论文阅读笔记：隐私

纳文琪

2019 年 3 月 1 日

1 Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data[1]

1.1 引言

最近研究表明，在大数据时代简单地匿名化数据集已无法再抵御针对隐私的攻击。保护隐私最直接的办法就是移除数据集中的 ID，然而，事实表明这并不奏效。本质上，一个 ID 表示的是被描述对象的一组特征，但是实际上，我们依靠一个人的多种特征而不是 ID 来识别这个人。

隐私研究的两个方面 隐私研究集中在两个方面：内容隐私（content privary）和行为隐私（interaction privacy）。

内容隐私 攻击者根据受害者的一些知识背景从匿名或加密的数据集中识别受害者身份。

行为隐私 攻击者更关注受害者的行为。

现代隐私研究主要关注两个方面：数据聚类（data clustering）和隐私框架（privacy framework）。

1.2 基础知识

隐私参与者 包括：数据生产者、使用者、管理者、攻击者。

数据操作 包括：收集、清洗（anonymizing）、交换（communicating）。

数据属性的类型 包括：

显示标识（explicit identifier**）** 比如身份证号、特定学校里的学号等；

准标识（quasi-identifier**）** 通过关联其他数据集就可以确定用户的属性，比如性别、年龄等。我们一般将一个记录里面的所有准标识字段称为“qid”。拥有相同 qid 的值的一组记录被称为一个等价类（equivalent class）。

敏感信息（sensitive information**）** 用户希望保护的那部分数据；

其他 用户的其他信息。

1.3 隐私研究的成果

1.3.1 数据聚类方面

数据聚类方面的研究成果主要包括：k-anonymity、l-diversity、t-closeness 等。

k-anonymity 首先，为保护隐私，我们必须将数据的 ID 全部移除，这样才能避免特定用户被识别。然而即便所有 ID 被移除，攻击者还是有可能通过诸如链接外部数据库等方式，根据 qid 来识别用户。k-anonymity 方法的基本原则是：确保含有相同 qid 组数据的记录在数据集中至少出现 k 次，也就是说，每个等价类至少有 k 个记录。这样就可以使得攻击者通过 qid 识别特定用户的概率变为 $\frac{1}{k}$，当存在一个很大的 k 值的时候，会对用户的识别产生一个很大的信息损失，从而达到隐私保护的作用。k-anonymity 方法主要是用于处理准标识字段上的隐私保护问题, 但不能处理敏感数据。攻

击者可能使用同质攻击（homogeneity attach）或背景知识攻击（backgroud knowledge attack）来破解。

l-diversity　l-diversity 方法可以克服 k-anonymity 方法的缺点。它要求数据集“对每一个 qid 的值，确保敏感数据字段至少有 l 个不同的值”。为了实现它，我们需要增大（还是减小?）敏感信息字段的颗粒度或增加噪声。某些特殊的时候，l-diversity 会起到反作用，他会释放更多的信息增益给攻击者。

t-closeness　可以修复 l-diversity 的脆弱性。它的基本思想是：对于任何一个等价类，保证它的值的分布被限定在 t 范围内。

1.3.2　隐私框架方面

微分隐私（differential privacy**）**　在了解用户背景知识的情况下，攻击者可能会通过多次进行统计查询来获得期望的信息。防范策略是：对两个差别很小的数据集进行查询，其结果差别也应该很小，这样就可以限制攻击者获得的信息增益。

微分可识别性（differential identifiability**）**

成员隐私（membership privacy**）**

1.4　隐私研究的学科

1.5　隐私研究的数学描述

匿名系统　是一个映射函数：$F = X \to Y$，$X = \{X_1, X_2, ..., X_n\}$ 是原始数据，$Y = \{Y_1, Y_2, ..., Y_m\}$ 是系统的输出，对于攻击者，其目的是建立一个映射：$G : Y \to \hat{X}$，尽可能地从输出还原原始数据。

隐私保护系统的两个目的　被描述为 utility 和 privacy，这也是匿名系统 F 的两个关键指标。

utility　使用 distorion D 来度量，抽象地表示为：

$$D = \lambda(X;Y) \tag{1}$$

D 有很多种度量方法，例如，可以使用均方来表示。

privacy　使用 leakage L 来度量，抽象地表示为：

$$L = \lambda(X;\hat{X}) \tag{2}$$

L 通常使用互信息来度量，即：

$$L = I(X,\hat{X}) \tag{3}$$

给定两个阈值 D_0 和 L_0，匿名系统可作为一个优化问题：

$$\begin{aligned} &optimize F \\ &\quad s.t.\ D \le D_0 \\ &\quad\quad L \le L_0 \end{aligned} \tag{4}$$

1.6 隐私度量（Privacy Measurements）

隐私的度量至今都没有太清晰的方法。现有以下几种度量方法：

1.6.1 相对度量（Relative Measurement）

其思想是：首先给定一个标准（benchmark），再度量研究对象到此标准的距离。比较流行的距离计算方式是 Kullback-Leibler 距离（相对熵）。

KL 距离是基于一阶统计的度量方法，而二阶方法可以度量得更加精确。

1.6.2 信息论度量（Information Theoretic Measurement）

对于一个投票系统，定义三个随机变量 V、S、E，分别表示投票者所投的票、来自投票系统以外的信息、攻击者由投票系统中获得的信息。

完美隐私（perfectly privacy） 定义为：在 S 条件下，V 和 E 独立。即：$p_{V|S}(v;s) = p_{V|S,E}(v;s,e)$。

隐私损失总量（amount of privacy loss） 定义为：

$$L = \max(H(V|S) - H(V|S,E)) \tag{5}$$

度量隐私可能纰漏的程度。

1.6.3 Unicity Measure

1.7 隐私数学模型

1.7.1 k-anonymity 模型

定义数据表 $T = \{t_1, t_2, ..., t_n\}$ 是数据行的集合，$A = \{A_1, A_2, ..., A_n\}$ 是数据的属性集，$t_i[A_j]$ 表示元组 t_i 的属性 A_j 的值，$C = \{C_1, C_2, ..., C_k\} \subseteq A$ 是子属性集。定义 $T[C] = \{t[C_1], t[C_2], ..., t[C_k]\}$ 是 t 在 C 上的映射，QI 为所有准标识符的集合。

我们说一个表 T 满足 k-anonymity，如果它满足，对于每一个元组 $t \in T$ 都存在 $k-1$ 个其他的元组 $t_{i_1}, t_{i_2}, ..., t_{ik-1} \in T$ 使得 $t[C] = t_{i_1}[C] = t_{i_2}[C] = ... = t_{i_{k-1}}[C], C \subseteq QI$。也就是说，任何一组具有相同属性值的准标识符在表中至少出现 k 次。

1.7.2 l-diversity 模型

1.7.3 t-closeness 模型

1.7.4 Differential Privacy 框架

2 k-anonymity: A model for protecting privacy [2]

2.1 背景

人们希望大量数据用于研究和分析，但同时又保证不泄露隐私，即同时满足“数据利用 (utility)”和“保护隐私 (privacy)”两个要求。通常，人们在发

布数据时，会将数据集的标识符（如姓名、身份证号等）删去，从而使得数据集中的个体 (individual) 不能够被攻击者“重新识别 (re-identify)”。然后，在很多情况下，利用非标识符字段仍然可以重新识别出个体。例如，研究显示，87% 的美国人可以通过邮编、性别和生日被识别出来。k-anonymity 提供了一种避免此问题的框架。

2.2 定义

纰漏（disclosure**）** 指的是数据明显地或通过推理被意外地泄露。

纰漏控制（disclosure control**）** 的目的是去识别或限制发布数据中的纰漏，也就是说，确保发布数据具有充分的匿名性。

准标识符（Quasi-identifier**）** 如果数据表中的一组属性，可以通过与外部属性连接重新识别数据表中的个体，则称这组属性为准标识符。

准标识符的形式化定义 令 U 为数据个体的全集，$T(A_1, ..., A_n)$ 为数据表，存在两个映射 $f_c : U \to T$ 和 $f_g : T \to U'$，其中 $U' \subseteq U$。表 T 的准标识符 Q_T 是一个属性集 $\{A_1, ..., A_j\} \subseteq \{A_1, ..., A_n\}$，满足 $\exists p_i \in U$ 使得 $f_g(f_c(pi)[Q_T]) = p_i$。

2.3 模型

k-anonymity 有表 T 和它的准标识符 Q_T，当且仅当 $T[Q_T]$ 中的每一行在 $T[Q_T]$ 中至少出现 k 次时，我们说 T 满足 k-anonymity。

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

图 1:

2.4 攻击 k-anonymity

未排序匹配 (Unsorted matching) **攻击** 如果发布的表的数据的顺序固定，就可以根据这种固定顺序，关联多个发布的表来获取敏感信息。避免这种攻击很简单，只需要对数据进行乱序操作后再发布即可。

互补发布攻击（Complementary release attack） 通常情况下，我们不会将所有发布的字段都列为准标识符，这样的话，如果同意数据表被发布了多个版本，攻击者就可能会通过关联多个发布的数据表来攻击 k-anonymity。

时间攻击 (Temporal attack) 如果数据集是动态变化的，由于每个时间的数据表都可能不一样，而不同时间点之间的数据表没有规则限制，因此攻击者有可能通过关联两个不同时间点数据集来识别数据表中的个体，从而实现对 k-anonymity 的攻击。两个不同时间点的数据表就与上面所讲的两个不同版本的数据集类似。

3 Protecting respondents identities in microdata release [3]

本文提出了一种实现 k-anonymity 的泛化算法。

3.1 **泛化数据** (Generalizing data)

3.1.1 **泛化关系** (Generalization relationships)

泛化关系 是关于全体值域的集合 Dom 上的一个偏序，记作 $\leq_D$。它满足两个条件：

- $\vee D_i, D_j, D_z \in Dom : D_i \leq_D D_j, D_i \leq_D D_z \Rightarrow D_j \leq_D D_z \vee D_z \leq_D D_j$ (每个 D_i 最多只有一个直接泛化域)
- 域中的所有值最终都被泛化到一个单一值（单一最大值）

域泛化结构 (domain generalization hierarchy, DGH_D 由泛化关系得到的一个全序结构。

值泛化关系 (value generalization relationships) 是与 DGR 类似的，关于某个值域 D 下的所有值的一个偏序，记作 $\leq_V$

值泛化结构 (value generalization hierarchy, VGH_D) 与 DGH_D 类似，可以用树来表示，其叶子节点就是域 D 中的值。

值泛化关系和值泛化结构可以看做是域泛化关系和域泛化结构更低层次的抽象。其示例如图 2 。

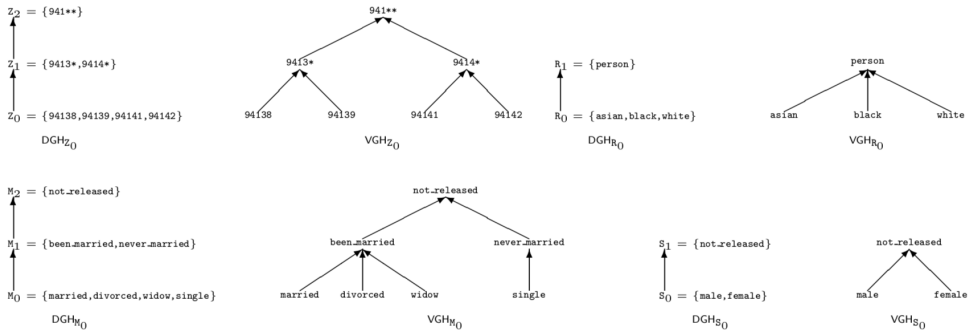

Figure 2: Examples of domain and value generalization hierarchies

图 2:

DGH **的格表示**　给定一个域元组 $DT =< D_1, ..., D_n >$，我们可以定义关于 DT 的域泛化结构为 $DGH_{DT} = DGH_{D_1} \times ... \times DGH_{D_n}$，即 DT 的 DGH 是其每个元素的 DGH 的笛卡尔积。由于每个 DGH_{D_i} 都是全序的，因此 DGH_{DT} 是一个关于 DT 的格 (lattice)，其最小值 (DT) 在底部，最大值在顶部。如图 3所示。

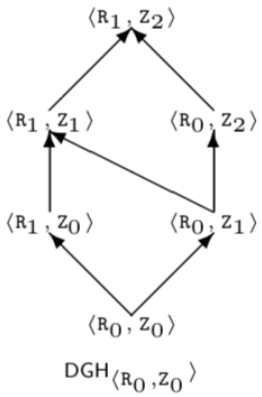

图 3: DGH 的格表示

泛化策略 (generalization strategy)　DGH_{DT} 格中的一组边及其节点组成的集合成为泛化策略。如图 4 所示。

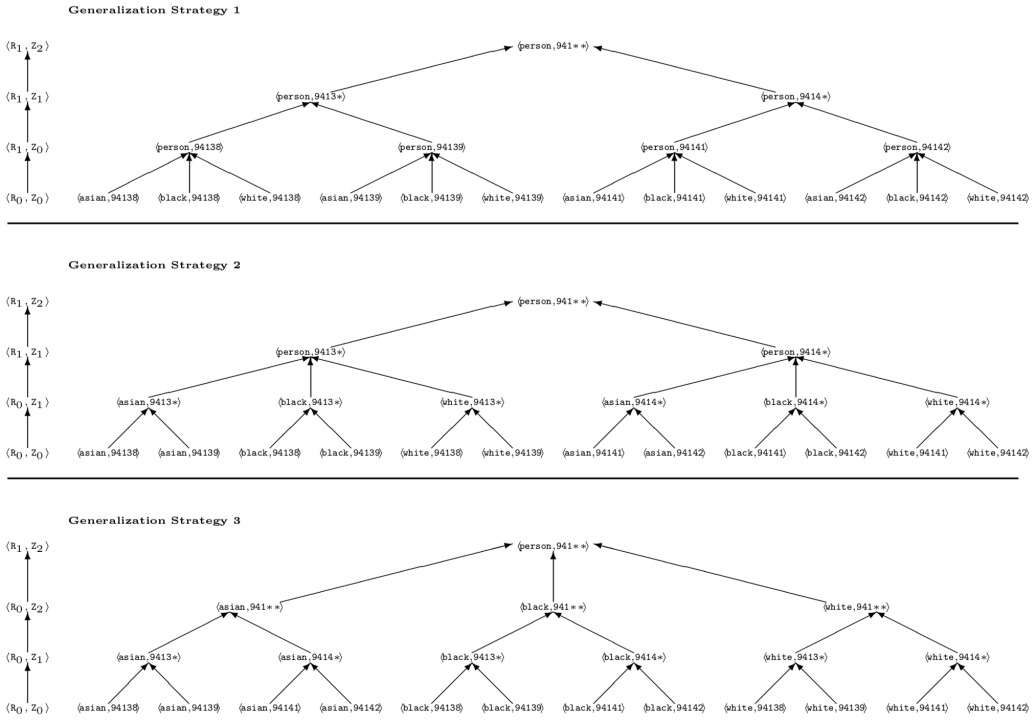

图 4: DGH 的格表示

3.1.2 Generalized table and minimal generalization

属性的值域 定义 $dom(A_i, T)$ 表示属性 A_i 在 T 中的值域。

泛化表 (Generalized Table) 令 $T_i(A_1, ..., A_n)$ 和 $T_j(A_1, ..., A_n)$ 是具有相同属性 (字段) 的两个表, 若满足以下条件则称 T_j 是 T_i 的泛化, 记作 $T_i \preceq T_j$:

- $|T_i| = |T_j|$(两个表中的数据量一致)
- $\forall A_z \in \{A_1, ..., A_n\} : dom(A_z, T_i) \leq dom(A_z, T_j)$ (每个属性满足一致的域泛化关系)
- 每个属性值满足一致的值泛化关系

距离向量 (distance vector) 两个表 T_i、T_j 满足 $T_i \preceq T_j$，从 T_i 到 T_j 的距离向量记作 $DV_{i,j} = [d_1, ..., d_n]$，其中 d_z 是 DGH_{D_z} 中 D_i 到 D_j 的距离。

距离向量的偏序关系　两个距离向量 DV、DV'，如果所有的 $d_i \leq d'_i$，则有 $DV \leq DV'$。因此，域泛化结构可以看做是距离向量的格，如图 5 所示。

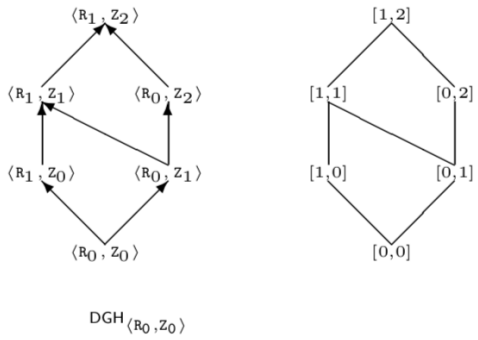

图 5: DGH 的距离向量格表示

k-minimal **泛化** (generalization)　两个表 T_i、T_j 满足 $T_i \preceq T_j$，如果 T_j 满足下列两个条件就称 T_j 是 T_i 的一个 k-minimal 泛化：

- T_j 满足 k-anonymity
- $\forall T_z : T_i \preceq T_z satisfies k-anonymity \Rightarrow \neg(DV_{i,z} \leq DV_{i,j})$(不存在任何一个比 T_j 小，而且满足 k-anonymity 的泛化表)。

3.2 Suppressing data

泛化的好处是可以在满足 k-anonymity 的条件下发布表中所有的数据条目，而 suppression 则通过移除某些条目来满足 k-anonymity，同时减少泛化所需的步骤。

带 suppression 的泛化表 (generalized table with suppression)　与泛化表相比，它不要求 $|T_i| = |T_j|$，而是要求 $|T_i| \geq |T_j|$，也就是说，泛化表的条目可以比原表少。

最小要求抑制 (minimal required suppression)　T_j 是 T_i 的满足 k-anonymity 的泛化，满足下列条件时，我们称对 T_j 执行了最小要求抑制 (也就是说，T_j

是满足 k-anonymity 的抑制中条目最多的):

$$iff\ \ \forall T_z : T_i \preceq T_j, DV_{i,z} = DV_{i,j}, T_z\ \ satisfies\ \ k-anonymity \Rightarrow |T_j| \geq |T_z|. \tag{6}$$

4 l-Diversity: Privacy Beyond k-Anonymity[4]

4.1 Introduction

4.1.1 k-anonymity 面临的两种形式的攻击

k-anonymity 面临两种形式的攻击:

同质化攻击（Homogeneity attack**）** 如同前面所提的的“互补发布攻击”、“时间攻击”，由于数据发布时，并不是所有字段都是 QI，而同一组 QI 的值相同的数据中，会出现敏感字段都相同的情况，而且这种情况并不罕见。在这种情况下，如果数据集仅执行 k-anonymity，则会由于敏感字段缺乏多样性而使得攻击者可以重新识别数据集个体，从而造成信息纰漏。同质化攻击问题的一种解决方案就是本文所讲的 l-diversity 模型。

背景知识攻击 (Background knowledge attack) 就算数据由于多样性的存在而不会受到同质攻击，但也有可能会由于背景知识（例如，可以从日常生活看出某人肯定不会患有某种特定疾病）的存在而是多样性容易排除，从而使得敏感信息泄露。

4.2 定义

数据泛化（Data generalization**）** 将一个数据集进行分组，各个分组之间互不重叠。其形式化定义为：域 $D^* = \{P_1, P_2, ...\}$ 是域 D 的泛化，如果 $\cup P_i = D$ and $P_i \cap P_j = , i \neq j$。$\phi_{D^*}(x)$ 表示 D^* 包含 x 的元素 P。

攻击者（Adversary**）的背景知识** 攻击者可能知道以下几类背景知识:

攻击者可能知道已发布的表 T 的泛化 T^*，也就是说，知道 T 的属性的值域。

攻击者可能知道某些个体存在于表中。

统计（demographic**）背景知识**　攻击者可以知道表中字段的值的分布。

4.3　贝叶斯优化隐私模型（Bayes-Optimal Privacy, BOP）

两个假设　为简化模型，给出两个假设：

- 数据表 T 是全集的一个随机样本集；
- 个体只有一个敏感字段。

另外，还假设一个比较糟糕的情况：攻击者知道准标识符 Q 和敏感字段 S 的联合分布。

度量　我们用两个指标对隐私纰漏进行度量：

先验置信（prior belief**）**　已知目标个体 t 的准标识符的值 q 的情况下，其敏感字段值为 s 的概率：

$$\alpha_{(q,s)} = P_f(t[S] = s | t[Q] = q) \tag{7}$$

后验置信（posterior belief**）**　已知目标个体 t 的准标识符的值 q，以及其泛化 t^* 的情况下，其敏感字段值为 s 的概率：

$$\beta_{(q,s,T^*)} = P_f(t[S] = s | t[Q] = q \wedge \exists t^* \in T^*, t \overset{*}{\rightarrow} t^*) \tag{8}$$

定理 3.1　令 $n_{q^*,s'}$ 是表中出现 (q^*, s') 对的个数。

定理 3.1 **的证明**

4.3.1　隐私原则

数据表发布后可能泄露隐私的两种方式

正纰漏 (Positive Disclosure)　如果发布数据表后，可以提升攻击者正确识别目标的概率，例如，对于一个 q，其对应的 s 的值都是同一个（同质化攻击），那就是正纰漏。

负纰漏 (Negative Disclosure)　如果发布数据表后，攻击者可以从中排除一些关于目标的错误的敏感值，例如，根据背景知识可以排除特定个体的一些敏感值，那就是负纰漏。

不提供信息原则 (Uninformative Principl)　隐私的理想定义应该满足不提供信息原则，即：表的发布并不能为攻击者带来除了背景信息以外的其他信息。此原则有很多实例化的方式，他们都属于贝叶斯优化隐私定义。其中一个是：

$(\rho_1, \rho_2) - Private$　给定表 T^* 和两个常数 ρ_1, ρ_2。如果 $\alpha_{(q,s)} < \rho_1 \wedge \beta_{(q,s,T^*)} > \rho_2$，或者 $\alpha_{(q,s)} < 1 - \rho_1 \wedge \beta_{(q,s,T^*)} > 1 - \rho_2$，我们就说发生了“$(\rho_1, \rho_2)$-privacy breach”。如果 (ρ_1, ρ_2)-privacy breach 没有发生，我们就说表 T^* 满足 $(\rho_1, \rho_2) - Private$。

4.3.2　BOP 的局限

BOP 存在着一些缺陷，使其无法在实践中使用。包括：

不充分知识　数据发布者不可能完全知道数据的分布。

对攻击者的无知　数据发布者不知道攻击者到底知道什么。

实例知识　理论定义不能定义那些无法被概率模型建模的知识。例如，Bob 的儿子可能会告诉 Alice，Bob 没有糖尿病。

多攻击者　可能会有多个不同知识水平的攻击者存在，他们具有的背景知识都不同，数据发布者都必须考虑到。

4.4 l-diversity: 一种实用的隐私定义

l-diversity 可以克服上述的 BOP 实践中存在的问题。

4.4.1 l-diversity 原则

等价类 $(q^* - block)$ 表 T^* 中的一组数据，他们的非敏感值都被泛化为 q^*。

l-diversity **原则** 如果有一个 $q^* - block$，它的隐私属性 S 包含至少 l 个良表示 (well-represented) 值，我们就说这个 $q^* - block$ 是 $l - diverse$ 的，如果表中所有的 $q^* - block$ 都是 $l - diverse$ 的，那么这个表就是 $l - diverse$ 的。

此原则主张确保每一个 $q^* - block$ 的敏感字段具有至少 l 个良表示值，但并没有说明良表示意味着什么。最简单的，我们可以将“l 个良表示值”理解为“l 个不同的值”。另外，定义不同的良表示值也可以得到不同的实例。

4.4.2 l-diversity 的实例

根据不同的良表示定义，可以得到不同的 l-diversity 定义。包括：可区分 l-diversity、熵 l-diversity、递归 l-diversity 等。

Distinct(**可区分**) l-diversity 要求每个等价类中至少有 l 个可区分的值。它存在一个缺陷：如果某一个取值的频率明显高于其他取值，这将使得观察者可以以较高的置信度认为这一等价类中的敏感属性都取这个值。

Entropy l-diversity 设 $p_(q^*, s)$ 为等价类中敏感值为 s 的概率，那么表满足 l-diversity 的条件是，每一个等价类都满足：

$$-\Sigma_{s\in S} p_(q^*, s)\log(p_(q^*, s)) \geq \log(l) \tag{9}$$

这要求整个表的熵至少是 $\log(l)$，也就是要求至少有 l 个可区分的值。这并不容易实现，特别是当某个敏感值非常常见的时候。

Recursive(c,l)-diversity 递归 l-diversity。

4.4.3 多敏感属性的情形

以上讨论都基于一个基本的假设：每条数据只有一个敏感字段。

Multi-Attribute l-diversity 若把 S_i 之外的所有 S 都当做 Q 还满足 l-diversity，那么表 T 就满足 l-diversity。

参考文献

[1] Shui Yu, "Big privacy: Challenges and opportunities of privacy study in the age of big data", IEEE access, vol. 4, pp. 2751–2763, 2016. 1

[2] Latanya Sweeney, "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002. 5

[3] Pierangela Samarati, "Protecting respondents identities in microdata release", IEEE transactions on Knowledge and Data Engineering, vol. 13, no. 6, pp. 1010–1027, 2001. 7

[4] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam, "\ell-diversity: Privacy beyond\kappa-anonymity", in null. IEEE, 2006, p. 24. 12