**数据发布中多敏感属性数据隐私保护方法研究**

**摘要**

**英文摘要**

**目录**

1. **绪论**

1.1数据发布研究背景及意义

1.2数据发布中隐私研究保护研究现状

1.3数据发布中的隐私规则

1.4本文主要研究内容与组织架构

1. **数据发布中的隐私保护**

2.1 隐私保护中的相关理论

2.2 隐私保护中数据发布匿名模型介绍

有损连接发布

2.3 隐私保护原则与信息度量

2.4 本章小结

1. **多维敏感属性数据发布中的隐私泄露**

3.1 多敏感属性数据发布中的问题研究分析

在对数据发布中的隐私保护技术研究初期，大多数敏感数据发布方法都是针对单一敏感属性的保护。但是，在实际的应用中，发布的数据大多数都会涉及到多个敏感属性，特别是这些敏感属性在某些情况下会存在一些关联关系，一些属性虽然对于发布个体不是直接的敏感属性，但是这些属性却和个体的敏感属性有着明显的特定关系，所以这样的属性也应该归类到个体的敏感属性被保护。例如表3-1-1 为将要发布的原始医疗信息，从表中可以看出，敏感属性主治医(Physician) 和疾病(Disease)之间存在着关联性，通过背景知识我们也可以知晓，一个主治医生专治哪几种疾病，其具体的关联性如表3-1-2 所示。

表3-1-1 原始医疗信息表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Tuple ID | Name | Age | Sex | Zipcode | Physician | Disease |
| t1 | Sam | 23 | M | 821071 | John | Flu |
| t2 | Anne | 44 | F | 821023 | John | Pneumonia |
| t3 | Mike | 56 | F | 821045 | John | Cancer |
| t4 | Lily | 35 | M | 821123 | Bob | Flu |
| t5 | Harry | 25 | F | 821031 | Bob | Pneumonia |
| t6 | Mona | 30 9 | M | 821035 | Anne | Gastritis |
| t7 | Tony | 40 | F | 821110 | Anne | Gastritis |
| t8 | Lucy | 37 | M | 821115 | Hugo | HIV |
| t9 | Tim | 60 | M | 821134 | Marry | Flu |

3-1-2 属性间关联表

|  |  |
| --- | --- |
| Physician | Disease |
| John | Flu,Pneumonia,Cancer |
| Bob | Flu,Pneumonia |
| Anne | Gastritis |
| Hugo | HIV |
| Marry | Flu |

现在假如我们需要对3-1-1表数据内容进行数据发布，并以l-diversity 匿名算法为例进行。我们对表3-1-1原始医疗信息进行匿名化处理，并通过分组算法得到数据发布表，结果如表3-1-3 所示。由表3-1-3 可以看出，由于医院中医生的主治哪些疾病是可以很容易获取的，也就是攻击者能够很容易获得表3-1-2内容的背景知识，若攻击者得知该个体的Physician 属性值为“John”再联合攻击者掌握的个体的准标识属性确定个体属于Group ID为3的分组。此时，攻击者推测出该个体的Disease 属性值的概率将高于1/3，这就违反了L多样多样性原则，造成隐私泄露的风险升高。这正是由于敏感属性Physician和Disease 间存在着关联性，即使数据发布表中的敏感属性满足l-diversity 匿名模型，依然存在隐私泄露风险。因此针对单敏感属性的隐私保护技术并不能直接用于多敏感属性数据发布，否则会给个体隐私数据的保护带来很大的挑战，所以对于多敏感属性的数据发布仍存在隐私泄漏风险。为了适应实际应用中的数据发布，面向多敏感隐私属性的数据发布方法的研究应受到重视。

表3-1-3 医疗发布数据3-diversity匿名表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tuple ID | Age | Sex | Zipcode | Physician | Disease |
| t1 | [20,30] | M | 8210\*\* | John | Flu |
| t5 | [20,30] | F | 8210\*\* | Bob | Pneumonia |
| t6 | [20,30] | M | 8210\*\* | Anne | Gastritis |
| t8 | [31.40] | M | 821\*\*\* | Hugo | HIV |
| t4 | [31,40] | M | 821\*\*\* | Bob | Flu |
| t7 | [31,40] | F | 821\*\*\* | Anne | Gastritis |
| t3 | [41,] | F | 821\*\*\* | John | Cancer |
| t2 | [41,] | F | 821\*\*\* | John | Pneumonia |
| t9 | [41,] | M | 821\*\*\* | Marry | Flu |

3.2 多敏感隐私属性的数据发布方法

因为多敏感属性数据隐私保护有一些特殊的要求，为了防止由于缺少整体性而造成失去隐私敏感属性的连锁攻击，许多学者提出了针对多敏感属性数据的隐私保护模型和方法。文献[3]提出了多敏感属性l-diversity 概念，并对其进行如下定义：

多敏感属性l-diversity. 设数据表T 中有若干个准标识符属性QI 和敏感属性SA，从QI 中任意选取一个属性将其作为唯一的敏感属性，其余的准标识符属性和敏感属性均作为准标识符属性，则此时若数据表T 均满足l-diversity，则说明数据表T 满足多敏感属性l-diversity。

由以上定义可知，多敏感属性l-diversity 规则要求每个敏感属性上的每一个敏感值与所有其他敏感属性上对应的敏感值的个数不少于l 个[12]，这一概念可以很好的解决多敏感属性数据发布的隐私保护问题。但是，当敏感属性个数增加时，每个等价类为了满足多敏感属性l-diversity 规则就必须包含更多的记录，这必然会导致数据表泛化程度加剧，从而造成大量的信息损失。

3.2.1基于多维桶分组技术的隐私数据发布方法

2008 年，杨晓春等人[2] 首次提出了以l-diversity 模型为基础的多维桶分组技术来解决多敏感属性数据发布的隐私保护问题。

多维桶分组技术的基本思路是，首先，将多个敏感属性看成一个高维复合敏感属性向量，也就是说，一个敏感属性对应一维；其次，使用多维桶的向量模型，将数据表中的记录映射到多维桶上；最后，按照某种方法在构造的多维桶上进行分组操作，使分组中的记录尽可能是在各维度上取值都不相同的桶中提取出来的。

多维桶分组技术重点是将多个敏感属性作为高维复合敏感属性来构造桶，并提出以下定义：

假设用户待发布的数据表为T{A1，A2，A3，···，Ap，S1，S2，···，Sd}，其中Ai（1ip）是待发布数据T的准标识属性（QI），p代表准标识属性的个数。Sj（1jd）是待发布数据T的敏感属性，d代表敏感属性的个数。设T中记录个数为n，即|T|=n，那么发布数据表中每条记录记为ti（1in），另t[X]标识记录t在X属性上的取值，其中X{A1，A2，A3，···，Ap，S1，S2，···Sd}。

（1）复合敏感属性.。待发布数据表T中所有的敏感属性构成一个复合敏感属性，记作S。其中第i个敏感属性可看作复合敏感属性的第i 维，记为Si， Dom（Si）为Si 的取值范围，指该敏感属性的所有取值，|Si|为D（Si）的基数，指该敏感属性取值的个数。

（2）复合敏感属性向量[2]. 待发布数据表T中任意记录t的全部敏感属性取值构造成向量模式<t[s1]，t[s2]，…，t[sd]>，称这样的向量模式为复合敏感属性向量。

（3）分组[4]. 一个分组是T 中记录的子集。T 中每一个记录属于且仅属于一个分组，T中所有记录的分组记为GT{G1，G2，G3，···，Gm}，其中m为最终分组数，并且 =T，并且=（1ijm）。

（4）单敏感属性L-多样性[5]。对于一个分组G，G中只包含单敏感属性的记录，假设v为G中出现频度最大的敏感属性取值，且c(v)为v在G中出现的次数，如果满足，|G|为G 中记录的个数。

（5）复合敏感属性L-多样性分组. 对于一个包含复合敏感属性的分组G，如果G中的任一一维敏感属性Si（1id）都满足单敏感属性L-多样性，则该分组满足复合敏感属性满足l-多样性。那么对于T中所有分组GT{G1，G2，G3，···，Gm}，如果其中每个分组Gi（1im）都满足复合敏感属性L-多样性性质，则称GT为T上的复合敏感属性L-多样性分组。

有了以上定义以后，我们知道待发布数据表T若存在多个敏感属性，仅仅满足单敏感属性的L-多样性原则无法保证所有敏感属性的隐私保护需求，针对多敏感属性的数发布，需要满足多敏感属性中每个属性都满足L-多样性原则，即需要得到待发布数据表T的复合敏感属性L-多样性分组。杨晓春等人提出基于有损连接技术的支持多敏感属性的隐私数据发布多维桶分组技术，目标是找到具有多敏感属性的待发布数据表T的分组方案，使得到的分组均满足复合敏感属性L-多样性。

多维桶的构造原理如下：复合敏感属性的每个维度对应多维桶的一维，将数据表T中的数据记录根据其复合敏感属性向量每一维的值分别映射到相应的桶中。设数据表T 的敏感属性个数为d，构造的d 维桶记为BUK（S1，S2，……，Sd），其中每个桶记为buk（s1，s2，……，sd），每个桶的大小记为size（buk（s1，s2，……，sd）），即含的记录数。以表 3-1-1 的原始数据为例，建立多维桶如表3-1-4 所示。

表 3-2-1 由表3-1-1构造的d维桶（d=2）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Flu | Pneumonia | HIV | Gastritis | Cancer |
| John | {t1} | {t2} |  |  | {t3} |
| Bob | {t4} | {t5} |  |  |  |
| Anne |  |  |  | {t6,t7} |  |
| Hugo |  |  | {t8} |  |  |
| Marry | {t9} |  |  |  |  |

在得到表3-1-4的多维桶后，分别采取MSB的贪心算法，包括最大桶优先算法，最大单维桶优先算法和最大多维桶优先算法来得到最后的满足复合敏感属性L-多样性分组的发布数据T’。

基于多维桶的分组算法为复合敏感数据发布的隐私保护提供了分组方案，并解决了多敏感属性下的隐私泄露问题，但仍然存在一些不足之处：

（1）多维桶分组技术在实现分组算法的时候，只考虑了最后分组对原数据的覆盖率问题，以贪心策略尽可能的得到更多的满足复合敏感属性L-多样性的分组，没有考虑到在实际数据发布中，每一维敏感属性的取值中，可能存在敏感度高低的问题，例如医疗数据发布表当中，个体敏感属性Disease取值中，HIV的敏感度肯定高于Flu。这就导致，多维桶分组算法在得到待发布数据表的所有分组后，可能会存在某个分组中，某个敏感属性的取值均是敏感属性度较高的情况，出现敏感信息分布切斜的现象，从而导致个体的敏感信息泄露。

（2）多维桶分组技术考虑到实际应用中某些元组包含着更重要的信息，需要尽量保留在发布的数据中，从而提出加权多维桶分组技术。加权多维桶分组技术只是考虑到部分重要数据的权值，在分组时先将权值高的记录先加入分组，这种方式虽然达到了保留重要数据的目的，但是由于重要数据一般都是敏感属性较高的数据，这种加权分组方式在实际应用中很有可能加剧敏感信息倾斜，造成个体隐私泄露。为解决这一问题，我们应该在保留高权值敏感信息的情况下同时保证高敏感度信息在分组中分布的均匀性。

3.2.2 基于L-覆盖性聚类分组的隐私数据发布方法

金华等人分析多维桶分组算法存在分组效率低，可能会存在由于每次分组选取桶的顺序问题造成大量不必要的数据记录遭到隐匿的问题，进一步对多敏感属性的医疗数据发布方法进行研究，提出了基于有损连接技术和相同敏感属性集的L-覆盖性聚类分组算法[6]。在多维桶提出的多敏感属性L-多样性的基础上给出以下定义：

（1）移除。对于一个分组G，若移除G中任意一条记录中的某一敏感属性值t[Si]（tG，1id）,需要将G中所有包换t[Si]的记录都删除。

（2）多敏感属性L-覆盖性。在一个分组G中，若至少需要移除L个敏感属性值，才能将G中所有记录移除，则称分组G满足多敏感属性L-覆盖性。

（3）相同敏感属性集。对于待发布数据表T中，包含同一敏感属性取值的所有记录组成的集合称为相同敏感属性集，记为SID（v）。待发布数据表T中，记录t的所有敏感属性值t[Si]的相同敏感属性集的并集称为记录相同敏感属性集，记为t.TSID，t.TSID=(t[Si])。对于一分组，中所有记录的记录相同敏感属性集的并集称为分组相同敏感属性集，记为.GSID,.GSID=。

（4）平均概率泄露度。由于需要处理完所有剩余记录，数据集T中肯定会存在仅仅满足多敏感属性L-覆盖性而不满足复合敏感属性L-多样性的分组。设得到的发布数据分组为GT{G1，G2，G3，···，Gm},||L，G中所有敏感属性的不同取值个数为n，每个敏感属性的取值为（1in）,在的出现的频率记为，则分组的概率泄露度定义为,则数据集的平均概率泄露度为=，其中|T|标识待发布数据集T中记录总数。

L-覆盖性聚类分组方法的主要思想是通过聚类的方法将满足L-覆盖性的记录进行分组，采用聚类思想，首先在数据集中顺序选取L个满足L-覆盖性的记录构成分组。然后对剩余记录分两步处理：第一步处理剩余记录中可以添加到其他分组而且仍满足多敏感属性L-多样性性质的记录；第二步是处理第一步剩余的记录，将剩余记录均匀地添加到分组较小的分组中，以降低平均概率泄露度。我们以表3-1-1为原始数据表为例，利用L-覆盖性聚类分组算法进行分组。另L=3，首先将数据集T中的每一条记录看做是一个分组，并计算他们各自的GSID。根据GSID的定义我们可以得到G1.GSID={t1,t2,t3,t4}，G2.GSID={t1,t2,t3,t5,t9}，G3.GSID={t1,t2,t3}，G4.GSID={t1,t4,t5,t9}，G5.GSID={t2,t4,t5}，G6.GSID={t6,t7}，G6.GSID={t6,t7}，G7.GSID={t6,t7}，G8.GSID={t8}，G9.GSID={t1,t4,t9}。按照L-覆盖性聚类算法顺序选取规则，首先选取G1，然后选择不在G1.GSID中含有的记录分组G6，得到G1,6.GSID={t1,t2,t3,t4,t6,t7}，然后选取不在G1,6中含有的记录分组G8，此时已得到一个分组{t1,t6,t8}，然后移除记录t1，t6，t8。算法依次循环进行得到另一个分组{t2,t7,t9}，剩余记录{t3,t4,t5}无法满足L-覆盖性（L=3），t5能够加入到第一个分组中得到{t1,t5,t6,t8}满足复合敏感属性L-多样性（L=3），

t3与t4加入到第二个分组得到{t2,t3,t4,t7,t9}，满足L-覆盖性（L=3）,最终得到的发布结果如表3-2-2所示。

表 3-2-2 L-覆盖性积累分组算法发布数据结果

|  |  |  |
| --- | --- | --- |
| Tuple ID | QIs | Group ID |
| t1 | ··· | G1 |
| t2 | ··· | G2 |
| t3 | ··· | G2 |
| t4 | ··· | G2 |
| t5 | ··· | G1 |
| t6 | ··· | G1 |
| t7 | ··· | G2 |
| t8 | ··· | G1 |
| t9 | ··· | G2 |

|  |  |  |
| --- | --- | --- |
| Group ID | Physician | Disease |
| G1 | John | Flu |
| Bob | Pneumonia |
| Anne | Gastritis |
| Hugo | HIV |
| G2 | John | Pneumonia |
| John | Cancer |
| Bob | Flu |
| Anne | Gastritis |
| Marry | Flu |

L-覆盖性分组算法解决了多敏感属分组后存在的剩余记录的问题，并提出了隐私平均概率泄露度的概念，但是含有敏感属性的数据发布中，最重要的就是保护个体记录的隐私不被泄露，所以该算法的不足之处也显而易见：

（1）满足L-覆盖性的复合敏感属性分组，并不一定满足复合敏感属性L-多样性，特别是在剩余记录存在大量相似敏感属性的时候，由于L-覆盖性算法在分组完成后为了将剩余记录全部加入已分组当中，会造成部分分组中敏感属性相同的记录出现的概率大与1/L。

（2）L-覆盖性分组算法总是顺序选取记录来进行分组，而且没有考虑敏感属性的度量问题，该算法虽然解决了剩余记录问题，但却在个性化分组和分组效率上存在比较大的局限性，分组效果通常不理想，在剩余记录较多的情况下，容易造成个体记录的隐私泄露。

3.3 多敏感属性的个性化隐私保护

通常情况下，对于待发布数据表T，都是由数据发布者来决定数据发布表中的敏感属性信息，通过设定一些约束参数来对待发布数据进行隐私保护。自从Xiao等人在2006年首次提出针对数据发布中的个性化匿名发布概念以后，个性化隐私保护就成为了数据发布中隐私保护研究的重要研究方向。

个性化隐私保护是指在进行隐私保护时，由数据发布者决定待发布数据表中的隐私属性，针对不同场景和发布数据指定不同的隐私保护策略与个体隐私保护的强度，从而满足不同的人对不同敏感属性的不同约束要求，在保护个体隐私的前提下，达到数据个性化发布的目的。所以个性化隐私保护可以很好的满足不同场景下对于不同隐私得保护的要求，并能够在一定程度上克服全局准标识属性匿名化编码造成的对敏感属性的保护“不足”和“过度”保护等问题。个性化数据法布中隐私保护目前的研究主要分为两个反面，一类是面向数据发布表中每条个体记录的个性化隐私保护；另一类则是针对数据发布表中所有敏感属性取值的个性化隐私保护。

1. 面向个体记录的个性化隐私保护方法

面向个体记录的个性化隐私保护策略的主要研究对象是针对待发布数据表中的每条个体记录，即研究的对象是个体。在数据发布者处理待发布数据时，为满足个性化数据发布的需求，需要从每个个体的实际隐私保护需求出发，对每条个体记录制定不同的个性化约束，对个体和与其相关的敏感属性之间的关联性进行一定的约束和限制。

面向个体记录的个性化隐私保护方法从本质上看来能够很好的制定数据发布表的个性化发布方案，但实际应用中待发布的数据集一般比较大，若要为数据集中每条记录都设定不同的个性化约束就需要非常大的任务量，因此，这种面向个体记录的单独设定个性化约束的数据发布方式虽然达到了良好的个性化需求，但在实际操作时缺乏可行性，存在一定的局限性。

（2）面向敏感属性值的个性化隐私保护方法

面向敏感属性值的个性化隐私保护方法主要是针对待发布数据表中的敏感属性的所有取值，以发布数据的敏感属性值为基础，数据发布者可以根据数据表中的敏感隐私信息不同的敏感值设定不同的个性化约束，实现个性化隐私保护。

面向敏感值的个性化隐私保护方法与面向个体记录的个性化隐私保护方法具有更高的可行性，可以很好的解决数据发布中隐私保护技术对隐私信息的过度保护从而造成发布数据的可用性降低或对隐私信息的保护不足而造成隐私泄露的问题。目前针对对于个性化隐私保护的技术的研究大都针对于单个隐私属性的数据发布，在待发布数据存在多敏感属性的情况下，单敏感属性的个性化隐私保护方法并不适用，因此针对多敏感属性的个性化隐私保护数据发布方法还需要更加深入的研究。

3.3.1 面向多敏感属性的个性化数据发布算法

（1）完全（a,k）-anonymity模型

韩建明等人[8]根据简单（a,k）-匿名模型和一般（a,k）-匿名模型提出完全（a,k）-匿名模型，该模型主要思想是根据不同敏感属性质的敏感度不同设置不同的频率约束a，以此实现对不同敏感属性的值在同意等价类分组中出现的频率进行控制，实现针对敏感属性值的个性化分组，达到个性化发布的目的。

简单（a,k）-匿名约束是面向一个特定的敏感值的，给定一个待发布数据表T，经过匿名分组后得到待发布数据表T’，对匿名表中任意一等价类分组G，给定一个敏感属性值v，（G，v）为等价类分组G中包含敏感属性值v的元组集合，如果v在等价类分组G中出现的频率都不大于a，即，则称敏感属性值v满足简单（a，k）匿名模型。一般（a，k）-匿名模型是将简单（a，k）-匿名模型从单个敏感属性值的约束扩展到对所有敏感属性值的约束，不仅仅限制单个的敏感属性值在任意等价类中出现的频率小于a，而是限定所有的敏感属性取值在任意等价类分组中出现的频率均小于a。

完全（a，k）-匿名模型是在简单（a，k）-匿名模型和一般（a，k）-匿名模型的基础上进行的推广。完全（a，k）-匿名模型针对待发布数据表T中的每一敏感属性v设定相应的频率约束av（0av1），要求得到的发布数据表T’中任意等价类分组中个敏感属性值v均满足（av，k）-匿名模型的约束。这里根据实际情况，若敏感属性值v的敏感属性越强，为了增大对其隐私保护力度，则相应的av就应该越小；敏感属性值的敏感度越弱，则对应的av就越大。如在医疗数据发布表中，敏感属性“疾病（Disease）”的取值“HIV”与“Flu”，明显“HIV”的敏感度强于“Flu”的敏感度，则对应的“HIV”的频率约束值aHIV应该小于“Flu”的频率约束值aFlu。例如表3-3-1的匿名数据发布表满足完全（a，k）-匿名模型，其中aHIV=0.4，aFlu=0.6，aPneumonia=0.4，aGastritis=0.4。

表 3-3-1 完全（av，3）-匿名表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tuple ID | QIs | Zipcode | Physician | Disease | Group ID |
| t1 | ··· | 8210\*\* | John | Flu | G1 |
| t5 | ··· | 8210\*\* | Bob | Pneumonia |
| t6 | ··· | 8210\*\* | Anne | Gastritis |
| t8 | ··· | 821\*\*\* | Hugo | HIV | G2 |
| t4 | ··· | 821\*\*\* | Bob | Flu |
| t7 | ··· | 821\*\*\* | Anne | Gastritis |
| t9 | ··· | 821\*\*\* | Marry | Flu |

在完全（a，k）-匿名模型中关于as的设定原则是，设定值as应该不小于敏感值s在原始待发布数据表中的出现的频率，否则难以生成满足完全（a，k）-匿名约束的发布匿名表。设待发布的数据表为T，|T|为数据表中元组个数，G为发布数据表中的等价类分组，S为敏感属性，vs为敏感属性的取值， avs为敏感属性vs的频率约束，则avs应该满足如下关系式：

1. 基于最小选择度优先的多敏感属性分组算法

杨静等人[9]在研究多敏感属性的隐私保护问题的时候，在传统单敏感属性L-多样性的基础上，利用拓扑空间中覆盖的思想定义了多敏感属性的L-多样性原则，引入了基于值域等级划分的个性化隐私保护方案，针对多敏感属性隐私保护提出了一种基于最小选择度优先的分组算法，在满足多敏感属性L-多样性原则的同时，实现敏感属性的个性化隐私保护需求。

该算法首先将敏感属性进行值域划分。将给定的敏感属性S，按照敏感属性S中的不同敏感属性取值的敏感度由高到低进行排序，然后对敏感属性取值划分成m个等级，记为CG（S）={LS1，LS2，LS3，···，LSm}。若CG（S）满足以下关系：=Dom（S）且SiSj=（1ijm），则就称CG（S）为S的一个值域等级划分。且SDegree（LSi）表示敏感属性S的在等级LSi的所取敏感度。所有敏感属性的值域等级划分完成后，发布数据表中每条个体记录的每个敏感属性都存在且只存在一个值域等级中。每条个体记录的记录敏感度为每个敏感属性的敏感度之和，即：

TDegree（t）=（t[Si]）

最小选择度优先的分组算法是根据个体记录的选择度执行的，在待发布数据表中，每条个体记录的选择度Select（tj）为tj中每一个敏感属性值v在待发布数据表T中出现的频率之和，考虑到个性化需求，所以单个个体记录的个性化选择度为PSelect（tj）表示为：

Select（tj）(tj)

其中为敏感属性在待发布数据表中出现的频率。（tj）是待发布数据表中个体记录tj中相异敏感属性值的集合。

最小选择度优先算法是一种启发式方法，基本策略是首先选择个性化选择度最小的元组作为等价类分组的初始元组，然后将具有不同敏感属性值的其他元组加入到当前等价类分组中，如果该等价类元组的元组数L，则将该等价类分组并入到待发布分组中，否则并入到待处理元组集中。循环上述步骤直至处理完待发布数据集中所有元组，从而得到最后满足多敏感属性隐私保护策略的发布数据所有分组。

完全（a，k）-匿名模型与基于最小选择度的数据发布算法都是针对多敏感属性数据发布中个性化发布的方法，完全（a，k）-匿名模型中虽然考虑了单个敏感属性的敏感度取值，但是对于含有高敏感度取值的元组没有进行个性化分组，可能造成含有高敏感度的元组（通常是比较总要的元组）被隐匿，造成整体数据集的可用性降低。基于最小选择度优先的启发式算法，结合了敏感属性值的敏感度和待发布数据的个体记录进行选择度量从而制定个性化分组，比较好的保留了选择度低的元组（敏感度高的元组），但算法只考虑了敏感属性值的敏感度而没有考虑敏感属性本身的敏感度问题。另外，最小选择度算法首先保留敏感度高的元组划入分组方法，有可能造成敏感度高的元组划分到同一等价类分组中，造成隐私属性倾斜，容易受到同质攻击。

3.4 本章小结

本章主要是针对多敏感属性数据发布中存在的问题进行了细致的分析，并从常规的多敏感属性数据发布和个性化多敏感数据发布两个方面展开讨论，讨论了常规多敏感数据发布中基于多维桶和L-覆盖性两种算法，个性化发布中的基于完全（a，k）-匿名模型的算法和基于最小选择度优先的数据发布方法，并分析各种算法解决的主要问题并分析这些算法在数据发布中存在的不足之处，引出下章本文提出的面向多敏感数据的发布算法以及个性化发布模型。

1. **面向多维敏感属性的数据发布**

4.1 多敏感属性隐私数据发布问题

敏感数据发布与共享环境中的个体隐私信息的安全问题一直是数据隐私研究的热点。在关系型数据发布中，Sweeney等人最早提出的k-匿名模型可保护但敏感属性下的隐私数据发布不受链接攻击[12,13]。文献[5]在分析k-匿名模型在某些情况下并不能保证隐私信息的安全。例如，在对数据表中准标识属性值概化后，具有相同准标识属性取值的大部分或者左右敏感属性的取值相同，那么攻击者只要确定个体记录属于哪个分组就能高概率地推断出个体的隐私信息，因此文献[5]提出了L-多样性概念，对匿名化数据表中，单个等价类分组中出现频率最高的敏感属性值的个数要求不大于1/L。

对于多敏感属性数据发布的研究，杨晓春等人提出的基于多维桶分组技术的方法对待发布数据表进行分组，是得到的分组在复合敏感属性的前提下满足L-多样性模型，达到隐私保护的要求。该方法较好地实现了对多敏感属性数据发布的隐私保护，但该分组方法效率较低，往往会因为选取桶的顺序问题造成分组效果不太理想，数据隐匿率较高，降低了发布数据的可用性。文献[6]提出的基于L-覆盖性分组算法，虽然解决了数据隐匿率等问题，但得到的等价类分组并不一定满足复合敏感属性L-多样性，特别是在剩余记录存在大量相似敏感属性的时候，由于L-覆盖性算法在分组完成后为了将剩余记录全部加入已分组当中，会造成部分分组中敏感属性相同的记录出现的概率大与1/L，存在隐私保护力度不足的问题。

因此，为了解决多敏感属性分组中分组效率不高与隐私保护力不足造成隐私泄露的问题，本章提出一种新的基于类二部图匹配的分组算法，对具有多维敏感属性的原始数据进行分组，使得各分组满足L-多样性，且降低数据隐匿率。

4.1.1 发布数据表中的属性定义

根据现有匿名研究[10,11]，待发布数据表T中的属性可以分为以下四类：

（1）标识符（Identifier，ID），能唯一标识数据表中个体记录或者机构具体身份的属性，例如身份证号码，手机号码，社会保险号等。

（2）准标识符（Quasi-Identifier，QID），数据表中联合起来能够标识个体记录的属性，如性别，出生日期，邮政编码，年龄等。

（3）敏感属性（Sensitive Attribute，SA），数据表中包含个体敏感信息的属性，如薪资，患病记录，职业和位置等。

（4） 非敏感属性（Non-Sensitive Attribute，NSA）数据表中除了以上三种属性之外的属性。

4.1.2 有损连接发布

Anatomy等在基于有损分解发布数据表的方式不需要对发布数据表进行泛化匿名操作，可以在保留原始数据表属性取值的前提下，保证个体记录的隐私信息不被泄露，通过分解准标识属性（QID）和隐私属性（SA）之间的对应关系，有损连接发布方式可以在直接发布原始数据表中的原始准标识属性值和敏感属性值的情况下保证发布数据满足L-多样性原则，保护用户隐私。

表 4-1-1 原始表

|  |  |  |
| --- | --- | --- |
| Tuple ID | QIDs | Sensitive Attribute |
| t1 | qid1 | sa1 |
| t2 | qid2 | sa2 |
| t3 | qid3 | sa3 |
| ··· | ·· | ··· |

以上我们有待发布数据表 4-1-1，假设元组{t1，t2，t3}为一个等价类分组，我们可以看到元组中的准标识属性值{qid1，qid2，qid3}和敏感属性{sa1­，sa2，sa3}是一一对应关系，经过有损分解，将待发布表分为准标识属性表和敏感属性表分成连个表发布，两个表之间仅通过分组编号相连，如表4-1-2所示。现在原有的一对一关系被分割成通过Group ID维持的一对多关系，攻击者无法保证通过准标识符唯一确定其敏感属性，若敏感属性Sensitive Attribute满足L-多样性，则真个数据发布表满足L-多样性原则。

表 4-1-2 有损分解发布表

|  |  |  |
| --- | --- | --- |
| Tuple ID | QIDs | Group ID |
| t1 | qid1 | G1 |
| t2 | qid2 |
| t3 | qid3 |
| ··· | ·· | ··· |

|  |  |
| --- | --- |
| Group ID | Sensitive Attribute |
| G1 | sa1 |
| sa2 |
| sa3 |
| ··· | ··· |

4.2 基于类二部图边选择的多敏感数据分组--BES

设待发布关系型数据表T={A1，A2，A3，···，Ap ，S1 S2 S3，···，Sd}，其中Ai{A1，A2，···，Ap}（1ip）为准标识属性，Sj{S1，S2，···，Sd}（1jd）为敏感属性。待发布数据表T中共有n条记录，即|T|=n，数据表中每条个体记录称为一个元组，元组标识为ti（1in）。令t[X]标识数据表中元组t在X属性上的取值。

定义4.1 分组[5]。将待发布数据表T中所有元组分为若干组，记为GT，GT={G1，G2，···，Gm}，并且 =T，并且=（1ijm）。则称GT为待发布数据T的分组。

定义4.2 复合敏感属性[2]。待发部数据表中的所有敏感属性组成复合敏感属性S={S1，S2···，Sd}，S为复合敏感属性集合，则|S|2。SiS（1id）标识数据表中的第i个敏感属性，Dom（Si）表示敏感属性Si的值域，|Si|标识Dom（Si）的基数，即Si所有可能取值的个数。

定义4.3 单敏感属性L-多样性。对于单敏感属性下得到的一个分组G，设SSet（G）为分组G中所有不重复的敏感属性取值。vi（i）为分组G中某一敏感值，viSSet(G), 标识vi出现的次数，若分组G中满足： ，则称该分组满足单敏感属性L-多样性。

定义 4.4 多敏感属性L-多样性。若一个包含多敏感属性分组G，如果其中每个元组的每一条记录上的敏感属性的取值都满足单敏感属性L-多样性，则称分组G满足多敏感属性L-多样性。

根据以上定义，针对一般多敏感属性隐私数据发布的问题主要是将待发布数据表T进行分组得到发布表T’，T’中所有分组是满足隐私保护要求的，本文要求是得到的分组均满足多敏感属性L-多样性，通过有损分解发布数据表T’,以下提出一种新的基于二部图边选择的分组算法（Bigraph-similar Edges Selection），得到多敏感属性L-多样性分组，发布数据表满足多敏感属性L-多样性模型。

4.2.1 BES基本思想

本章提出的基于类二部图的边选择的分组技术（BES）-是为解决多敏感隐私属性数据发布中的分组问题，目的是找到多敏感属性待发布数据表T上的分组方案，得到分组GT，使得每个分组均满足多敏感属性L-多样性。因此，BES分组方法首先需要将多敏感属性数据表T中所有元组映射到二部图上。

定义4.5 二部图。无向图BG = <V,E>的结点集V能够划分为两个子集V1,V2，满足V1∩V2 =，且V1∪V2 = V（全集），使得BG中任意一条边的两个端点，一个属于V1，另一个属于V2，则称G为二部图或二分图（Bigraph）。

针对待发布医疗数据表3-1-1中，共有两个敏感属性Physician与Disease，取值集合分别为：

SPhysician={John，Bob，Anne，Hugo，Marry}

SDisease={Flu，Pneumonia，Cancer，Gastritis，HIV}

其中SPhysician∩SDisease =,则我们可以将每一维敏感属性作为图的一个点集，每个敏感属性的取值为图上一点，根据表3-1-1的各元组可得到如下图的二部图：

图 4-2-1 表3-1-1数据表敏感属映射的类二部图

图中每条边即代表从待发布数据表T中的一个元组中的敏感属性取值，如若待发布数据表中的敏感属性超过2个，例如在表3-1-1的中添加第三个敏感属性Marital-status，敏感属性取值SMarital-status={Divorced，Separated，Never-married，Married}。则可根据表数据构造出图4-2-1类二部图（Bigraph-similar）：

图 4-2-2 3个敏感属性映射的类二部图

定义4.6 元组边（t.E）。发布数据表T中ti(1in),tiT。由ti映射到类二部图的所有边称为该元组边的元组边，记为ti.E。V(ti.E)标识该元组边经过的所有点的集合。所有元组边构成的集合称为类二部图的元组边集TE=。例如图4-2-1中，元组t1的元组边为t1.E={(John,Flu)},t1元组边经过的点集V（t1.E）={John，Flu}。

根据待发布数据表T映射得到的类二部图表示为BG，BG=<VS,TE>，其中VS={VS1, VS2, ···, VSd}，其中VSi（1id）表示敏感属性Si的所有属性值映射到类二部图中的点的集合。例如图4-2-1中，敏感属性“Disease”对应的点集合VDisease={ Flu，Pneumonia，Cancer，Gastritis，HIV}。在得到所有元组映射的类二部图后，基于边选择的方法采用某种策略尽可能选择多的元组边作为一个分组，且这些元组边的敏感属性点集Vti没有交集。如上文所述，BES方法在发布数据时采用分组内有损连接的方式发布数据，因此在不破坏L-多样性的前提下，分组越小造成的数据利用度越高，信息损失也越小。理想情况下若得到的分组大小为L，并满足多敏感属性L-多样性，就要求每个分组中的每条个体记录的敏感属性值值出现一次。处理剩余记录时，在满足多敏感属性L-多样性的前提下，可能某些分组的记录条数会超过L，会造成附加的有损连接信息损失。这里采用文献[2]定义的附加信息损失度=，其中m为得到的分组数。

定义4.7 不相交边选择。在类二部图BG=<VS,TE>中，在为当前分组G选取一条元组边t.E时，若该元组边与已加入当前分组的所有元组边均不相交,即

，则称为不相交边选择。

基于类二部图边选择的分组方法是一种以固定分组大小，采用贪心策略在类二部图上依次对元组边作不相交边选择，选取L个元组边对应的元组构成分组，重复进行，尽可能多地得到大小为L的分组，最后在不破坏多敏感属性L-多样性的前提下，将剩余元组加入到已有分组中。最后将不包含在任何分组的个体记录从发布的数据中隐匿掉。采用数据隐匿率（Suppress ratio）来衡量分组后隐匿的数据记录占发布数据表中的比例。数据隐匿率定义，Ns为隐匿的个体记录数，由上可知道SuppRatio越小，隐私数据越少，理想情况下，SuppRatio=0。本文将数据隐匿率和附加信息损失度一起作为发布数据的衡量标准。

4.2.2算法描述

BES分组算法主要分为3个步骤：

（1）将待发布数据表的多个敏感属性提取出来，构建类二部图并得到所有的元组边。

（2）分组阶段在已有的类二部图中依次做不相交边选择，得到每一个敏感属性取值都互不相同的L个元组边构成一个有效分组。循环进行，直到剩余的元组边中无法再做不相交边选择操作。

（3）剩余记录处理阶段。将第（2）步中剩余的元组依次遍历，在不破坏原有分组满足多敏感属性L-多样性的前提下将记录添加到分组中，最终将不属于任何分组的个体记录隐匿。

算法. 基于类二部图的边选择分组算法

输入：待发布数据表T{A1，A2，A3，···，Ap，S1，S2，···，Sd}，多样性参数L

输出：准标识属性表QIT，敏感属性表ST

步骤：

1. 提取T中敏感属性值，构建类二部图BG，得到所有记录的元组边TE
2. while TE中的元组边不为空
3. 遍历TE-> t[i···n].E
4. if ti.E能在当前分组G上作不相交边选择
5. 将ti将入到当前分组G
6. if 当前分组G中元组数等于L
7. 将当前分组G加入到已完成分组集合GS中，并将G中所有

元组对应的元组边从TE中移除

1. 停止遍历，重新进入到while循环
2. if 当前分组G中元组数不足L
3. 将分组G中所有元组对应元组放入待处理元组集合RT，并将其

对应的元组边从TE中移除

1. 遍历RT->t[i···|RT|]
2. for G[j…|GS|]
3. if ti加入到GSj中GSj依然满足多敏感属性L-多样性
4. 元组ti加入到GSj中，并将ti从RT中移除
5. 隐匿RT中所有剩余的元组
6. 将GS中所有分组以QIT，ST形式发布

算法的第1步是预处理阶段，得到后面步骤所需要的元组边；第2步到底10步是分组阶段，整个过程只有一个while循环；得到所有分组后，算法第11步到14步是对RT中的剩余元组的处理，看是否能加入到已有分组中且不影响原有分组的多敏感属性L多样性，从而减小元组的隐匿数量。算法第13/14步是隐匿数据和数据发布阶段。

4.2.3 BES算法实例应用

我们依然以表3-1-1作为待发布数据表为例，通过BES分组算法进行分组。假设多样性参数L=3。首先得到的类二部图如图4-2-1，再的到的元组边为TE={t1.E{(John，Flu)}，t2.E{(John，pneumonia)，t3.E{(John，Cancer)}，t4.E{(Bob，Flu)}，t5.E{(Bob，Pneumonia)}，t6.E{(Anne，Gastritis)}，t7.E{( Anne，Gastritis)}，t8.E{(Hugo，HIV)}，t9.E{(Marry，Flu)}}，为能够进行不相交边操作，我们需要得到元组边的边点集V（t.E）,根据定义4.7我们可以得到V（t1.E）={John，Flu}，V（t2.E）={John，Pneumonia}，V（t3.E）={John，Cancer}，V（t4.E）={Bob，Flu}，V（t5.E）={Bob，Pneumonia}，V（t6.E）={Anne，Gastritis}，V（t7.E）={ Anne，Gastritis}，V（t8.E）={Hugo，HIV}，V（t9.E）={Marry，Flu}。遍历TE，首先选取ti到当前分组G中，依次遍历找到t5，因为，能够进行不相交边选择操作，所以将t5加入到当前分组G中，继续遍历TE得到t6有且，满足不相交边选择条件，所以讲t6将入到当前分组G，得到一个满足L=3条件的有效分组{t1，t5，t6}，并将它们对应的元组边从TE中移除。重复以上过程继而得到{t2，t4，t7}，{t3，t8，t9}。

（这里怎么搞啊，要改表数据还是增加两条数据？感觉增加数据要好点T\_T，明天来增加两条数据吧。。。）

4.3 实验结果及分析

4.3.1 实验数据

4.3.2 实验结果对比分析

4.4 本章小结

1. **面向多敏感属性的个性化发布模型**

目前关于多敏感属性个性化L-多样性模型中，针对敏感属性值的权值定义或者登记划分，都是将记录中的敏感属性值拆分成单个的敏感属性取值，然后在作处理，从而实现单个的高敏感度敏感属性值的个性化保护。或者直接将个体记录作为一个整体，对每个个体记录指定个性化发布约束，虽然个性定制对隐私信息的保护较好，但是效率低，可行性不高，且都没有考虑到敏感属性本身的敏感度问题（例如疾病的敏感度大于收入的敏感度）。本文拟在考虑单个敏感属性值的敏感度的同时也考虑个体记录的整体敏感度，由于个体记录的整体敏感度由组成该个体记录的所有敏感属性值的敏感度决定，本文尝试结合单个敏感属性的个性化约束与个体整体记录的个性化约束结合的方式，整体个体记录的约束性由个体的每一项隐私属性决定，最后分组时考虑个体记录的敏感度，而数据发布时仅仅需要为单个敏感属性值指定个性化约束，从而达到的隐私保护的目的。

5.1 多维敏感属性数据的个性化发布

5.2 多敏感属性(L，a)-diversity 个性化数据发布模型

5.3 最大权边类二部图边选择分组算法

5.3.1 算法基本思想

5.3.2 算法描述

5.3.3 算法实例应用

5.4 L-拆分类二部图边选择分组算法

5.3.1 算法基本思想

5.3.2 算法描述

5.4 实验结果对比及分析

5.5 本章小结

1. **总结与展望**

**参考文献**

1. Anatomy: Simple and Effective Privacy Preservation
2. 数据发布中面向多敏感属性的隐私保护方法，杨晓春等人
3. Machanavajjhala A,Kifer D,Gehrke J. et al. l-diversity: privacy beyond k-anonymity[C]// Proceedings of the 22nd International Conference on Data Engineering(ICDE’06). New York: ACM,2006:24-35
4. 孙岚, 郭旭东, 王一蕾, 吴英杰. 个性化隐私保护轨迹发布算法[J]. 系统工程与电子技术.2014,36(12):2550-2555
5. Machanavajjhala A，Gehrke J，and Kefer D．／-diversity：Privacy beyond k-anonymity//Proceedings of the 22nd International Conference on Data Engineering．Atlanta，Georgia,
6. 面向多敏感属性医疗数据发布的隐私保护技术 – 金华 刘善成等人
7. Xiao Xiaokui,Tao Yufei. Personalized Privacy Preservation
8. 面向敏感值的个性化隐私保护，韩建明等人
9. 一种基于最小选择度优先的多敏感属性个性化l-多样性算法。杨静等人
10. 龚奇源，杨明，罗军舟．面向关系一事务数据的数据匿名方法
11. Qiyuan Gong，Ming Yang，Zhenguo Chen，Junzhou Luo．Utility Enhanced Anonymization for Incomplete Microdata
12. Sweeney L．K—anonymity：A model for protecting privacy．International Journal on Uncertainty，Fuzziness，and Knowledge-Based Systems，2002，10(5)：557—570
13. Samarati P，Sweeney L．Generalizing data to provide anonymity when disclosing information