广西科学院学报
Journal of Guangxi Academy of Sciences

2009,25(4):273~277
Vol. 25,No. 4 November 2009

# 基于匿名方法的数据发布隐私泄露控制技术研究进展
# Technique Advances for Anonymization-based Controlling Privacy Disclosure in Data Publishing

滕金芳,钟 诚
TENG Jin-fang,ZHONG Cheng

(广西大学计算机与电子信息学院,广西南宁 530004)
(School of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, 530004, China)

**摘要**:介绍主要的数据匿名保护模型,总结分析基于泛化和隐匿、基于聚类、基于交换和分解的数据匿名化算法的研究成果和存在的问题,指出数据发布的匿名化技术还需要在同质性攻击和背景知识攻击、动态性数据的隐私保护、个性化的隐私保护、数据发布自适应机制、面向应用的隐私保护和多属性数据集的隐私保护等几个方面做进一步的研究。

**关键词**:数据发布 匿名化 隐私 泄露 保护 $k$-匿名
**中图法分类号**:TP309.2 **文献标识码**:A **文章编号**:1002-7378(2009)04-0273-05

**Abstract**: This paper introduces the main anonymity preservation models, reviews the research advances and analyzes the limitations of the anonymization algorithms that generalization and suppression-based, clustering-based and swapping-based. This paper also indicates that the anonymization technique of data publishing need to be further researched, such as the homogeneity attack and the background knowledge attack, the privacy preservation of dynamic data, the personalized privacy preservation, adaptive mechanism of data publication, application-oriented privacy protection, and multiple sensitive attributes dataset privacy protection.

**Key words**: data publishing, anonymization, privacy, disclosure, preservation, $k$-anonymity

随着互联网技术的迅速发展,人们在共享资源的同时,也出现了大量隐私泄露问题。这使得隐私保护技术的研究受到人们的高度关注,其中数据发布环境中的隐私泄露问题正成为隐私保护领域的一个热门研究课题。

数据发布中的数据集为一个二维表,每个数据记录均与现实中的某一个体对应,包含多个属性值。这些属性可以分为标识符、准标识符和敏感属性3类。在发布数据时,为了保护个体的隐私,往往会直接删除原始数据集中唯一标识个体身份的标识符(如姓名、身份证号码等)。但是,这样做并不能完全有效地保护隐私,攻击者可能通过数据集中的准标识符(如属性组[出生日期,性别,居住地邮编])与外部公开的数据源进行链接,推导出个体的身份和敏感属性值,从而造成隐私泄露。这种情况被称为链接攻击[1]。文献[1]研究表明,若对选民登记表和医疗信息表采用这种链接攻击,则超过87%的美国公民的身份被确定。针对数据发布中存在的链接攻击问题,一个重要的方法是采用匿名化技术防止数据窥探者的推理行为。即在数据发布前对数据进行预处理,将原始数据集划分为若干个等价组,使得组内记录的准标识符属性和敏感属性不再具有一对一的关系,由此切断两者之间的推理联系,以保护个体的隐私信息。

数据匿名化的主要目标是在保证数据可用性的同时,通过适当损失一些属性值所包含的信息来提高数据的安全性。因此,匿名化原始数据集,必然会造成信息的损失。数据的可用性和数据的安全性是

**收稿日期**:2009-10-10
**作者简介**:滕金芳(1970-),女,硕士研究生,讲师,主要从事网络信息安全研究。

相互矛盾的，两者之间需要找到折中平衡。目前，数据匿名化的研究工作主要是设计更有效的匿名保护模型，以及针对特定匿名保护模型设计出性能更好的匿名化算法。

## 1 数据匿名保护模型研究现状

目前主要的两种匿名保护模型是 $k$-匿名模型（$k$-anonymity）和 $l$-多样模型（$l$-diversity）。其中 $l$-多样模型是针对 $k$-匿名模型进行改进的，它可以提供更强有力的匿名保护。

### 1.1 $k$-匿名模型

Samarati 和 Sweeney 在文献[2]中提出了 $k$-匿名技术，它要求数据集中每个记录的准标识符属性值都至少跟其他 $k-1$ 个记录的准标识符属性值相同，使得个体无法与记录一一对应，从而达到保护个体隐私的目的。Sweeney 在文献[1]中提出了 $k$-匿名模型，以防止数据发布中的链接攻击。同时，Sweeney 在文献[3]中论述了实现 $k$-匿名保护的泛化和隐匿方法。

文献[4]指出：在某些情况下，$k$-匿名并不能保证隐私信息的安全，$k$-匿名模型存在同质性攻击和背景知识攻击的问题。背景知识攻击是指攻击者利用预先知道的某些额外（背景）信息来进行攻击。$k$-匿名模型虽然防止了个体身份的泄露，但是并不能完全防止个体敏感属性的泄露。例如，假设等价组的记录具有相同的敏感属性值，攻击者虽然无法从等价组中推断出与个体对应的具体记录，但是很容易知道个体的敏感属性值，从而出现了所谓的同质性攻击问题。

### 1.2 $l$-多样性模型

针对 $k$-匿名模型存在的缺陷，文献[4]提出了 $l$-多样性模型，该模型除了要求满足 $k$-匿名之外，还要求每个等价组具有 $l(l \geqslant 2)$ 个不同的敏感属性值。虽然 $l$-多样性模型通过提高等价组敏感属性值多样性的方法来解决了同质性攻击问题，但是在某些特殊情况下，$l$-多样性模型依然存在缺陷。例如，由于背景知识攻击的复杂性，$l$-多样性模型并没有很好的办法来解决背景知识攻击问题；而且，$l$-多样性模型仅仅考虑到有主关键字的数据集的情形，在无主关键字的情形下，攻击者仍有可能推断出个体的隐私信息。

### 1.3 其他模型

在 $k$-匿名模型和 $l$-多样性模型的基础上，人们先后又提出了一些匿名模型。其中，文献[5]提出了 $(a,k)$-匿名模型，它要求等价组中任意一个敏感属性值的出现频率小于参数 $a$，以避免某些敏感值出现频率过高的情形。文献[6]给出一种 $p$ 敏感 $k$-匿名模型，该模型要求等价组内任何敏感属性至少有 $p$ 个不同的值。针对数值型敏感属性，文献[7]提出了 $(k,e)$-匿名模型，此模型要求等价组内敏感属性值的范围至少为 $e$。文献[8]提出的 $t$-封闭性模型则要求敏感属性值在等价组上的分布与在整体上的分布尽可能接近，两者间的差异不大于参数 $t$。

## 2 数据匿名化算法研究现状

文献[9]和文献[10]分别证明了采用 $k$-匿名技术由原始数据集导出最优的匿名数据集是一个 NP-hard 问题。因此，数据匿名化算法研究的重点是，寻找近似最优算法以达到有效地匿名化原始数据。

### 2.1 基于泛化和隐匿的匿名化算法

所谓泛化是指在数据集中用抽象的属性值来代替原来具体的属性值，使其意义变得广泛[2,3]。所谓隐匿是指直接删除数据集中某些属性值或记录[2,3]。

数据泛化一般采用基于泛化层次结构的策略，数据泛化层次包括域泛化层次和值泛化层次。数据泛化必须按照预先定义的泛化层次结构由下而上逐步使用抽象属性值代替原有属性值。通过泛化和隐匿技术来实现数据匿名化一般需要经过多次的数据泛化，才能满足匿名要求。数据泛化的次数越多，信息的损失量越大，数据的可用性也会降低。

目前主要是采用全域泛化来实现数据集的 $k$-匿名化，全域泛化是指对整个属性域进行泛化。经典的全域泛化算法主要有：(1)Datafly 算法[11]。该算法采用贪心方法，以属性为单位对所有记录进行匿名处理。但是该算法只能对记录级数据进行隐匿操作，而且满足匿名要求的记录继续参与泛化，因而信息损失量大。(2)MinGen 最小泛化算法[3]。该算法给出了最小泛化和最小失真的定义，其基本思想是从所有满足 $k$-匿名要求的泛化表中求出具有最高精度的 $k$-匿名表。该算法无法应用于规模较大的数据集。(3)Incognito 算法[12]。这是由 Lefevre 等人提出的一种高效全域泛化方法，该算法提高了单一约束全域泛化的效率，但是没有能够很好解决信息损失问题。

由于全域泛化算法受到泛化层次结构的限制，泛化过程中常常会产生不必要的信息损失。所以针对这个缺陷，有研究者提出了改进的泛化算法，这些泛化算法主要有：(1)Iyengar 提出的 GA（Genetic

Algorithm)算法[13]。它采用不完全随机搜索策略,提出一种遗传算法以解决 $k$- 匿名中的数据挖掘分类问题。(2) 由底向上的泛化方法[14]。它采用启发式搜索策略,利用信息熵来计算信息损失,递归地修剪分类树直到满足 $k$- 匿名需求为止。(3)自顶向下的泛化方法[15]。它采用面向分类的匿名化策略,提高了分类结果的准确率,但是该方法没有考虑到匿名化造成的信息损失。(4)基于多维空间划分的 $k$-匿名算法[16]。该算法只能对连续型数据进行划分,对分类型数据难以实现,实际应用效果较差。

虽然基于泛化和隐匿的匿名化算法实现了数据集的匿名化,但是此类算法仍然存在如下不足:(1)为了提高隐私信息的安全性,将等价组中不同的准标识符属性值泛化为相同值,造成了信息的损失。(2)全域泛化对数据集的整个属性域进行泛化,使得同一属性域中的所有数据都位于相同的泛化层次,导致一些不必要的信息损失。(3)全域泛化需要预定义泛化层次结构,而泛化层次的构造没有一个统一的标准。(4)如何处理好不同类型数据的泛化还没有很好的解决办法。

### 2.2 基于聚类的匿名化算法

聚类是目前应用较广的数据分析方法,将聚类方法引入到数据集 $k$- 匿名中的主要思想是:用聚类方法将原始数据集划分为若干类,每个类至少包含 $k$ 个记录,使得类内的记录尽可能地相似。然后对同一类内所有记录进行泛化处理,使其在准标识符上具有相同的属性值,生成等价组,从而实现数据集的匿名化。

基于 $k$- 匿名模型,文献[17]提出一种基于聚类的方法以解决数据匿名化问题,但是,该方法仍然存在隐私泄露。文献[18]在实现 $k$- 匿名化的过程中引入权重,但是如何设置适合的权重值仍然困难。为了解决数据匿名化问题,文献[19]提出了 $r$-gather 和 $r$-cellular 两种聚类算法。求解基于聚类的匿名化最优算法也是一个 NP-hard 问题。但是由于基于泛化和隐匿的匿名化算法受到泛化层次结构的限制,因此,通过基于聚类的匿名化算法获得的匿名数据集具有更高的精度和可用性。

为了达到既降低信息损失量又能保证数据的正确性和可用性的目的,如何合理定义不同类型数据的距离、类质心以及信息损失的度量方法;如何设置不同属性的权重值;如何实现类内的记录最大程度地相似;如何选择更好的泛化方法来生成等价组。是进一步深入研究基于聚类的匿名化算法时需要解决的问题。

### 2.3 基于交换和分解的匿名化算法

数据交换是将某些数据项在不影响数据一致性和整体统计特性的同时互换数据以达到保护隐私信息的目的[20]。

文献[7]提出了 $(k,e)$- 匿名方法,它首先将数据集划分为若干组,等价组内敏感属性值的范围至少为 $e$,然后将组内记录之间的准标识符属性与敏感属性进行交换,记录准标识符属性值不变,但是每个记录的准标识符属性与敏感属性之间构成多对多的关系,提高了隐私数据的不确定性。

文献[21]提出了一种新的不采用泛化和隐匿技术的匿名方法 Anatomy,其主要思想是:首先将数据集划分成满足 $l$- 多样性匿名模型的等价组,然后将结果分成两张数据表发布,其中一张表主要包含准标识符属性,另一张表主要包含敏感属性,两张数据表之间通过等价组 ID 关联,利用两张表之间的有损连接来防止隐私的泄露。文献[22]基于有损连接对隐私数据进行保护的思想,提出了一种针对多敏感属性隐私数据发布的多维桶分组技术。此外,文献[23]也提出了一种基于有损连接的隐私数据发布算法。

基于交换和分解的匿名化算法的特点是通过进行交换或分组,使得等价组内记录的准标识符属性和敏感属性形成多对多的关系,保证了隐私数据的安全性。而且,该类算法不需要对准标识符属性进行泛化,较好的保持了数据完整性,降低了信息的损失,能够较好的满足面向聚集查询的应用要求。

## 3 基于匿名方法的数据发布隐私保护研究展望

在信息化社会中,数据的高度共享虽然提供了信息交流的便利,但是它同时增加了隐私信息泄露的风险。数据发布环境中存在的隐私泄露问题严重影响数据拥有者的权益,因此匿名化技术成为数据发布环境中隐私保护的主要研究内容之一。尽管目前匿名化技术研究取得了不少重要的成果,但是仍然存在许多理论和技术问题需要进一步研究解决。

### 3.1 同质性攻击和背景知识攻击的研究

文献[4]指出 $k$- 匿名模型存在同质性攻击和背景知识攻击的问题。针对同质性攻击和背景知识攻击两种攻击,$l$- 多样性模型要求发布数据的每个等价组中的敏感属性值具有多样化以防止隐私信息的泄露。然而,在无主关键字的情形下,攻击者仍可能

推断出个体的隐私信息。并且由于背景知识攻击的复杂性，缺乏用于评估攻击者背景知识的有效方法，所以目前对背景知识攻击的研究还不成熟。因此，需要对同质性攻击和背景知识攻击做进一步的研究。

### 3.2 动态数据的隐私保护研究

现有的大多数匿名化模型和算法都假设数据是静态的，而现实中的数据往往是动态变化的，这些模型和算法并不能完全保护动态数据的隐私安全。如果采用已有的匿名方法对动态更新后的数据进行再次发布，那么将会造成隐私数据质量下降，并且很有可能会形成多个发布版本之间的推理通道，从而导致隐私信息的泄露[24]。因此对动态数据的隐私保护需要进一步研究。

### 3.3 个性化的隐私保护研究

文献[25]提出了个性匿名的概念和个性泛化方法，所谓个性匿名是指对数据表中的记录提供不同粒度的隐私保护程度，以避免出现保护不足和过分保护的情况。目前大部分数据发布方法对数据集中的记录提供相同的隐私保护粒度，而不考虑不同个体的隐私需求，缺乏隐私保护个性化。因此个性化的隐私保护将会成为今后的一个研究热点。

### 3.4 数据发布自适应机制的研究.

大部分隐私数据发布方法没有考虑攻击者的背景知识，不同背景知识的攻击者的攻击方法不太一样且对数据发布造成的隐私侵犯威胁程度也不一样。因此，依据不同背景的攻击者，自适应地采取不同的隐私保护策略[26]，是隐私数据发布研究值得重视的方向。

### 3.5 面向应用的隐私保护研究

文献[27]提出了基于可用性的匿名方法，通过对属性加权来衡量不同属性在不同应用中的重要性。目前数据发布往往忽略了不同应用领域数据使用者的需求差别，对外发布的数据都是一样的。然而隐私保护的应用领域是广泛的，因此，在确保隐私信息不被泄露的前提下，需要人们研究提出更有效的隐私数据发布方法以最大限度地满足不同应用对数据精度的不同需求。

### 3.6 多属性数据集的隐私保护研究

目前的匿名化技术研究主要针对包含较少准标识符属性以及单一敏感属性的数据发布。然而现实当中，数据集中往往包含较多的准标识符属性和敏感属性，如果将现有的匿名化方法应用于多准标识符属性和多敏感属性的数据集，将会造成大量的信息损失和隐私的泄露。多属性数据集发布的隐私保护是公认的难题，需要人们进行深入的研究。

**参考文献：**


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

[2] Samarati P,Sweeney L. Protecting privacy when disclosing information: $k$-anonymity and its enforcement through generalization and suppression: proceedings of the IEEE Symposium on Research in Security and Privacy[R]. Technical report,CMU SRI, 1998:1-19.

[3] Sweeney L. Achieving $k$-anonymity privacy protection using generalization and suppression [J]. International Journal on Uncertainty,Fuzziness and Knowledge-based Systems,2002,10(5):571-588.

[4] Machanavajjhala A,Gehrke J,Kifer D. $L$-diversity: Privacy beyond $k$-anonymity: proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE'06)[C]. Atlanta,GA,USA:IEEE Press,2006: 24-36.

[5] Wong R C,Li J,Fu A W,et al. $(a,k)$-Anonymity:an enhanced $k$-anonymity model for privacy-preserving data publishing: proceedings of 12th ACM SIGKDD international conference on knowledge discovery and data mining[C]. New York: ACM Press, 2006: 754-759.

[6] Truta T M,Vinay B. Privacy protection: $p$-Sensitive $k$-anonymity property: proceedings of the 22nd International Conference on Data Engineering Workshops [C]. Washington DC: IEEE Computer Society,2006:94.

[7] Koudas N,Srivastava D,Yu T,et al. Aggregate query answering on anonymized tables: proceedings of the 23rd International Conference on Data Engineering[C]. Istanbul:IEEE Computer Society,2007:116-125.

[8] Li N,Li T. $t$-closeness:privacy beyond $k$-anonymity and l-diversity:proceedings of IEEE 23rd International Conference on Data Engineering [C]. Istanbul: IEEE Computer Society,2007:106-115.

[9] A Meyerson,R Williams. On the complexity of optimal $k$-anonymity: proceedings of the ACM SIGMOD-SIGACT-SIGART Principles of Database Systems[C]. New York:ACM Press,2004:223-228.

[10] Aggarwal G,Feder T,Kenthapadi K,et al. $k$-anonymity: algorithms and hardness [R]. Stanford University:Technical report,2004.

[11] Sweeney L. Guaranteeing anonymity when sharing medical data, the datafly system: proceedings of the 1997 American Medical Informatics Association Annual Fall Symposium[C]. Nashville, TN: [s. n.], 1997:51-55.

[12] Lefevre K,Dewitt D J,Ramakrishnan R. Incognito:

efficient full-domain $k$-anonymity: proceedings of the 24th ACM SIGMOD International Conference on Management of Data[C]. New York: ACM Press, 2005:49-60.

[13] Iyengar V. Transforming data to satisfy privacy constraints: proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM Press, 2002:279-288.

[14] Wang K, Yu P, Chakraborty S. Bottom-up generalization: a data mining solution to privacy protection: proceedings of the 4th IEEE International Conference on Data Mining[C]. Washington DC: IEEE Computer Society, 2004:249-256.

[15] Fung B, Wang K, Yu P. Top-down specialization for information and privacy preservation: proceedings of the 21st International Conference on Data Engineering [C]. Washington DC: IEEE Computer Society, 2005, 205-216.

[16] Lefevre K, Dewitt D, Ramakrishnan R. Mondrian multidimensional $k$-anonymity: proceedings of the 22nd International Conference on Data Engineering [C]. Washington DC: IEEE Computer Society, 2006: 25-34.

[17] Ji-WonByun, Ashish Kanira, Elisa Bertino, et al. Efficient $k$-anonymity using clustering technique[R]. CERIAS Technical Report 2006-10, West Lafayette, Indiana: Purdue University, 2006.

[18] Jiuyong Li, Raymond Chi-Wing Wong, Adawai-Chee Fu, et al. Achieving $k$-anonymity by clustering in attribute hierachical structures: proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery[C]. Krakow: [s. n.], 2006:405-416.

[19] Gagan Aggarwal, Tomas Feder, Krishnaram Kenihapadi, et al. Achieving anonymity via clustering: proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems[C]. New York: ACM Press, 2006:153-162.

[20] Duncan G, Feinberg S E. Obtaining information while preserving privacy: a markov perturbation method for tabular data: proceedings of Joint Statistical Meetings [C]. Anaheim, CA, 1997.

[21] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation: proceedings of the 32nd International Conference on Very Large Databases [C]. Seoul, Korea: Sept, 2006:139-150.

[22] 杨晓春,王雅哲,王斌,等. 数据发布中面向多敏感属性的隐私保护方法[J]. 计算机学报, 2008, 31(4): 574-586.

[23] Wong R, Liu Y, Yin J, et al. $(a,k)$-anonymity based privacy preservation by loss join: proceedings of the Advances in Data and Web Management, Joint 9th Asia-Pacific Web Conference, and 8th International Conference on Web-Age Information Management, Huangshan, Anhui, China [C]. Lecture Notes in Computer Science 4505, Springer, 2007:733-744.

[24] 刘喻,吕大鹏,冯建华,等. 数据发布中的匿名化技术研究综述[J]. 计算机应用, 2007, 27(10): 2361-2364.

[25] Xiaokui Xiao, Yufei Tao. Personalized Privacy Preservation: proceedings of the ACM SIGMOD International Conference on Management of Data[C]. New York: ACM Press, 2006:229-240.

[26] 陈珂. 开放式环境下敏感数据安全的关键技术研究[D]. 杭州: 浙江大学, 2007.

[27] Xu J, Wangw, Pei J, et al. Utility-based anonymization using local recoding: proceedings of the 12th ACM SIGK DD international conference on Knowledge discovery and data mining [C]. New York: ACM Press, 2006:785-790.

(责任编辑:韦廷宗)

(上接第272页)

[4] Shameem Akhter, Jason Roberts. Multi-Core programming: increasing perform-ance through software multi-threading[M]. 北京: 电子工业出版社, 2007.

[5] 卢开澄. 计算机密码学[M]. 北京: 清华大学出版社, 2003.

[6] 王晶,樊晓桠,张盛兵,等. 多核多线程结构线程调度策略研究[J]. 计算机科学, 2007, 34(9): 256-258.

[7] EI-Moursy A, Garg R, Albonesi D H, et al. Compatible phase co-Scheduling on a CMP of multi-threaded processors: proceedings of IEEE 20th International Parallel and Distributed Processing Symposium, April 25-29, 2006[C]. Rhodes Island, Greece, pp, 10-22.

[8] Cay S Horstmann, Gary Cornell. JAVA 核心技术:第二卷[M]. 第7版. 北京: 机械工业出版社, 2006.

(责任编辑:尹 闯)