数据发布中的隐私保护研究综述 *

兰丽辉[1,2], 鞠时光[1], 金 华[1], 刘善成[1]
(1. 江苏大学 计算机科学与通信工程学院,江苏 镇江 212013; 2. 吉林师范大学 计算机学院,吉林 四平 136000)

摘 要: 如何在发布涉及个人隐私的数据时保证敏感信息不泄露,同时又能最大程度地提高发布数据的效用,是隐私保护中面临的重大挑战。近年来国内外学者对数据发布中的隐私保护(privacy-preserving data publishing, PPDP)进行了大量研究,适时地对研究成果进行总结,能够明确研究方向。对数据发布领域的隐私保护成果进行了总结,介绍了常用的隐私保护模型和技术、隐私度量标准和算法,重点阐述了 PPDP 在不同场景中的应用,指出了 PPDP 可能的研究课题和应用前景。

关键词: 数据发布; 隐私保护; 匿名技术; 信息度量
中图分类号: TP309 **文献标志码**: A **文章编号**: 1001-3695(2010)08-2822-06
doi:10.3969/j.issn.1001-3695.2010.08.004

Survey of study on privacy-preserving data publishing

LAN Li-hui[1,2], JU Shi-guang[1], JIN Hua[1], LIU Shan-cheng[1]
(1. *School of Computer Science & Telecommunication Engineering, Jiangsu University, Zhenjiang Jiangsu* 212013, *China*; 2. *School of Computer Science, Jilin Normal University, Siping Jilin* 136000, *China*)

Abstract: When publishing the data set that is involved in personal privacy, the publisher should guarantee individual sensitive information security, simultaneously, as much as possible to improve the data usefulness and it is the great challenge in the privacy protection faces. In the recent years, the domestic and foreign scholars have conducted the extensive research. Carrying on the summary to the research results at the right moment, it can be clear about the research direction. This paper surveyed privacy protection achievements in the PPDP field, introduced the typical privacy protection model and technology, privacy metrics and algorithms. Focused on PPDP in different application scenarios, and pointed out that the PPDP possible research topics and application prospects.

Key words: data publishing; privacy preservation; anonymity technology; information metric

0 引言

随着信息技术的快速发展,使得关于个人数据信息收集的种类和数量呈指数增长。基于知识决策、信息共享、科学研究等的需要,要求数据收集者(个人、企业、政府等)将收集到的数据进行发布。但信息中可能涉及个人隐私,如果将收集到的原始数据直接进行发布,会使个人的敏感信息泄露。为了保证个人敏感信息的安全,要在发布数据的同时进行隐私保护。本文所指的隐私保护主要是针对个人敏感信息的安全;另外为了行文方便,在文中将数据发布中的隐私保护简写为 PPDP。如不特殊说明,文中所指的隐私保护都是针对数据发布而言。

为了更好地进行数据发布中的隐私保护研究,假设如下的条件:

a)发布环境。要求个体(数据持有者)向数据收集者提交真实数据,同时假设数据收集者即为数据发布者,并且不会进行隐私攻击,但是,使用发布数据的对象(数据接收者)可能进行隐私攻击;另外,数据发布者发布的是数据,而不是数据挖掘的结果。

b)属性分类。将待发布的数据表中的记录属性划分为四类:(a)显式标志符(EID),用来直接识别个体身份的信息,如身份证号、姓名等;(b)准标志符(QID),一些属性的组合,能够在背景知识(其他外部信息)的帮助下用来识别个体;(c)敏感属性(SA),隐私信息;(d)非敏感属性(NSA),不在上述三类中的属性。

c)背景知识。隐私攻击者除了能够访问发布的数据表外,还能从其他渠道获得一些关于目标对象的信息。通常情况下,假定攻击者可以获知目标对象的准标志符属性,如出生日期、性别、出生地、邮政编码等;也有一些攻击者可能通过文献资料、技术文档等获知发布的数据集采用的隐私模型、实现算法等。将攻击者可能获知的关于发布数据的信息称为背景知识。

PPDP 的关键在于实现隐私保护的同时保证数据的可用性。围绕这一思想,近年来对 PPDP 进行了大量研究。本文对数据发布领域的隐私保护成果进行了总结,介绍了常用的隐私保护模型和技术,以及隐私度量标准和隐私保护算法,重点阐述了 PPDP 在不同场景中的应用,对 PPDP 存在的问题、研究前

收稿日期: 2010-01-20; **修回日期**: 2010-03-09 **基金项目**: 国家自然科学基金资助项目(60773049);江苏省科技创新资金资助项目(sbc20080655)

作者简介:兰丽辉(1976-),女,吉林乾安人,讲师,博士研究生,主要研究方向为数据库安全、隐私保护(lanlihuicaoyue@163.com); 鞠时光(1955-),男,江苏镇江人,教授,博导,博士,主要研究方向为空间数据库、数据库安全;金华(1975-),男,江苏镇江人,讲师,博士研究生,主要研究方向为数据库安全、隐私保护; 刘善成(1985-),男,江苏盐城人,硕士研究生,主要研究方向为隐私保护.

景和新兴领域的应用进行了说明。

1　隐私保护模型

1980 年 Cox 最先提出用匿名的方法实现隐私保护，随后在 1986 年 Dalenius 针对人口普查记录集的隐私保护应用了匿名技术。但是，2002 年 Sweeney 在研究中发现，通过链接攻击仍然能使隐私泄露。为了避免受到隐私攻击，学者们提出了隐私保护模型作为指导发布者进行数据发布的原则。在众多隐私保护模型中比较典型和常用是 k-anonymity、l-diversity 和 t-closeness。其中，k-anonymity 是最早提出的隐私保护模型，随后很多模型受到其思想启发而提出。文献中对上述三个模型的介绍较多，除此之外还有一些其他的隐私保护模型。隐私保护模型的提出都是针对可能存在的隐私泄露情况，下面依据隐私泄露的几种情况对隐私保护模型进行介绍。

1）避免身份识别的隐私保护模型

攻击者能够在背景知识的帮助下确定发布的数据集中与目标对象匹配的记录，致使个人隐私信息泄露。为了防止攻击者进行记录链接攻击，就要使其不能通过背景知识惟一确定目标对象在发布表中所对应的记录。通常攻击者通过 QID 作为背景知识进行记录链接攻击，如果通过 QID 使攻击者不能惟一确定一条记录，就能够实现隐私保护的目的。

基于上述思想，文献[1,2]提出了 k-anonymity，文献[3]提出了 (X,Y)-anonymity，文献[4]提出了 multirelational k-anonymity。此类模型对 QID 进行泛化（第 2 章介绍）后，把记录划分成若干个等价类，每个等价类中至少 k 条记录，这样链接到某条记录的概率不超过 $1/k$，保证了记录安全。但是，如果每个等价类中记录的敏感属性取值相同或者某些敏感值出现的频率很高，则仍然存在隐私泄露的可能。

2）避免敏感属性泄露的隐私保护模型

攻击者无须准确匹配目标对象在发布表中的记录，根据准标志符，按照其所在的等价类能够推断出其敏感属性的取值。为了防止攻击者进行该类攻击，就要使其不能通过背景知识确定目标对象在发布表中敏感属性的可能取值。发布者应使得按 QID 得到的记录分组中敏感属性取值多样化，分布尽可能均匀。

基于上述思想，文献[5]提出 l-diversity 模型，要求每个 QID 组中对应的敏感属性至少有 l 个 well-represented 取值。文献[6]提出 t-closeness 模型，要求每个 QID 组中敏感属性值的分布和整个表中敏感属性值的分布接近，解决了偏斜攻击问题。文献[7]提出 (α,k)-匿名，要求等价类中至少有 k 条记录，且每个等价类中敏感属性的每个取值出现频率不超过 α。文献[8]提出个性化隐私保护，允许记录所有者自己设定隐私保护级别。该方法发布的数据缺损少、效用高，但是记录所有者应预先知道表（组）中记录敏感值的分布情况，这很难做到。如不能预先知道，就可能会设置较高的保护级别，影响数据的效用。

3）避免高概率推断的隐私保护模型

对于前两类攻击，攻击者已知目标对象的记录存在发布表中这一信息；而在有些时候，比如医院发布了一个重大疾病的数据表，攻击者只要确定目标对象在发布的表中，其实就意味着隐私泄露。攻击者通过访问发布的数据表，能够以很高的概率推断目标对象的记录是否存在发布的数据集中或者以较高的概率判断目标对象敏感属性取值。

为了解决这类概率攻击问题，发布者应尽可能保证攻击者在访问发布数据表前后得到的关于目标对象的信息相同，即应该实现“无信”原则，但由于背景知识的不确定性，不可能做到百分之百的无信。文献[9]提出的 δ-presence 模型以不超过 δ% 的概率推断目标对象的记录是否存在发布集中；Blum 等人[10]提出了适用于非交互查询模型的分布式隐私保护模型，Dwork[11]提出了 ε-差分隐私模型；Rastogi 等人[12]提出了 (d,γ)-隐私模型。它们均可以防止攻击者以较高概率推断目标对象的敏感属性取值。

由于背景知识的不确定性，在数据发布过程中隐私泄露不能完全避免，不存在完美的隐私保护。隐私保护模型在选择时应根据具体的应用场景和用户提出的隐私要求而定，没有一种模型可以避免所有可能的隐私攻击。

2　隐私保护技术

匿名化是最早提出的隐私保护技术，将发布数据表中涉及个体的标志属性删除之后发布。下面对基于匿名技术的方法进行介绍。

2.1　泛化与隐匿相结合的技术

泛化是采用较多的一种隐私保护方法。为了避免攻击者通过目标对象的 QID 进行隐私攻击，在数据发布中将 QID 中的属性具体取值用更概括、抽象的值替代。泛化方法要求构建 QID 中每个属性的分类树，对于不满足隐私要求的取值可以用分类树中双亲节点的值代替。根据隐私保护要求，可以采用全局编码方案（如全域泛化、子树泛化等）和局部编码方案（如单元泛化）。根据泛化的属性的数量又可分为单维泛化和多维泛化。

泛化不可避免地会带来数据的缺损，为了在保证隐私安全的同时又能最大程度地保证数据的效用，可以采用带有隐匿的泛化技术。隐匿即对不满足隐私保护要求的数据项删除，不进行发布。但是，过多的隐匿数据项也会影响数据的效用，所以通常在具体实现算法时，要对隐匿的数据项数量设定一个上限值。

2.2　基于有损连接的分割技术

泛化由于改变了原始的数据，数据缺损相对较大。分割技术不改变 QID 和敏感属性的值，而是通过降低两者之间的联系进而实现隐私保护。数据发布者将待发布的数据集分割成两个表，一个是包含 QID 的表（QIT），一个是包含敏感属性的表（ST），这两个表中有一个共同属性 GroupID。该方法没有改变原始数据，查询精确度高。但是，数据接收者在使用数据时，通过对分割的两个表进行链接会产生多余的记录，而且该方法不适用于连续的数据发布。

2.3　泛化与分割的结合——ANGEL

ANGEL 是最新提出的一种隐私保护方法，是将泛化与分割方法结合在一起。泛化方法实现简单，但数据缺损较大。为了提高数据的效用，可以针对泛化的结果集发布边缘视图，对泛化的信息进行补充。对泛化表和边缘视图采用分割的方法发布，这样既避免了泛化带来的较大的数据缺损，又实现了对视图边缘数据的安全发布。

2.4 微聚集技术

通过某种启发式方法将数据集划分为若干类,要求每个等价类中至少包含 k 个元组,类内数据最大程度地相似、类间数据最大程度地不同,然后用类质心来代替类内所有元组,从而实现数据集的 k-匿名化。由于微聚集用类质心取代类内元组的值,等价类内同质性越大,信息的损失量越小。

除了上述提到的匿名技术外,还有与泛化操作相对的细化方法、与隐匿方法相对的披露方法等;另外,使用数据加密和干扰技术也可以实现隐私保护。加密技术将敏感信息进行加密已保证其不泄露,安全性和准确度都很高,但是计算开销较大。干扰技术采用添加噪声、数据交换、合成伪数据等方法来实现隐私保护,实现效率高,但数据缺损较大。基于匿名的发布技术则折中了这两类方法的优缺点,相对容易实现,数据有效性高,是 PPDP 广泛应用的技术。

3 隐私度量标准和算法

具体的实现隐私保护需要依据隐私模型设计相应的算法。通常在算法实现过程中要考虑发布数据的效用和隐私保护度,攻击者的背景知识等因素。基于 k-anonymity 模型实现的算法较多,大多数算法的差异是泛化策略,搜索空间的剪枝、结束条件等。下面按照算法采用的信息度量标准进行介绍。

3.1 实现最小信息缺损的算法

最小信息缺损原则通过比较原始数据和匿名数据的相似度来衡量隐私保护的效果,信息缺损越小说明发布的数据集的有效性越高。但是,这种度量原则需要考虑 QID 中每个属性的每个取值的泛化和隐匿带来的信息缺损,计算代价较高,适用于对单个属性进行度量。采用该标准度量的算法,如 Sweeny 提出的 MinGen 算法、Samarati[13] 提出的针对泛化空间进行折半查找的 binary search 算法、LeFevr 等人[14] 提出的 incognito 算法。此类算法根据给定的最小信息度量标准找出最优的 k-匿名解决方案,主要采用全域泛化和记录隐匿操作来实现,要求对 QID 的全域泛化空间进行穷举搜索,以实现最优泛化。由于搜索空间代价较高,该类算法仅适用于较小规模的数据集的发布。

3.2 采用 ILoss 度量的算法

ILoss 是 Xiao Xiao-kui 等人在 2006 年提出的隐私保护度量标准,要求检查每条记录的 QID 中每个属性的取值泛化带来的信息缺损,进而计算出每条记录泛化后的信息缺损,根据每条记录的缺损计算整个表发布导致的信息缺损。他们提出的基于个性化隐私保护模型的 greedy personalized 算法就是以 ILoss 作为度量标准。

3.3 采用可辨识标准度量的算法

可辨识度量是通过对比原始数据和匿名发布后数据的可辨识程度进行度量。隐私保护的目的就是让攻击者不能识别(区分)目标对象的记录,所以原则上讲,数据发布后识别度越低隐私泄露的可能就越小。采用该标准度量的算法,如 LeFevr 等人[15] 提出的 Mondrian 多维泛化算法,Xu 等人[16] 提出的自底向上泛化和自顶向下细化的贪心算法,Li Ning-hui 等人提出的采用 t-closeness 模型的 incognito 算法。此类算法根据该度量标准进行最优或最小泛化,实现隐私保护的最优化(或最小化)匿名解决方案。

3.4 采用隐私/效用折中度量算法

数据发布中的隐私保护既要考虑尽可能不泄露隐私,也要最大可能地保证数据的有效性。下面给出的是两个折中的度量标准。

$$\mathrm{IGPL}(s)=\frac{IG(s)}{(PL(s)+1)} \tag{1}$$

$$\mathrm{ILPG}(g)=\frac{IL(g)}{(PG(g)+1)} \tag{2}$$

其中:式(1)采用的是从上到下的细化解决方案,$IG(s)$ 代表细化操作 s 获得的信息,$PL(s)$ 代表泄露的隐私;式(2)采用的是从下至上的泛化解决方案,$IL(g)$ 代表进行泛化操作 g 带来的信息缺损,$PG(g)$ 代表获得的隐私保护。针对实际的算法,$IG(s)$、$PL(s)$、$IL(g)$ 和 $PG(g)$ 可以根据具体的公式进行计算得出。

Wang Ke 等人[17] 提出的 TDS2P 算法、Fung 等人[18] 提出的 TDS 算法、Wang Ke 等人[19] 提出的 bottom-up generalization 算法分别采用了 IGPL 和 ILPG 的标准。除了提到的度量标准之外,还用一些其他的针对实际应用而提出的标准。例如,在全域泛化框架中要进行最小泛化,文献[20]提出的 DA 标准可以用来指导搜索泛化空间,优先选择 QID 中取值数量多的属性进行泛化。还有针对某一特定数据挖掘任务而设定的度量标准。例如,如果数据发布的目的是分类建模,那么分类准确度会直接影响建模后的分析结果,文献[21]提出的分类度量标准可以检测分类的正确性。算法在具体设计时,要根据发布数据的应用场景、有效性要求等选取度量标准。

4 应用场景

随着用户对数据发布要求的不断提高、新技术的不断出现,数据发布扩展到多个不同的应用场景,本章介绍几个典型的应用场景。

4.1 静态数据集中发布

早期大多数隐私保护研究都是针对集中式数据发布进行。所谓集中式数据发布即指将数据收集后统一保存在数据库服务器上,直接对一个数据表进行一次发布。前面提到的保护模型,算法大都适用于该场景。静态发布的一个特例是:根据不同数据接收者的数据分析要求,数据发布者可以在同一时刻发布针对同一基本表的多个视图。每个视图在发布中都各自满足隐私保护要求,但是如果将发布的多个视图进行连接就可能导致信息泄露。

例如,表 1 中 T 是原始数据,数据发布者根据两个用户的不同要求分别发布了表 2 中 T1 和表 3 中 T2。假设发布者不想攻击者通过{age,birthplace}属性链接到 disease 属性。攻击者通过连接 T1 和 T2 得到表 4 中 T3,可以推断出{30,US}的个体患有 cancer 的概率为 100%。

表 1　原始数据表 T

name	age	job	birthplace	disease	class
Mary	30	lawyer	US	cancer	C1
John	30	lawyer	US	cancer	C1
Sandy	40	carpenter	France	HIV	C2
Deak	40	electrician	UK	cancer	C3
Wood	50	electrician	France	HIV	C4
Alice	50	clerk	US	HIV	C4

表 2　T 的发布 T1

age	job	class
30	lawyer	C1
30	lawyer	C1
40	carpenter	C2
40	electrician	C3
50	electrician	C4
50	clerk	C4

表 3　T 的发布 T2

job	birthplace	disease
lawyer	US	cancer
lawyer	US	cancer
carpenter	France	HIV
electrician	UK	cancer
electrician	France	HIV
clerk	US	HIV

表 4　T1 和 T2 的连接 T3

age	job	birthplace	disease	class
30	lawyer	US	cancer	C1
30	lawyer	US	cancer	C1
40	carpenter	France	HIV	C2
40	electrician	UK	cancer	C3
50	electrician	France	HIV	C4
50	clerk	US	HIV	C4
30	lawyer	US	cancer	C1
30	lawyer	US	cancer	C1

针对存在的链接攻击,文献[22]提出了检测发布的多个视图链接是否违背 k-匿名的方法;文献[23]提出在对基本表进行匿名发布的基础上,可以发布边缘信息来增加发布数据的效用,并在 k-anonymity 和 l-diversity 隐私保护模型中应用了边缘发布的思想,提出了检测是否违背匿名要求的方法;文献[24]提出了在差分隐私模型中针对边缘数据的发布可能违背匿名原则的情况,并给出了检测和解决方法。

4.2　动态数据集发布

随着数据集的插入、删除和更新操作的不断进行,发布者应把变化后的数据及时进行发布,即重发布。重发布中由于先后发布的数据集结构完全相同,攻击者根据对比不同的发布可能推断出目标对象的隐私信息。例如,表 5 中 T1 是满足 2-多样性的发布,表 6 中 T2 是向 T1 中插入一条记录后的发布。假如攻击者知道目标对象的信息{female,lawyer}在 T2 中,不在 T1 中,通过对比 T1 和 T2 攻击者可以断定其感染的疾病是 HIV。表 7 中 T3 是从 T1 中删除记录{professional,female,diabetes},插入记录{professional, female, fever}后的发布。假如攻击者知道目标对象的信息{female,engineer}既在 T1 中又在 T2 中,对比 T1 和 T2 攻击者可以确认其感染了 cancer。

表 5　满足 2-多样性的发布 T1

job	sex	disease
professional	female	cancer
professional	female	diabetes
artist	male	fever
artist	male	cancer

表 6　T1 中插入记录后的发布 T2

job	sex	disease
professional	female	cancer
professional	female	diabetes
professional	female	HIV
artist	male	fever
artist	male	cancer

表 7　T1 中删除记录后发布 T3

job	sex	disease
professional	female	cancer
professional	female	fever
artist	male	fever
artist	male	cancer

针对重发布中隐私泄露问题,文献[25]最先提出了采用 l-diversity模型解决记录插入后的重发布问题。考虑了记录插入后可能受到的隐私泄露威胁,在后续发布的版本中包括了之前发布的所有记录。要求当前发布的 t_p 满足 l-多样性,而且对比之前任意一个发布的 t_i 和 t_p,应保证两个发布相差的数据集中每一条记录所属的等价类中至少有 $k-1$ 条。如果新插入的记录不能满足上述条件则延迟发布,这样可能导致一段时间内没有新记录发布,而且需要较大的缓存来保存这些延缓发布的记录。

文献[26]提出了 m-invariance 方法,可以实现记录插入和删除的动态发布。在具体实现 m-不变性规则时,采用了插入伪记录和对伪记录进行泛化来满足 m-不变性的要求。为了最大程度地保证发布数据的效用,除了发布数据表之外还发布了一个辅助的信息表,用来发布有关伪数据的统计信息。

文献[27]提出了 BCF-匿名方法。该方法指出在满足 k-匿名模型发布的数据集中,有些记录不可能是目标对象的候选记录,也就是说即使与目标对象同属一个等价类,攻击者能够在背景知识的帮助下确认这些记录不可能是目标对象的记录。因此,实际上不能实现完全的 k-匿名,等价类中不可区分的记录数量小于 k。BCF-匿名将这些记录数量 m 确定,用$|\text{qid}|-m$得到的数值才是真正的不可区分的记录的大小。

文献[28]则指出在后续的发布中,用户的 QID 和敏感属性值可能发生改变(如某个人搬家了,邮编就会改变,一个病人可能从一种疾病中康复又转为其他的疾病)。前三个方案中都假设用户的记录一经确定就不会发生改变。针对 m-invariance 存在的问题提出了 HD-compositon 规则,实现动态数据集的隐私保护(插入、删除、更新),实现了基于用户级的隐私保护,而不是在记录一级进行保护。

文献[29]也针对重发布中敏感值可以改变的记录如何保证在所有发布的版本中不泄露其隐私,提出全局担保的解决方案。前提是在每一个发布中每人只对应一条记录,而且除了 QID 之外,攻击者没有额外的背景知识,敏感属性的取值从一个发布到另一个发布的更新不受限制。

4.3　数据聚合发布

数据聚合(data mashup)是网络聚合技术的应用之一,是将一个或多个信息源整合起来的信息服务。由于从多个站点获取信息整合后提供给用户,在信息整合和发布的过程中存在着隐私泄露的风险。数据聚合发布不同于安全多方计算,安全多方计算各方共享的是计算结果而不是数据集。聚合发布的目的是使各方都能访问发布后的完整数据集,但是除此之外,各方都不应该获知其他数据持有者的信息。

在实际应用中可能存在多个组织要合作发布数据集的情况,比如有两家信誉卡公司要将双方各自的客户信息集中在一起进行信息欺诈检测或者双方需要把各自用户的信息提交给银行。但是由于涉及隐私和业务竞争的原因,双方都不愿将自己拥有的信息毫无保留地提交。例如表 8 中 T 是经过压缩处理得到的原始数据集。其中,party A 表示 A 用户拥有的数据集,party B 表示 B 用户拥有的数据集,shared 表明共有属性。将 A 和 B 拥有的数据集直接连接后会发现该记录集中最后一条记录{F,lawyer,44K}是惟一的记录,会导致信息泄露。如果在连接之前将 lawyer 和 accountant 按图 1 中的分类树泛化为 professional,就不会发生信息泄露。

表 8　原始数据表 T

shared		party A		party B		
SSN	class	sex	…	job	salary	…
1 ~ 3	0Y3N	M	…	janitor	30k	…
4 ~ 7	0Y4N	M	…	mover	32k	…
8 ~ 12	2Y3N	M	…	carpenter	35k	…
13 ~ 16	3Y1N	F	…	technician	37k	…
17 ~ 22	4Y2N	F	…	manager	42k	…
23 ~ 25	3Y0N	F	…	manager	44k	…
26 ~ 28	3Y0N	M	…	accountant	44k	…
29 ~ 31	3Y0N	F	…	accountant	44k	…
32 ~ 33	2Y0N	M	…	lawyer	44k	…
34	1Y0N	F	…	lawyer	44k	…

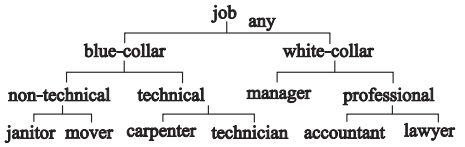

图1　属性job的分类树

解决上述问题有两种方法:a)先将数据聚合到一起,然后共同泛化,但这种方法会使得双方的隐私都泄露给对方;b)先各自对拥有的数据集进行泛化,然后聚合在一起进行发布。但这种方法在泛化时划分的等价类要涉及两个或多个数据发布者,实际操作很困难。Wang 等人提出了 TDS2P 算法,针对双方参与的数据聚合发布给出了解决方案,能实现与方法 a)相同的发布效果,而且没有在多个发布者中泄露其他发布者的隐私。具体实现时要求参与聚合的两方各自对数据表按信息度量标准选取属性进行自顶向下的细化操作,并根据本地属性的泛化操作指导对方进行泛化,直到所有 QID 中的属性泛化完成,返回泛化的结果集。Mohammed 等人[30]对该算法又进行了深入的研究,进一步拓展了网络聚合应用。Jiang 等人[31]按照方法 b)的思想并结合加密技术实现了数据聚合的安全发布,具体方法为:每一数据发布方都先按照 k-anonymity 要求对本地的数据表进行匿名泛化;根据每一条记录的 ID 标志把满足 k-匿名的本地表进行连接,对比每个 QID 组中交集的记录数量,如果大于等于 k,则直接将本地表进行连接发布即可,整体满足 k-匿名要求。如果数量小于 k,则各方需要对本地数据表继续进行泛化,然后再比较,直到满足隐私保护要求为止。为了保证在进行连接时不泄露隐私,要求对 ID 进行加密处理,防止各方按照记录 ID 定位个人信息,窃取隐私。

4.4　社会网络数据发布

由于信息技术的不断发展和 Web 2.0 技术的广泛应用,越来越多的社会网络网站形成,如以社交为目的的在线社区 Myspace 和 Facebook;以职业和商务服务为目的的商务网络 linked in 和 we@ link(若邻网);以找回联络为目的的校友网络校内网;以婚恋为目的的在线配对网络百合网。社会网络将实体及实体间的联系融合在一起,其中包含着丰富的信息。社会网络数据在发布过程中也要进行隐私保护,比较简单的方法就是采用匿名技术。可以用数据结构中的图来描述社会网络,$G=(V,E,L,L_V,L_E)$。其中,V 表示顶点集,E 表示边集,L 表示标签集,L_V表示顶点的标签集,L_E表示边的标签集。如图 2 所示,(a)中给出的是顶点带有标签的电子邮件通信的社会网络,(b)是将标志用户信息的姓名隐匿,并对标签属性进行泛化后的发布[32]。虽然进行了匿名发布,但是攻击者可以将目标顶点在图中所处的位置作为背景知识进行推断。例如,如果攻击者知道 Sunny 在该网络中的对应顶点的度为 4,与其有通信的人在网络中的对应顶点的度分别为 2、3 和 2,那么攻击者可以惟一地识别出目标个体 Sunny 在匿名社会网络中的对应顶点,因为在(b)中只有一个顶点符合 Sunny 的结构特征。已有的隐私保护模型和算法都是针对传统的关系型数据,不能将其直接移植到社会网络发布中。原因在于:攻击者的背景知识更加复杂也更难模拟;不能通过简单地对比匿名前后的网络进行信息缺损判断,社会网络数据的发布信息缺损度量标准复杂;由于背景知识和度量标准的复杂性使得设计社会网络数据的匿名处理方法更具有难度和挑战性。目前,国外学者对社会网络数据发布中的隐私保护研究较多,提出了一些解决的方案,但该领域的研究尚属初期阶段,面对复杂的社会网络应用有更多的问题需要提出和解决。

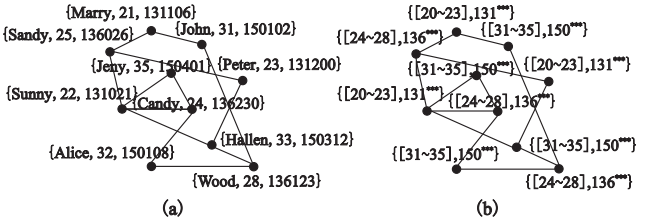

图2　电子邮件通信的社会网络

除了上述提到的应用场景之外,针对具体领域的数据发布,如医疗数据发布中可能存在多敏感属性、电子商务中的高维数据、基于位置的移动对象数据、文本信息等的安全发布都会带来一些新的应用场景。

5　结束语

数据发布中的隐私保护是隐私保护研究领域的一个分支,是近年来新兴的研究课题,对该领域的研究已经取得了不少成果。本文针对数据发布中的隐私保护模型、隐私保护技术、隐私保护算法、隐私保护应用场景和可能的研究方向及应用前景等方面的内容进行了综述,重点介绍了在不同应用场景中的隐私保护。PPDP 的应用广泛,具有很大的研究空间。

a)集中式静态数据集发布的隐私保护实现较多,技术比较成熟,可以考虑将已有的隐私保护模型和算法进行改进应用到新的场景中。另外,针对待发布的数据集如何正确选择准标志符属性;如何解决最小化攻击问题;如何针对发布的数据集进行查询,保证在查询中不泄露隐私。这些问题虽然已有学者提出解决的方法,但是在具体的应用场景中仍然是值得研究的课题。

b)针对非关系型数据的隐私保护,大多数的研究都是针对传统的关系型数据进行。在实际应用中,还有很多非关系型数据在发布中仍面临隐私攻击导致信息泄露的可能。例如在商业活动和公共服务中发布的高维数据、移动对象数据、文本数据等,这些数据通常包含丰富的信息,可能会导致敏感信息的泄露。

c)新兴领域的隐私保护技术运用。随着新技术的不断涌现,如社会网络、RFID、生物信息学、网络聚合应用等,这些领域中的隐私保护也是值得研究的课题。

目前,如何将隐私保护取得的研究成果应用在现实的数据发布中是面临的一个重要问题。可以考虑为个人提供隐私保护工具,如保护隐私的 Web 浏览器、电子商务中最小信息缺损的协议等;在工程实施的过程中纳入隐私保护,还有些技术的

应用需要提出正式的规范和检验工具。

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

[3] WANG Ke, FUNG B C M. Anonymizing sequential releases[C]// Proc of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2006:414-423.

[4] NERGIZ M E, CLIFTON C, NERGIZ A E. Multirelational k-anonymity[J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21(8):1104-1117.

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

[6] LI Ning-hui, LI Tian-cheng, VENKATASUBRAMANIAN S. t-closeness: privacy beyond k-anonymity and l-diversity[C]//Proc of the 23rd International Conference on Data Engineering. Istanbul: IEEE Computer Society, 2007:106-115.

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

[8] XIAO Xiao-kui, TAO Yu-fei. Personalized privacy preservation [C]//Proc of ACM SIGMOD Conference on Management of Data. Chicago: ACM, 2006:229-240.

[9] NERGIZ M E, ATZORI M, CLIFTON C W. Hiding the presence of individuals from shared databases[C]//Proc of ACM SIGMOD International Conference on Management. New York: ACM, 2007:665-676.

[10] BLUM A, LIGETT K, ROTH A. A learning theory approach to non-interactive database privacy[C]// Proc of the 40th Annual ACM Symposium on Theory of Computing. New York: ACM, 2008:609-618.

[11] DWORK C. Differential privacy[C]// Proc of the 33rd International Colloquium on Automata, Languages and Programming. 2006:1-12.

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

[13] SAMARATI P. Protecting respondents' identities in microdata release [J]. IEEE Trans on Knowledge and Data Engineering, 2001, 13(6):10-27.

[14] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Incognito: efficient full-domain k-anonymity[C]//Proc of ACM SIGMOD International Conference on Management. New York: ACM, 2005:49-60.

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

[16] XU Jian, WANG Wei, PEI Jian, *et al*. Utility based anonymization using local recoding[C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2006:785-790.

[17] WANG Ke, FUNG B C M, DONG Guo-zhu. Integrating private databases for data analysis[C]//Proc of IEEE International Conference on Intelligence and Security Informatics. Atlanta: Springer-Verlag, 2005:171-182.

[18] FUNG B C M, WANG K, WANG Ling-yu, *et al*. Privacy-preserving data publishing for cluster analysis[J]. Data & Knowledge Engineering, 2009, 68(6):552-575.

[19] WANG Ke, YU P S, CHAKRABORTY S. Bottom-up generalization: a data mining solution to privacy protection[C]//Proc of the 4th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2004:249-256.

[20] SWEENEY L. Datafly: a system for providing anonymity in medical data[C]//Proc of the 11th IFIP TC11 WG11.3 International Conference on Database Securty XI: Status and Prospects. London: Chapman & Hall, 1998:356-381.

[21] IYENGAR V S. Transforming data to satisfy privacy constraints[C]// Proc of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002:279-288.

[22] YAO Chao, WANG X S, JAJODIA S. Checking for k-anonymity violation by views[C]//Proc of the 31st International Conference on Very Large DataBases. Trondheim: VLDB Endowment, 2005:910-921.

[23] KIFER D, GEHRKE J. Injecting utility into anonymized datasets [C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM, 2006:217-228.

[24] BARAK B, CHAUDHURI K, DWORK C, *et al*. Privacy, accuracy, and consistency too: a holistic solution to contingency table release [C]//Proc of the 26th ACM Symposium on Principles of Database Systems. New York: ACM, 2007:273-282.

[25] BYUN J W, SOHN Y, BERTINO E, *et al*. Secure anonymization for incremental datasets[C]//Proc of the 3rd VLDB Workshop on Secure Data Management. [S.l.]: Springer-Verlag, 2006:48-63.

[26] XIAO Xiao-kui, TAO Yu-fei. m-invariance: towards privacy preserving re-publication of dynamic datasets[C]//Proc of ACM SIGMOD Conference on Management of Data. New York: ACM, 2007:689-700.

[27] FUNG B C M, WANG Ke, FU A W C, *et al*. Anonymity for continuous data publishing[C]//Proc of the 11th International Conference on Extending Database Technology. New York: ACM, 2008:264-275.

[28] BU Ying-yi, FU A W C, WONG R C W, *et al*. Privacy preserving serial data publishing by role composition[J]. Proceedings of the VLDB Endowment, 2008, 1(1):845-856.

[29] WONG R C W, FU A W C, LIU Jia, *et al*. Preserving individual privacy in serial data publishing[J/OL]. (2009). http://arxiv.org/pdf/0903.0682.

[30] MOHAMMED N, FUNG B C M, WANG Ke, *et al*. Privacy-preserving data mashup[C]//Proc of the 12th International Conference on Extending Database Technology. New York: ACM, 2009:228-239.

[31] JIANG Wei, CLIFTON C. Privacy-preserving distributed k-anonymity [C]//Proc of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security. Berlin: Springer, 2005:166-177.

[32] 魏琼. 数据发布中的隐私保护方法研究[D]. 武汉:华中科技大学,2008.

[33] FUNG B C M, WANG Ke, YU P S. Top-down specialization for information and privacy preservation[C]//Proc of the 21st IEEE International Conference on Data Engineering. Washington DC: IEEE Computer Society, 2005:205-216.

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

[35] TAO Yu-fei, CHEN He-kang, XIAO Xiao-kui, *et al*. ANGEL: enhancing the utility of generalization for privacy preserving publication [J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21(7):1073-1087.

[36] 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):847-860.

[37] FUNG B C M, WANG Ke, CHEN Rui, *et al*. Privacy-preserving data publishing: a survey on recent developments[J]. ACM Computing Surveys, 2010, 42(4):1-55.

[38] 杨晓春,刘向宇,王斌,等. 支持多约束的 k-匿名化方法[J]. 软件学报,2006,17(5):1222-1231.

[39] 韩建民, 岑婷婷, 虞慧群. 数据表 k-匿名化的微聚集算法研究[J]. 电子学报,2008,36(11): 2021-2029.