# 第5章 隐私保护的数据发布

**学习要求**：了解交互式和非交互式的数据发布框架；掌握显示标识属性、准标识属性的概念；掌握 $k$-匿名、$l$-多样化的概念；理解数据脱敏和数据溯源的概念及有关的实现方法；掌握保留格式加密的定义、基本构造方法，了解保留格式加密的基本模型，了解基于保留格式加密的数据库水印的应用。

**课时**：2 课时

**建议授课进度**：5.1 节～5.2 节用 1 课时，5.3 节～5.4 节用 1 课时

## 5.1 基本概念

人们正生活在一个大数据的时代，越来越多的设备和传感器通过数字网络相连，数据收集者们通过其中的应用程序大量收集个人数据，并将其提供给有需求的数据分析者。分析者可以利用各种工具对获取的数据进行挖掘，以此产生能够支持商业计划、政府决策、科学研究、广告投放等应用的策略，实现商业利益和科研价值，最终使大众受益。

### 5.1.1 发布框架

在隐私保护数据发布领域中，数据发布者从数据拥有者采集到应用中的数据，例如，医疗数据、金融数据、电信数据、访问数据、社会调查数据等。然后，将数据发送给数据接收者。这个模式中包括将数据公布于众，或者将数据发送给申请的单位、机构或者个人等，使数据用于科学研究或者支持决策，服务于公众，如图 5-1 所示。

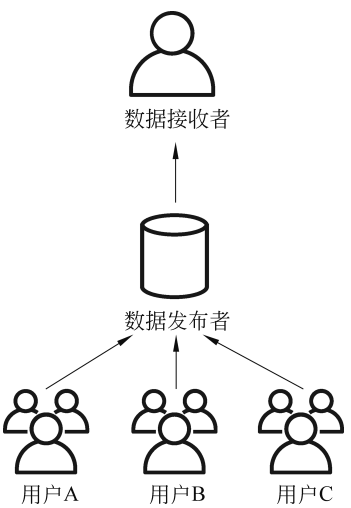

图 5-1 数据发布示意图

在数据发布应用的第一个阶段数据收集，假设是诚实的模型，数据所有者将数据发送给诚实的数据发布者。然而，在第二个阶段，数据发布阶段是非诚实模型，数据接收者是不诚实的，数据接收者可能是一个攻击者。例如，某医药公司获得一份某医院的电子医疗信息，但是无法保证所有的员工都是诚实的。会有人员通过发布的数据获取其中的敏感信息，称为攻击者。攻击者设法获取的敏感信

息所对应的个体，称为攻击对象。

隐私保护的数据发布技术（privacy preserving in data publishing，PPDP）是数据发布者将原始数据表进行匿名化操作，然后再对它进行发布，以保护数据中的敏感信息，避免隐私泄露。

数据发布流程框架主要分为两种，即交互式和非交互式数据发布框架。

如图 5-2 所示，交互式数据发布通常表现为数据的在线查询发布，较多出现在政府机关和研究机构的对外数据发布中，供有兴趣的用户查询。例如，美国的联邦经济数据研究网站，能够提供一系列经济数据在不同时间周期内的聚合查询和批量查询。

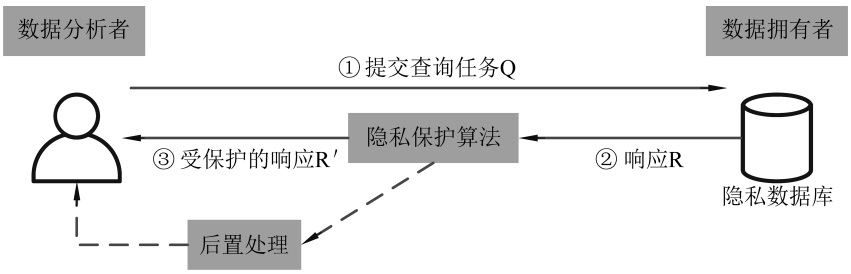


图 5-2　交互式数据发布框架

如图 5-3 所示，非交互式数据发布通常表现为离线发布，例如，数据挖掘竞赛发布的公开测试集，交通管理局发布的周期性的路况信息等。数据拥有者先通过隐私保护算法对需要发布的数据集进行完整的匿名处理，然后数据分析者根据已发布的数据集进行各种需要的查询。在非交互式数据发布中，由于数据拥有者并不知道数据分析者会对匿名数据集进行何种查询，因此，设计隐私保护算法需要同时满足隐私性以及较高的可用性。

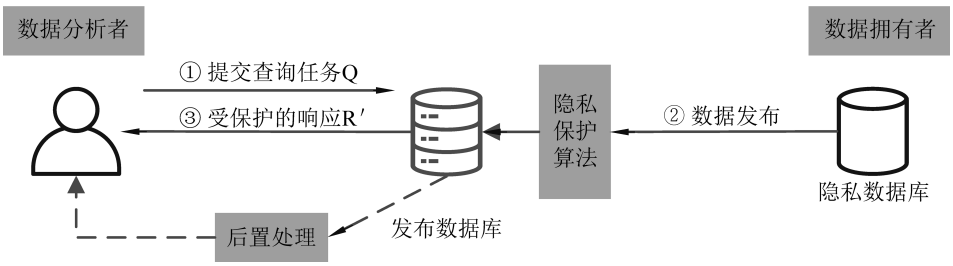


图 5-3　非交互式数据发布框架

## 5.1.2　属性分类

假设原始数据是经过预处理的结构化数据，在 PPDP 最基本的格式中，数据发布者有一个格式表：D（显式标识属性，准标识属性，非敏感属性，敏感属性）。

（1）**显式标识属性（identifier attribute）**：也称显式标识符或标识符，是能唯一标识单一个体的属性，如姓名、身份证号码。

（2）**准标识属性（quasi-identifier attribute，QI）**：是组合起来能唯一标识单一个体的属性，如性别和年龄的组合等。

（3）**等价类**：准标识属性完全相同的多条记录，称为一个等价类。

（4）**敏感属性（sensitive attribute）**：包含敏感数据的属性，尤其是涉及个体隐私的细节信息，如疾病、病人患病记录、个人薪资、地理位置等。

数据的发布者不能把原始数据直接发布，要避免数据接收者把数据表中的敏感属性与个体链接起来。敏感属性包含个体隐私的信息，是数据接收者进行数据挖掘、数据分析的对象，不能被移除。

### 5.1.3　背景知识

数据发布隐私保护需要关注的一个重要问题就是攻击者可能拥有的各种背景知识，这些知识可以包括外部数据、常用知识、有关匿名算法的知识和过去发布的数据，这些信息可以通过关联已发布的数据集来推测匿名数据集中的个人敏感属性。

(1) 外部数据。主要包括公开可获得的数据，如选民登记记录，电影评分统计等；攻击者容易获得的关联数据，如目标用户隔壁邻居的年龄和地址等。这些外部数据可能包含除原始数据中敏感属性外的所有类型的信息。通过这些从外部数据获得的额外信息，攻击者可以在匿名数据中推敲目标个体存在的元组，并进一步发现目标个体的敏感值。

(2) 常用知识。这是关于目标个体敏感信息分布的额外信息，可以从许多来源获得。例如，攻击者可能有一个常识：冬天很容易感冒，或者对手可能从他的同事那里听说另一位同事的工资超过一万元。如果目标个体可能患有某些疾病或其工资数目在某一个固定的范围内，那么攻击者就可以利用这些非关联的常识信息排除匿名数据集中的一些个体，从而以更高的概率推断出目标个体。

(3) 基于隐私保护算法的知识。攻击者可能知道当前匿名数据集所使用匿名算法的机制，因为生成匿名数据的算法很可能会在数据发布时公布。在某些情况下，这些算法本身就可能披露敏感信息。

(4) 过时数据。在数据发布的场景中，有些需要数据拥有者在固定时间周期内进行多次发布，以确保数据集的实时性。那么这种方式下攻击者可以获得所有先前发布的数据，并使用这些数据来排除目标个体的可能候选元组或敏感属性值。

### 5.1.4　相关攻击

很显然，攻击者有了背景知识，如果发布数据表仅仅简单移除了显式标识属性是不够的，隐私信息仍然有可能被准标识属性联合起来定位获得。

Sweeny 等[①]在 2002 年说明了美国公众可以从公开的选民数据集获取姓名、社会保障号、年龄、邮政编码这些人口统计信息。这将导致 87% 的美国人遭受“链接攻击(link attack)”。这意味着他们能够被准标识属性联合起来唯一确定。

如图 5-4 所示，公开数据集中包含姓名、家庭住址、政治面貌、注册日期、出生日期、性别、邮政编码。公众可以获得数据集所含个体的这 8 个属性信息。另外一张表，是医院的医疗记录，它仅仅从原始的医疗记录中移除了显式标识属性，公开了诊断结果、就诊日期、处方、出生日期、性别、邮政编码等属性。由于人们可以从公开数据集中获取与医疗记录相重叠的属性出生日期、性别、邮政编码，从而可以唯一确定个体的敏感属性，造成隐私的泄露，即诊断结果、就诊日期等。

① LATANYA S. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05,2002: 571-588.

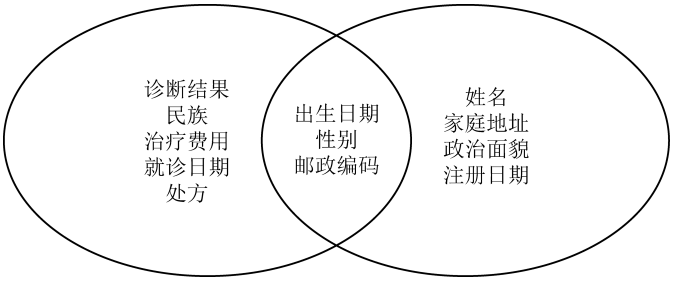


图 5-4　链接攻击示例

总之，如果攻击者有包含背景知识的数据，包含了个体的准标识属性值，通过连接这两张表，能推断出一些敏感属性值，可以细分为以下 3 种类型的攻击。

（1）**记录链接**：当攻击者能够将记录的所有者与发布的数据表中的相应的记录相对应时，称为记录链接。例如，通过准标识符确定一条记录的所有者身份。如图 5-5 所示，Doug 可以通过准标识符<Job，Sex，Age>唯一确定。

（2）**属性链接**：当攻击者能够将记录的所有者与发布的数据表中的敏感属性相对应时，称为属性链接。如图 5-5 所示，Emily 和 Glady 可以通过准标识符<Job，Sex，Age>确定得了 HIV，即泄露她们得了 HIV。

（3）**表格链接**：当攻击者能够将记录的所有者与发布的数据表本身相对应时，称为表格链接。如果能确定所有者出现在了某表中，如疾病表，会泄露该所有者存在疾病这一信息。

| Job | Sex | Age | Disease |
|---|---|---|---|
| Engineer | Male | 35 | Hepatitis |
| Engineer | Male | 38 | Hepatitis |
| Lawyer | Male | 38 | HIV |
| Writer | Female | 30 | Flu |
| Writer | Female | 30 | HIV |
| Dancer | Female | 30 | HIV |
| Dancer | Female | 30 | HIV |

(a)

| Name | Job | Sex | Age |
|---|---|---|---|
| Alice | Writer | Female | 30 |
| Bob | Engineer | Male | 35 |
| Cathy | Writer | Female | 30 |
| Doug | Lawyer | Male | 38 |
| Emily | Dancer | Female | 30 |
| Fred | Engineer | Male | 38 |
| Gladys | Dancer | Female | 30 |
| Henry | Lawyer | Male | 39 |
| Irene | Dancer | Female | 32 |

(b)

图 5-5　删除标识符的数据发布示意

(a)病人数据表；(b)扩展数据表

目前，已有的各种隐私保护方法都是为降低某些隐私泄露危险、抵御攻击者的攻击模型而产生的，在数据的发布过程中，数据集可能遭受来自攻击者的隐私威胁，除了链接攻击之外，还有同质性攻击(homogeneous attack)、敏感性攻击(sensitivity attack)、概率攻击(probability attack)等。

### 5.1.5　匿名化方法

为了完成数据表的隐私保护的安全发布，需要对其数据进行匿名化操作，常用的方法有泛化、抑制、解剖、扰动等。

(1) 泛化是用一个更加泛化的值代替具体的值。对于分类型属性,泛化是用父类级别代替子类级别。对于数值型属性,泛化是用数值所在的区间代替具体的数值。

(2) 抑制是抑制某个数据项,不发布这个数据项。对于分类型属性,抑制是泛化到分类树的根节点这种特殊的情况;对于数值型属性,抑制是泛化到属性值域这个最大的区间的特殊情况。泛化的逆操作称为细化,抑制的逆操作称为公开。

(3) 解剖是指不修改原始数据表中的准标识属性或者敏感属性,而是将数据表分割成两张表发布,一张是准标识属性表,一张是敏感属性表。这两张表中的数据通过等价类的标号链接,两张表中属于同一个等价类的记录具有相同的等价类标号。同一个等价类的敏感属性值如果相同,那么在敏感属性表中只出现一次,也就是敏感属性表中属于同一等价类的数值都是不同的。因而,同一个等价类中的记录链接到类内的敏感属性值的概率是相等的。

(4) 扰动是防止统计泄露中的一种针对数据的操作。它是保持数据的一些统计性质不变的前提下,对数据进行添加噪声,数据交换,或者人工数据合成操作。生成的数据已经不再是真实数据,它不会与真实的数据链接起来,从而保护数据的隐私信息。扰动对于数值型统计查询(如聚合查询)很有用,因为它可以保留原始数据的统计信息。而且基于差分隐私(differential privacy, DP)保护算法的扰动数据集能够达成最理想的隐私保护效果。但在非数值型数据集中,由于准标识符和敏感信息之间的关系失真太多,因此,数据挖掘算法从扰动数据中学习的知识模型可能精度较差。

## 5.2 *k*-匿名模型

本节以基本的 *k*-匿名模型为例,讲解数据发布过程中的攻防博弈。

### 5.2.1 *k*-匿名

如果仅仅是将显示标识属性删除,是不够的。如图 5-4 所示,攻击者很容易通过记录链接等攻击,推断出用户得的疾病情况。

***k*-匿名模型要求在所发布的数据表中,对于每条记录都至少存在其他 *k* − 1 条记录,使得它们在全体准标识属性上取值相等**,即这个模型要求每个等价类的记录不少于 *k*。

实现 *k*-匿名的方法就是泛化或者抑制。

如图 5-6 所示,对 Age 进行了泛化,用年龄段来代替年龄。这样,就得到了 4 个等价类,即<Engineer, Male,[35-40)>,<Lawyer, Male, [35-40)>,<Writer, Female, [30-35)>和<Dancer, Female, [30-35)>,分别满足了 2——匿名、1——匿名、2——匿名和 2——匿名。

注意,为了满足匿名模型,需要使等价类中记录的数量至少为 *k* 条,因此 *k* 越大,隐私保护越好,由此带来的数据损失也就越大。然而,这个匿名模型只针对准标识属性有约束,并没有约束敏感属性。

### 5.2.2 *l*-多样化

1) 同质性攻击

如果在一个等价类中全部敏感属性的取值相等,那么虽然攻击者不能确定哪条记录属

| Job | Sex | Age | Disease |
|---|---|---|---|
| Engineer | Male | [35-40) | Hepatitis |
| Engineer | Male | [35-40) | Hepatitis |
| Lawyer | Male | [35-40) | HIV |
| Writer | Female | [30-35) | Flu |
| Writer | Female | [30-35) | HIV |
| Dancer | Female | [30-35) | HIV |
| Dancer | Female | [30-35) | HIV |

(a)

| Name | Job | Sex | Age |
|---|---|---|---|
| Alice | Writer | Female | [35-40) |
| Bob | Engineer | Male | [35-40) |
| Cathy | Writer | Female | [30-35) |
| Doug | Lawyer | Male | [35-40) |
| Emily | Dancer | Female | [30-35) |
| Fred | Engineer | Male | [35-40) |
| Gladys | Dancer | Female | [30-35) |
| Henry | Lawyer | Male | [35-40) |
| Irene | Dancer | Female | [30-35) |

(b)

图 5-6　Age泛化后的结果

(a)病人数据表；(b)扩展数据表

于攻击对象，但是，能以100%的概率确定攻击对象的记录的敏感属性。因此，这个模型仅能够从一定程度上抵御记录链接，不能够抵御属性链接。同质性攻击是等价类中的敏感值都相等，而导致的属性链接。它是由于等价类中的敏感值缺少多样性而造成的。

在图5-6中，仅仅对Age进行泛化还不够，很显然，＜Engineer，Male，[35-40)＞的2个等价类具有相同的属性、＜Dancer，Female，[30-35)＞的2个等价类也具有相同的属性。

2）$l$-多样化匿名模型

如果数据表中的每个等价类有至少$l$个敏感属性值，那么称数据表是$l$-多样化的。

如图5-7所示，继续将Job进行泛化，用高级别的分类来代替，如用Artist来代替Dancer、Writer；用Professional代替Lawyer和Engineering。这样，就得到了2个等价类，分别为3-匿名和4-匿名，均为2-样化，就可以抵御同质性攻击。

| Job | Sex | Age | Disease |
|---|---|---|---|
| Professional | Male | [35-40) | Hepatitis |
| Professional | Male | [35-40) | Hepatitis |
| Professional | Male | [35-40) | HIV |
| Artist | Female | [30-35) | Flu |
| Artist | Female | [30-35) | HIV |
| Artist | Female | [30-35) | HIV |
| Artist | Female | [30-35) | HIV |

(a)

| Name | Job | Sex | Age |
|---|---|---|---|
| Alice | Artist | Female | [35-40) |
| Bob | Professional | Male | [35-40) |
| Cathy | Artist | Female | [30-35) |
| Doug | Professional | Male | [35-40) |
| Emily | Artist | Female | [30-35) |
| Fred | Professional | Male | [35-40) |
| Gladys | Artist | Female | [30-35) |
| Henry | Professional | Male | [35-40) |
| Irene | Artist | Female | [30-35) |

(b)

图 5-7　$l$-多样化示意

(a)病人数据表；(b)扩展数据表

虽然$l$-多样化和$k$-匿名模型在有关防止属性泄露方面上迈出了关键性的一步，但它不足以防止(敏感)属性泄露，因为它容易遭受倾斜攻击和相似性攻击。

以倾斜攻击为例，在满足多样化的一个匿名表中，如果某个敏感属性值在全局出现的频率很低，而在某个等价类中出现的频率远高于全局的频率，那么这个等价类中被攻击者链接为此敏感属性值的概率远高于全局的概率，这就是倾斜攻击。图 5-7 满足了 2-多样化匿名，HIV 在全局出现的概率是 50%，但是在第 2 个等价类中 HIV 出现的概率是 75%，因而使得第 2 个等价类中的记录更容易被链接到 HIV 这种疾病。

总之，当总体分布是偏态分布时，满足 $l$-多样化并不会阻止属性公开。

### 5.2.3　t-相近

$t$-相近模型是一个首次提出敏感属性值的分布的隐私保护方法，它考虑了等价类内敏感属性的分布，要求每个 $k$-匿名组中敏感属性值的统计分布与该属性在整个数据集中的总体分布“接近”。

一个等价类满足 $t$-相近模型，则等价类中敏感属性值的分布与在数据表的分布差异不超过 $t$。如果数据表的每个等价类都满足 $t$-相近，则称这个数据表满足 $t$-相近。

$t$-相近是基于 $l$-多样化组的匿名化的进一步细化，用于通过降低数据表示的粒度来保护数据集中的隐私。这种减少是一种折中，它会导致数据管理或挖掘算法的一些有效性损失，从而获得一些隐私。因为，满足这个模型的匿名表中，由于每个等价类与全局等价类的分布的差异不大(不超过阈值 $t$)，使得匿名表丢失了很多准标识属性与敏感属性之间的相关信息，这可能正是数据接收者进行数据挖掘和科学研究所需要的信息。

### 5.2.4　其他模型

数据发布的过程中，如何保护隐私和确保可用性，总是存在矛盾，而相关研究也是在这个矛盾中逐步前进。

传统的数据发布隐私保护技术通过删除能够唯一识别个体身份的信息(标识符属性)实现匿名发布，典型的解决办法就是 $k$-匿名模型。如前面所述，虽然 $k$-匿名隐私模型切断了个体与数据表中某条记录之间的联系，但是却没有切断个体与敏感信息之间的联系，因此 $l$-多样化模型、$t$-相近模型等相继提出。

1) 差分隐私模型

基于 $k$-匿名模型及其改进策略的匿名保护模型大都沿用了属性的泛化操作，对发布数据的可用性造成较大影响。同时，大数据发布环境下的组合攻击、前景知识攻击等新型攻击方式对 $k$-匿名模型及其改进方法提出了严峻挑战。Dwork 等提出的差分隐私模型借鉴了密码学中语义安全的概念，通过在发布数据或查询结果中添加随机噪声来达到隐私保护的效果。差分隐私模型允许攻击者拥有无穷的计算能力和任何有用的背景知识，而且不需要关心攻击者的具体攻击策略。在最坏的情况下，即使攻击者获得了除某一条记录之外的所有敏感数据，差分隐私模型仍然可以保证攻击者无法从查询输出结果判断该条记录是否在数据集内。由于具备严格的数学特性，差分隐私被认为是一种非常可靠的保护机制，得到了大量研究学者的关注。基于差分隐私模型的数据发布主要针对敏感数据的统计信息进行保护。

2) $m$-不变性模型

传统的静态数据集隐私保护方法无法直接应用于动态数据集重发布过程中，因此，

需要研究适用性较强且能够保护动态数据集隐私安全的数据匿名方法。$k$-匿名、$l$-多样化等模型都是面向静态数据集的隐私保护而提出的，无法保证动态数据集的隐私安全。动态数据集的隐私保护问题所面临的挑战是：隐私保护模型不仅要保护数据集的当前快照和以往发布的快照，而且在攻击者将所有发布数据集联合后也能保护数据集的隐私安全。针对动态数据集的重发布的隐私保护问题，$m$-不变性模型被提了出来。该模型要求数据拥有者每个周期发布的匿名数据表中，每个等价类都包含至少 $m$ 条记录，且他们的敏感值各不相同，且每条记录 $t$ 在其发布周期$[t_1, t_2]$($t_1 \leqslant t_2$)内的归属等价类具有相同的敏感属性值集合。

虽然 $m$-不变性模型能够维护数据重发布下的隐私安全，但该模型仅关注了数据集对记录的插入和删除两种操作，但在动态更新记录属性值时，$m$-不变性模型便无法较好地保持数据集的隐私安全；此外，$m$-不变性匿名模型还要考虑 $m$ 值选取的合理性问题，$m$ 值选取不当便会导致向数据集中添加假数据降低数据的可用性。

## 5.3 数据脱敏与溯源

### 5.3.1 数据脱敏

数据脱敏(data masking)是指对某些敏感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。在涉及客户安全数据或者一些商业性敏感数据的情况下，在不违反系统规则条件下，对真实数据进行改造并提供测试使用，如身份证号、手机号、卡号、客户号等个人信息都需要进行数据脱敏。

1989 年，Adam 等[①]就提出数据脱敏的概念，脱敏的方法有替换、遮蔽、加密等，比如，将手机号部分数字通过用 * 号替换实现脱敏等。5.2 节讲述的一些匿名化方法也可以用来脱敏。一些数据脱敏的方法示例如表 5-1 所示。

**表 5-1 数据脱敏方法示例**

| 名称 | 描　述 | 示　例 |
|---|---|---|
| 掩码 | 利用“ * ”等符号遮掩部分信息，并且保证数据长度不变，容易识别出原来的信息格式，常用于身份证号、手机号等 | 12300001234→<br>123 * * * * 1234 |
| 替换 | 一般会有一个字典表，通过查表进行替换 | 张三→X 李四→Y |
| 混合掩码 | 将相关的列作为一个组进行屏蔽，以保证这些相关列中被屏蔽的数据保持同样的关系，例如，城市、省、邮编在屏蔽后保持一致 | |
| 截断 | 舍弃某些必要信息保证数据的模糊性 | 13800001234→13800 |
| 加密 | 利用加密算法对数据进行变化 | 13800001→IQ5XRW== |

数据脱敏按模式可以分成静态数据脱敏和动态数据脱敏。其主要区别在于是否对敏感数据信息采取实时的脱敏操作。

① ADAM N, WORTHMANN J C. Security-control methods for statistical databases: a comparative study[J]. ACM Computing Surveys (CSUR) 21.4 (1989): 515-556.

（1）静态数据脱敏是数据存储时脱敏，存储的是脱敏数据。一般用在非生产环境，如开发、测试、外包和数据分析等环境。

（2）动态数据脱敏在数据使用时脱敏，存储的是明文数据。一般用在生产环境，动态脱敏可以实现不同用户拥有不同的脱敏策略。

### 5.3.2　数据溯源

数据溯源是数据发布后流转过程中发生泄密后的回溯泄密节点的操作。如图 5-8 所示，数据溯源通常通过向数据中加入水印，在数据泄密后，通过提取数据中的水印来完成泄密节点的溯源。很显然，实现数据溯源的关键就是水印不能被攻击者检查出来或者破坏掉，也就是水印的鲁棒性要好。

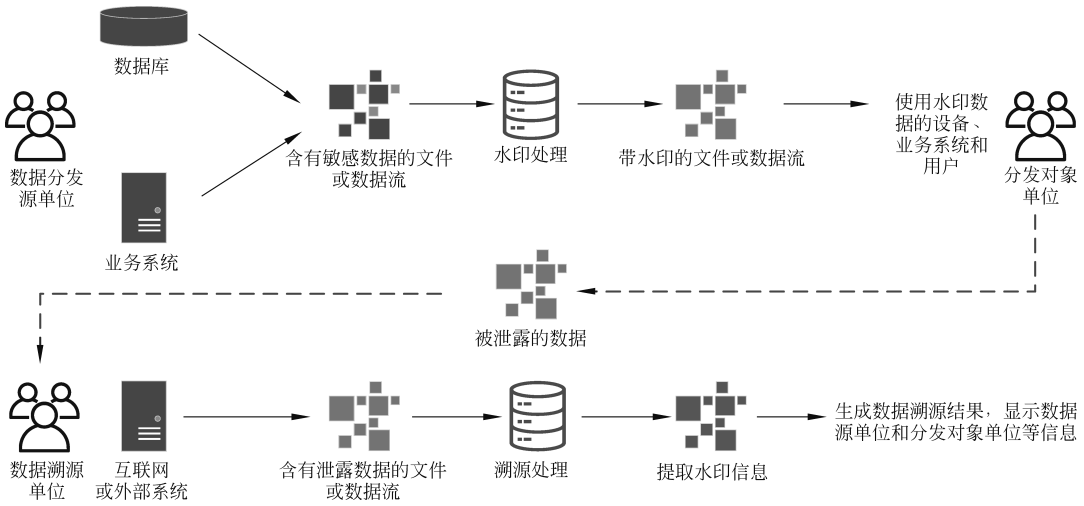


图 5-8　数据溯源示意

#### 1. 基于标注技术的溯源方法

对于文件而言，有很多冗余空间，可以隐秘地写入一些流转过程产生的标注信息。

具体来说，可以按时间序，在每次文件流转或修改的时候增加标注信息，标注信息包含当前文件的哈希值等鉴别信息、时间、源属主等。

根据应用场景选择标注信息嵌入机制：

（1）对于非文本型具有特定格式的文件，可采用信息隐藏技术嵌入文件中，随文件流转。文件无论修改与否均适用；

（2）将标注信息存储到第三方存储系统中，只适用于文件修改的场景。

#### 2. 基于数据库水印的溯源方法

对于数据库存储的数据而言，很难找到冗余空间，添加水印的难度很大，而且鲁棒性不够高，容易被擦除。因为，数据库存储对数据提出了严苛的限制。

即使如此，仍有一些数据库水印算法提出，包括伪行、伪列等，如表 5-2 所示。

表 5-2　数据库水印算法示例

| 应用场景 | 算法名称 | 原理说明 | 重点突破 |
| --- | --- | --- | --- |
| 对单条数据的查询 | 伪行算法 | 增加伪行实现水印嵌入 | 原始数据规模、数据属性关系、数据仿真度 |
| 对数据的统计查询 | 伪列算法 | 增加伪列实现水印嵌入 | 数据重复性、数据仿真度 |
| 文本型数据查询 | 文本属性算法 | 添加不可见字符实现水印嵌入 | 规则确定、防擦除 |
| 数值数据查询 | 数值属性可逆算法 | 替换最低有效位实现水印嵌入 | 精度失真比例、执行性能 |

## 5.4 保留格式加密及应用

脱敏后的数据通常会被用来做数据分析等任务，为了满足数据分析后结果脱敏的需求，需要有可逆脱敏的技术支持。保留格式加密(format-preserving encryption，FPE)能确保密文与明文具有相同的格式，可以提供可逆脱敏的能力。

目前，NIST 已经接受 FPE 算法，并颁布了两种标准算法：FF1 算法和 FF3 算法。

### 5.4.1 基本定义

基于 FPE 已有的研究成果，可以从两个角度对 FPE 进行定义：基本 FPE 和一般化 FPE。基本 FPE 描述了 FPE 要解决的问题，即确保密文属于明文所在的消息空间；一般化 FPE 则强调 FPE 问题的复杂性在于待加密消息空间的复杂性。

**定义 5-1(基本 FPE)**　FPE 可以简单描述为一个密码 $E$：$K\times X\to X$，其中，$K$ 为密钥空间，$X$ 为消息空间。

基本 FPE 强调明文和密文处于相同的消息空间，因此具有相同的格式。以 $n$ 位信用卡号的保留格式加密为例，密文要求和明文一致都是由十进制数字组成的长度为 $n$ 的字符串，即两者均为消息空间$\{0,1,\cdots,9\}^n$ 内的元素。根据基本 FPE 的定义，分组密码也是一种特殊的 FPE，它是由分组长度 $n$ 决定的$\{0,1\}^n$ 字符串集合上的置换。然而，FPE 要处理的消息空间远比分组密码复杂得多，比如，格式为 YYYY-MM-DD 的日期型消息空间，不仅有长度为 10 的字符串长度限制，还需要满足特定位置是字符-、年、月、日在合理范围内等格式要求。

为了更准确地描述 FPE 问题，定义集合 $\Omega$ 为格式空间，任意一个格式 $\omega\in\Omega$，可确定消息空间的一个与格式 $\omega$ 相关的子空间 $X_\omega$。FPE 与集合$\{X_\omega\}_{\omega\in\Omega}$有关，称 $X_\omega$ 为由格式$\omega$ 确定的消息空间的一个分片，每个分片都是一个有限集。当给定密钥 $k$，格式 $\omega$ 和调整因子 $t$ 后，FPE 就是一个定义在 $X_\omega$ 上的置换 $E_k^{\omega,t}$。

**定义 5-2(一般化 FPE)**　FPE 可以描述为一个密码 $E$：$K\times\Omega\times T\times X\to X\cup\{\perp\}$，其中，$K$ 为密钥空间，$\Omega$ 为格式空间，$T$ 为调整空间，$X$ 为消息空间。所有空间都非空，且 $\perp\notin X$。

为了有效地研究分析加密模型，可通过算法三元组$\mathcal{E}_{\mathrm{FPE}}=(\mathrm{Gen},\mathrm{Enc},\mathrm{Dec})$来描述一般化 FPE。