分析ε值就可以反映隐私保护强度,实现比较简单,效率较高.同时,在使用该方法度量相互独立的数据元组时,无需考虑攻击者拥有的背景知识,节省了分析攻击者背景知识的开销.基于差分隐私不仅能度量小规模数据,而且使用大数据计算技术能够实现大规模多维数据的度量.但在数据元组相互关联时,差分隐私仅通过隐私保护预算值并不能反映差分隐私技术的关联隐私信息的保护强度,有必要利用差分隐私的优点,结合互信息等其他隐私度量方法来量化隐私水平.该方法主要适应于面向差分隐私保护技术的隐私保护强度的度量.

3.3 隐私保护技术与隐私度量方法的关系分析

前面分析了 3 类隐私保护技术和 4 种隐私度量方法.结合各类隐私保护的代表性技术、处理数据的本质特征和 4 类隐私度量方法的度量特点,可分析出各类隐私保护技术适用的隐私度量方法,见表 3.

Table 3 The type of measurement method applies to privacy protection techniques

表 3 隐私保护技术适用的隐私度量方法

隐私保护技术	代表性技术	处理数据特征	隐私度量方法
基于数据匿名的隐私保护技术	*k*-匿名,*l*-多样性,*t*-邻近性,…	数据中显示标识符被抑制,数据中准标识符被泛化,数据中敏感属性值被约束	基于匿名的隐私度量方法,基于信息熵的度量方法,基于集对分析理论的度量方法
基于数据失真的隐私保护技术	ε-差分隐私,贝叶斯差分隐私…	数据中添加噪声	基于差分隐私的度量方法,基于信息熵的度量方法
基于密码学的隐私保护技术	收敛加密,基于属性的加密,同态加密	数据被编码成密文形式	–

4 云数据隐私度量的未来研究方向

上文依据 4 种类型总结了隐私度量方法的相关研究工作,但面向云数据的隐私度量的研究当前还处于发展初期,在云计算环境中还面临着诸多挑战.从现有的研究分析来看,未来云数据隐私度量的研究工作主要体现在隐私度量的过程、效果和方法 3 个方面.

4.1 隐私度量过程的研究

(1) 攻击者背景知识的动态、精准量化研究.为准确度量隐私信息泄露的程度,在设计隐私度量方法时需要充分考虑攻击者背景知识对隐私信息泄露的影响.这需要预测攻击者可能拥有的背景知识的具体内容和背景知识的量,已有的研究通常先假设攻击者可能拥有的背景知识,再利用关联规则等方法进行量化.事实上,假设的内容与攻击者的实际情况存在差异.因此,现有方法量化的背景知识与攻击者实际拥有的背景知识还存在差异.尤其是在云计算环境下,随着计算能力的提高,攻击者的背景知识和计算推理能力随时间变化飞速增长,对于背景知识随时间的变化而快速变化的动态、精准量化问题还有待进一步解决.

(2) 隐私保护强度与数据可用性之间的权衡研究.目前隐私保护方法需要牺牲数据可用性来获得较高的隐私保护强度,同时降低了云平台数据共享的服务质量.根据用户的实际情况分析隐私保护强度与云数据可用性的重要性,可以引用博弈论模型等方法在两者之间做出权衡.通过将博弈论与隐私度量相结合设计一种最优权衡的隐私保护度量机制,在给定能容忍的最大数据信息损失和计算开销的条件下,提供最优的隐私保护.如何权衡云数据可用性与有效保障用户隐私之间的关系亦是一个值得深入研究的问题.

4.2 隐私度量效果的研究

(1) 隐私度量粒度可控性研究.虽然现有的隐私度量方法能够实现依据不同隐私保护技术提供隐私度量,但在云服务中,同一云数据可能共享给多个数据使用者.因此,用户依据数据使用者的级别、数据的用途、不同时间段和安全级别等对云数据隐私度量的粒度需求均不相同.根据上述情况对云数据隐私度量进行细粒化处理,构造出结合这些因素变化的细粒度隐私度量方案是一个具有挑战性的问题.

(2) 隐私度量的效率优化研究.随着云计算、大数据技术在人们生活中的不断渗透,越来越多的数据外包到云服务提供商进行计算、共享、存储等.针对大规模多维的云数据产生速度快、流通量大等问题,如何提升云数据度量方法的效率、减少隐私度量的处理开销是云数据隐私度量方法可行性的关键.

4.3 隐私度量方法的研究

(1) 多种隐私度量方法的组合研究.面向云数据的隐私度量,由于用户拥有大量不同类型的数据,可能需要先进行聚类处理[66],再进行隐私度量;或这些数据可能经过不同的隐私保护方法进行处理,针对不同隐私保护方法的度量指标也不同,从而度量结果也不同,如匿名程度、信息熵值和差分隐私的隐私保护预算等,这可能需要融合多种隐私度量技术,如标准熵、标准互信息等.云计算环境中,针对已有的单一隐私度量方法存在的优缺点,有必要进一步对已有的隐私度量方法进行扩展和抽象,或者将多种隐私度量方法进行组合研究,进而设计出度量应用范围更广的隐私度量方法.

(2) 隐私度量的标准化研究.不同的云数据隐私度量方法能够度量不同隐私保护技术的隐私保护强度.现有的隐私度量方法主要建立在成熟隐私保护技术的基础上进行有针对性的度量.针对特定的隐私保护技术需要构建特定的隐私度量模型,不具备通用性且目前缺少一致的隐私度量标准,而隐私度量标准化可以减少可观察的特征,更容易定义和建模隐私度量.因此,设计一种标准化的云数据隐私度量模型,实现对不同类型的云数据隐私保护技术进行全面的、通用的隐私度量.

(3) 云数据全生命周期、动态演化的隐私度量研究.现有云服务已经渗透到人们日常生活的各个方面,然而,一旦采用云服务,数据的所有权与管理权分离,用户将会失去对数据的安全控制权.因此,从数据发布、存储、共享、使用到销毁的全生命周期各阶段均存在隐私信息泄露的风险[1].迫切需要引入隐私计算理论[67],结合云数据全生命周期和数据动态演化过程实施隐私度量.实现数据发布、存储、共享、使用、销毁的全生命周期各阶段

动态演化的隐私度量将对云数据隐私保护具有重要意义.

5 结束语

云服务为人们的生活提供了诸多便利,随之而来的云数据隐私保护问题不容小觑.目前,存在许多面向云数据的隐私保护技术,但如何有效评价面向云数据的隐私保护技术的可靠性依然是隐私保护领域面临的挑战性问题.其中,隐私保护强度是评估隐私保护技术的核心指标,研究隐私度量对隐私保护具有重大意义.本文主要对现有隐私度量方法的研究成果进行综述:首先,对隐私保护技术和隐私度量进行了概述,并给出了面向云数据的隐私保护技术性能评价指标和一种综合评价框架;然后,提出了一种面向云数据的隐私度量抽象模型,结合相关研究工作,对基于匿名的、基于信息熵的、基于集对分析理论和基于差分隐私这 4 类隐私度量方法的基本思想、实现原理等进行详细的分析、归纳和总结;最后分别总结了 4 类隐私度量方法主要的优缺点:基于匿名的隐私度量方法原理简单,可以结合攻击者及其背景知识进行度量,但该方法主要适用于匿名数据的度量;基于信息熵的隐私度量方法应用范围广,但度量结果易受异常值和错误数据等的影响;基于集对分析理论的隐私度量方法原理简单、应用范围广,但构建数据集之间的集对关系较复杂并未能结合攻击者及背景知识进行度量;基于差分隐私的隐私度量方法在数据元组独立时,无需考虑攻击者的背景知识,但其主要适用于加噪后的数据.最后,指出了面向云数据隐私度量的未来发展趋势.随着互联网以及新技术的发展,人们对信息隐私保护问题越来越关注,无论是理论研究还是在实际应用领域,隐私度量对隐私保护的研究都具有重大意义.从现有研究分析可知,当前面向云数据的隐私度量的研究还处于发展初期,隐私度量的研究与实际应用还需进一步研究与探索.

References:

[1] Xiong JB, Li FH, Liu XM, Yao ZQ, Chen P. A full lifecycle privacy protection scheme for sensitive data in cloud computing. Peer-to-Peer Networking and Applications, 2015,8(6):1025–1037. [doi: 10.1007/s12083-014-0295-x]

[2] Zhou SG, Li F, Tao YF, Xiao XK. Privacy preservation in database applications: A survey. Chinese Journal of Computes, 2009,32(5):847–858 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2009.00847]

[3] Machanavajjhala A, Gehrke J. On the efficiency of checking perfect privacy. In: Proc. of the ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database. Chicago: ACM Press, 2006. 163–172. [doi: 10.1145/1142351.1142375]

[4] Liu YH, Zhang TY, Jin XL, Cheng XQ. Personal privacy protection in the era of big data. Journal of Computer Research and Development, 2015,52(1):229–247 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2015.20131135]

[5] Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. IEEE Access on Theoretical Foundations for Big Data Applications, 2016,4:1821–1834. [doi: 10.1109/ACCESS.2016.2558446]

[6] Zhang X, Liu C, Nepal S, Yang C, Gou WC. A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud. Journal of Computer & System Science, 2014,80(5):1008–1020. [doi: 10.1016/j.jcss.2014.02.007]

[7] Mohammadian E, Noferesti M, Jalili R. FAST: Fast anonymization of big data streams. In: Proc. of the ACM Conf. on Big Data Science and Computing. Beijing: ACM Press, 2014. [doi: 10.1145/2640087.2644187]

[8] Yu JD, Dong X, Lou Y, Li ML. Differentially private wireless data publication in large-scale WLAN networks. In: Proc. of the IEEE Conf. on Parallel and Distributed Systems. Melbourne: IEEE Press, 2015. 290–297. [doi: 10.1109/ICPADS.2015.44]

[9] Li SD, Dou JW, Wang DS. Survey on homomorphic encryption and its applications to cloud security. Journal of Computer Research and Development, 2015,52(6):1378–1388 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2015.20131494]

[10] Fun B, Wang K, Chen R, Yu P. Privacy-Preserving data publishing: A survey of recent development. ACM Computing Surveys, 2010,42(4):1–53. [doi: 10.1145/1749603.1749605]

[11] Bayardo RJ, Agrawal R. Data privacy through optimal *k*-anonymization. In: Proc. of the Int'l Conf. on Data Engineering. Washington: ACM Press, 2005. 217–228. [doi: 10.1109/ICDE.2005.42]

[12] Lu QW, Wang CM, Xiong Y. Personalized privacy-preserving trajectory data publishing. Chinese Journal of Electronics, 2017, 26(2):285–291 (in Chinese with English abstract). [doi: 10.1049/cje.2017.01.024]

[13] Xiao X, Tao Y. Personalized privacy preservation. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Chicago: ACM Press, 2006. 229–240. [doi: 10.1145/1142473.1142500]

[14] Jiang HW, Zeng GS, Ma HY. Greedy clustering-anonymity method for privacy preservation of table data-publishing. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):341–351 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5015.htm [doi: 10.13328/j.cnki.jos.005015]

[15] Fun B, Wang K, Yu P. Top-Down specialization for information and privacy preservation. In: Proc. of the Int'l Conf. on Data Engineering. Tokoyo: ACM Press, 2005. 205–216. [doi: 10.1109/ICDE.2005.143]

[16] Gong QY, Yang M, Lou JZ. Data anonymization approach for microdata with relational and transaction attributes. Ruan Jian Xue Bao/Journal of Software, 2016,27(11):2828–2842 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5099.htm [doi: 10.13328/j.cnki.jos.005099]

[17] Xiong JB, Yao ZQ, Ma JF, Liu XM, Li Q, Ma J. PRIAM: Privacy preserving identity and access management scheme in cloud. KSII Trans. on Internet and Information Systems, 2014,8(1):282–304. [doi: 10.11959/j.issn.1000-436x.2016176]

[18] Chen BC, Ramakrishnan R, Lefevre K. Privacy skyline: Privacy with multidimensional adversarial knowledge. In: Proc. of the Int'l Conf. on Very large Data Bases. Vienna: ACM Press, 2007. 770–781.

[19] Li TC, Li NH. Injector: Mining background knowledge for data anonymization. In: Proc. of the Int'l Conf. on Data Engineering. NewYork: ACM Press, 2008. 446–455.

[20] Cai ZP, He Z, Guan X, Li YS. Collective data-sanitization for preventing sensitive information inference attacks in social networks. IEEE Trans. on Dependable and Secure Computing, 2016,(99):1–14. [doi: 10.1109/TDSC.2016.2613521]

[21] Du W, Teng Z, Zhu Z. Privacy-MaxEnt: Integrating background knowledge in privacy quantification. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Vancouver, 2008. 459–472.

[22] Wang CM, Gou YJ, Gou YH. Privacy metric for user's trajectory in location-based services. Ruan Jian Xue Bao/Journal of Software, 2012,23(2):352–360 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3946.htm [doi: 10.3724/SP.J.1001.2012.03946]

[23] Li T, Li NH, Zhang J. Modeling and integrating background knowledge in data anonymization. In: Proc. of the IEEE Int'l Conf. on Data Engineering. Shanghai, 2009. 6–17. [doi: 10.1109/ICDE.2009.86]

[24] Mao YX, Chen TB, Shi BL. Efficient method for mining multiple-level and generalized association rules. Ruan Jian Xue Bao/Journal of Software, 2011,22(12):2965–2980 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3907.htm [doi: 10.3724/SP.J.1001.2011.03907]

[25] Sweeney L. *k*-Anonymity: A model for protecting privacy. Int'l Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002,10(5):557–570.

[26] Machanavajjhala A, Gehrke J, Kifer D. *l*-Diversity: Privacy beyond *k*-anonymity. In: Proc. of the IEEE Int'l Conf. on Data Engineering. Atlanta: IEEE Press, 2006. 24–35.

[27] Li NH, Li TC, Venkata S. *t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity. In: Proc. of the IEEE Int'l Conf. on Data Engineering. Istanbul: IEEE Press, 2007. 106–115. [doi: 10.1103/ICDE.2007.367856]

[28] Li NH, Li TC, Nkatasubramanian S. (*n*,*t*)-Closeness: A new privacy measure for date publishing. IEEE Trans. on Knowledge and Date Engineering, 2010,22(7):943–956. [doi: 10.1109/TKDE.2009.139]

[29] Zhang JP, Xie J, Yang J, Zhang B. A *t*-closeness privacy model based on sensitive attribute values semantics bucketization. Journal of Computer Research and Development, 2014,51(1):126–137 (in Chinese with English abstract). [doi:10.7544/issn1000-1239.2014.20130688]

[30] Gkountouna O, Terrovitis M. Anonymizing collections of tree-structured data. IEEE Trans. on Knowledge and Data Engineering, 2015,27(8):2034–2048.

[31] Yuji Y, Kouichi I. *k*-Presence-Secrecy: Practical privacy model as extension of *k*-anonymity. IEICE Trans. on Information & System, 2017,(4):730–740. [doi: 10.1587/transinf.2016DA0015]

[32] Li XY, Zhang CH, Jung T, Qian JW, Chen LL. Graph-Based privacy-preserving data publication. In: Proc. of the IEEE Int'l Conf. on Computer Communications. San Francisco: IEEE Press, 2016. 1–9. [doi: 10.1109/INFOCOM.2016.7524584]

[33] Shannon C. A mathematical theory of communication. The Bell System Technical Journal, 1948,27(3):379–423.

[34] Clauß S, Stefan S. Structuring anonymity metrics. In: Proc. of the ACM Conf. on Computer and Communications Security. Alexandria: ACM Press, 2006. 55–62.

[35] Peng CG, Ding HF, Zhu YJ, Tian YL, Fu ZF. Information entropy models and privacy metrics methods for privacy protection. Ruan Jian Xue Bao/Journal of Software, 2016,27(8):1891–1903 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5096.htm [doi: 10.13328/j.cnki.jos.005096]

[36] Zhang HL, Shi YL, Zhang SD, Zhou ZM, Cui LZ. A privacy protection mechanism for dynamic data based on partition-confusion. Journal of Computer Research and Development, 2016,53(11):2454–2464 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20150553]

[37] Diaz C, Troncoso C, Danezis G. Does additional information always reduce anonymity. In: Proc. of the ACM Workshop on Privacy in the Electronic Society. Alexandria: ACM Press, 2007. 72–75.

[38] Lai LF, Ho SW, Poor HV. Privacy-Security trade-offs in biometric security systems. Part II: Multiple use case. IEEE Trans. on Information Forensics & Security, 2011,6(1):140–151. [doi: 10.1109/TIFS.2010.2098872]

[39] Asoodeh S, Alajaji F, Linder T. Notes on information-theoretic privacy. In: Proc. of the IEEE Conf. on Communication, Control and Computing. Monticello: IEEE Press, 2015. 1272–1278.

[40] Calmon F, Makhdoumi A, Médard M. Fundamental limits of perfect privacy. In: Proc. of the IEEE Int'l Symp. on Information Theory. HongKong: IEEE Press, 2015. 1796–1800.

[41] Alvim M, Andrés M, Chatzikokolakis K, Pierpaolo D, Palamidessi C. On the information leakage of differentially-private mechanisms. Journal of Computer Security, 2015,23(4):427–469.

[42] Calmon F, Fawaz N. Privacy against statistical inference. In: Proc. of the IEEE Conf. on Communication, Control and Computing. Monticello: IEEE Press, 2012. 1401–1408.

[43] Humbert M, Ayday E, Hubaux JP, Telenti A. Addressing the concerns of the lacks family: Quantification of kin genomic privacy. In: Proc. of the ACM Conf. on Computer and Communications Security. Berlin: ACM Press, 2013. 1141–1152. [doi: 10.1145/2508859.2516707]

[44] Humbert M, Ermanayda Y, Hubaux JP, Telenti A. Quantifying interdependent risks in genomic privacy. ACM Trans. on Privacy & Security, 2017,20(1):1–30. [doi: 10.1145/3035538]

[45] Zhao KQ. Disposal and description of uncertainties based on the set pair analysis. Information and Control, 1995,24(3):162–166 (in Chinese with English abstract). [doi: 10.13976/j.cnki.xk.1995.03.006]

[46] Yan Y, Hao XH, Wang WJ. A set pair analysis method for privacy metric. Engineering Journal of Wuhan University, 2015,48(6):883–890 (in Chinese with English abstract). [doi: 10.14188/j.1671-8844.2015-06-027]

[47] Dwork C. Differential privacy. In: Proc. of the Int'l Colloquium on Automata, Languages and Programming. Berlin: Springer-Verlag, 2006. 1–12. [doi: 10.1007/11787006_1]

[48] Dwork C, Lei J. Differential privacy and robust statistics. In: Proc. of the ACM Symp. on Theory of Computing. Bethesda: ACM Press, 2009. 371–380.

[49] Dwork C, Mcsherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. Theory of Cryptography, 2006,7(8):265–284. [doi: 10.1007/11681878_14]

[50] Chen R, Acs G, Castelluccia C. Differential private sequential data publication via variable-length *N*-grams. In: Proc. of the ACM Conf. on Computer and Communication Security. Raleigh: ACM Press, 2012. 638–649.

[51] Zhang WJ, Li H. A differentially-private mechanism for multi-level data publishing. Chinese Journal of Network and Information Security, 2015,1(1):58–65 (in Chinese with English abstract). [doi: 10.11959/j.issn.2096-109x.2015.00008]

[52] Jorgensen Z, Yu T, Cormode G. Conservative or liberal? Personalized differential privacy. In: Proc. of the IEEE Int'l Conf. on Data Engineering. Seoul: IEEE Press, 2015. 1023–1034. [doi: 10.1109/ICDE.2015.7113353]

[53] Chen R, Fung BCM, Yu P, Desai B. Correlated network data publication via differential privacy. The Int'l Journal on Very Large Data Bases, 2014,23(4):653–676.

[54] Kifer D, Machanavajjhala A. Pufferfish: A framework for mathematical privacy definitions. ACM Trans. on Database Systems, 2014,39(1):671–683.

[55] Yang B, Sato I, Nakagawa H. Bayesian differential privacy on correlated data. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Melbourne: ACM Press, 2015. 747–762. [doi: 10.1145/2723372.2747643]

[56] Zhu TQ, Xiong P, Li G, Zhou W. Correlated differential privacy: Hiding information in non-IID data set. IEEE Trans. on Information Forensics and Security, 2015,10(2):229–242. [doi: 10.1109/TIFS.2014.2368363]

[57] Wu XT, Dou WC, Ni Q. Game theory based privacy preserving analysis in correlated data publication. In: Proc. of the Australasian Computer Science Week Multi-Conf. Geelong: ACM Press, 2017. 73–82. [doi: 10.1145/3014812.3014887]

[58] Barthe G, Kopf B. Information-Theoretic bounds for differentially private mechanisms. In: Proc. of the Computer Security Foundations Symp. Washington: IEEE Press, 2011. 191–204.

[59] Alvim M, Andres M, Chatzikokolakis K, Degano P, Palamidessi C. Differential privacy: On the trade-off between utility and information leakage. In: Proc. of the Int'l Conf. on Formal Aspects of Security and Trust. Leuven: ACM Press, 2012. 39–54.

[60] Wang W, Ying L, Zhang J. On the relation between identifiability, differential privacy, and mutual-information privacy. IEEE Trans. on Information Theory, 2016,62(9):5018–5029. [doi: 10.1109/TIT.2016.2584610]

[61] Cuff P, Yu LQ. Differential privacy as a mutual information constraint. In: Proc. of the ACM SIGSAC Conf. on Computer and Communications Security. Vienna: ACM Press, 2016. 43–54. [doi: 10.1145/2976749.2978308]

[62] Wagner I, Eckhoff D. Technical privacy metrics: A systematic survey. arXiv Preprint arXiv:1512.00327, 2015.

[63] Wan S, Li FH, Niu B,Sun Z, Li H. Research progress on location privacy-preserving techniques. Journal on Communications, 2016,37(12):124–141 (in Chinese with English abstract). [doi: 10.11959/j.issn.1000-436x.2016279]

[64] Toth G, Hornak Z, Vajda F. Measuring anonymity revisited. In: Proc. of the NORDIC Workshop on Secure IT Systems. Helsinki: ACM Press, 2004. 85–90.

[65] Murdoch S. Quantifying and measuring anonymity. In: Data Privacy Management and Autonomous Spontaneous. Berlin: Springer-Verlag, 2014. 3–13. [doi: 10.1007/978-3-642-54568-9_1]

[66] Wu DP, Yang BR, Wang HG, Wang CB, Wang RY. Privacy-Preserving multimedia big data aggregation in large-scale wireless sensor networks. ACM Trans. on Multimedia Computing, Communications and Applications, 2016,12(4). [doi: 10.1145/2978570]

[67] Li FH, Li H, Jia Y, Yu NH, Weng J. Privacy computing: Concept, connotation and its research trend. Journal on Communications, 2016,37(4):1–11 (in Chinese with English abstract).

附中文参考文献:

[2] 周水庚,李丰,陶宇飞,肖小奎.面向数据库应用的隐私保护研究综述.计算机学报,2009,32(5):847–858. [doi: 10.3724/SP.J.1016.2009.00847]

[4] 刘雅辉,张铁赢,靳小龙,程学旗.大数据时代的个人隐私保护.计算机研究与发展,2015,52(1):229–247. [doi: 10.7544/issn1000-1239.2015.20131135]

[9] 李顺东,窦家维,王道顺.同态加密算法及其在云安全中的应用.计算机研究与发展,2015,52(6):1378–1388. [doi: 10.7544/issn1000-1239.2015.20131494]

[14] 姜火文,曾国荪,马海英.面向表数据发布隐私保护的贪心聚类匿名方法.软件学报,2017,28(2):341–351. http://www.jos.org.cn/1000-9825/5015.htm [doi: 10.13328/j.cnki.jos.005015]

[16] 龚奇源,杨明,罗军舟.面向关系-事务数据的数据匿名方法.软件学报,2016,27(11):2828–2842. http://www.jos.org.cn/1000-9825/5099.htm [doi:10.13328/j.cnki.jos.005099]

[22] 王彩梅,郭亚军,郭艳华.位置服务中用户轨迹的隐私度量.软件学报,2012,23(2):352–360. http://www.jos.org.cn/1000-9825/3946.htm [doi:10.3724/SP.J.1001.2012.03946]

[24] 毛宇星,陈彤兵,施伯乐.一种高效的多层和概化关联规则挖掘方法.软件学报,2011,22(12):2965–2980. http://www.jos.org.cn/1000-9825/3907.htm [doi:10.3724/SP.J.1001.2011.03907]

[29] 张健沛,谢静,杨静,张冰.基于敏感属性值语义桶分组的 *t*-closeness 隐私模型.计算机研究与发展,2014,51(1):126–137. [doi: 10.7544/issn1000-1239.2014.20130688]

[35] 彭长根,丁红发,朱义杰,田有亮,符祖峰.隐私保护的信息熵模型及其度量方法.软件学报,2016,27(8):1891–1903. http://www.jos.org.cn/1000-9825/5096.htm [doi: 10.13328/j.cnki.jos.005096]

[36] 张宏磊,史玉良,张世栋,周中民,崔立真.一种基于分块混淆的动态数据隐私保护机制.计算机研究与发展,2016,53(11):2454–2464. [doi: 10.7544/issn1000-1239.2016.20150553]

[45] 赵克勤.集对分析对不确定性的描述和处理.信息与控制,1995,24(3):162–166. [doi: 10.13976/j.cnki.xk.1995.03.006]

[46] 晏燕,郝晓弘,王万军.一种隐私保护度量的集对分析方法.武汉大学学报,2015,48(6):883–890. [doi: 10.14188/j.1671-8844.2015-06-027]

[51] 张文静,李晖.差分隐私保护下的数据分级发布机制.网络与信息安全学报,2015,1(1):58–65. [doi: 10.11959/j.issn.2096-109x.2015.00008]

[63] 万盛,李凤华,牛犇,孙哲,李晖.位置隐私保护技术研究进展.通信学报,2016,37(12):124–141. [doi: 10.11959/j.issn.1000-436x.2016279]

[67] 李凤华,李晖,贾焰,俞能海,翁健.隐私计算研究范畴及发展趋势.通信学报,2016,37(4):1–11. [doi: 10.11959/j.issn.1000-436x.2016078]

熊金波(1981－),男,湖南益阳人,博士,副教授,CCF 专业会员,主要研究领域为云数据安全,隐私保护技术.

马蓉(1992－),女,硕士生,CCF 学生会员,主要研究领域为云数据安全,隐私保护技术.

王敏燊(1994－),男,硕士生,CCF 学生会员,主要研究领域为信任评估,数据安全.

姚志强(1967－),男,博士,教授,CCF 高级会员,主要研究领域为信息安全.

田有亮(1982－),男,博士,教授,博士生导师,主要研究领域为算法博弈论,数据安全,隐私保护.

林铭炜(1985－),男,博士,副教授,CCF 专业会员,主要研究领域为存储系统,嵌入式系统.