arXiv:2304.09546v1 [cs.DB] 19 Apr 2023

差分隐私查询处理的敏感度估计

张美凡[a]，刘欣[a]，尹丽华*[a]

[a]*Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, 510006, , China*

摘要

差分隐私已成为数据分析、查询处理和机器学习中一种广泛采用的隐私保护方法，它通过向查询结果添加噪声来避免隐私泄露。敏感度，即删除或插入一个元组对查询结果产生的最大影响，决定了所添加噪声的大小。计算一些简单查询（如计数查询）的敏感度较为容易，但计算包含连接操作的复杂查询的敏感度则颇具挑战。此类查询的全局敏感度可能大到无界，从而损害查询答案的准确性。弹性敏感度和残差敏感度为局部敏感度提供了上界，可用于减少噪声，但它们要么准确性较低，要么计算开销较高。我们提出了两种快速的查询敏感度估计方法，分别基于采样和草图；与最先进的方法相比，这两种方法兼具竞争力的准确性和更高的效率。

关键词：差分隐私、连接查询、近似查询处理。

1. 引言

差分隐私通过向查询答案添加随机噪声，为查询处理提供强有力的隐私保障。通常，噪声由查询的敏感度决定；敏感度是指在仅相差一条记录的两个数据集上，

*通讯作者*URL:* yinlh@gzhu.edu.cn(尹丽华*)

查询结果之间的最大差异。例如，查询“SELECT Count(*) FROM data WHERE Salary >5000;”的敏感度为 1，因为插入或删除一个元组对结果的影响最多为 1。然而，包含连接运算符的查询，其敏感度更为复杂，因为添加或删除一个元组可能会对查询结果产生显著影响。对于图 1 所示关系上的查询 $q{=}Count(R_1(A,\ B)\ \bowtie\ R_2(B,\ C)\ \bowtie\ R_3(C,\ D))$，要在确保隐私、准确性和效率的同时计算此类查询的敏感度并不容易。此类查询的全局敏感度无界，因为它取决于查询而非数据。局部敏感度同时取决于数据和查询，虽然引入的误差较小，但不满足差分隐私。弹性敏感度（ES） [1] 和残差敏感度（RS） [2] 等最先进的敏感度定义采用局部敏感度的平滑上界来解决这一问题，但准确性和效率仍不尽如人意。下面通过一个示例说明 ES 和 RS 的主要思想及其不足。

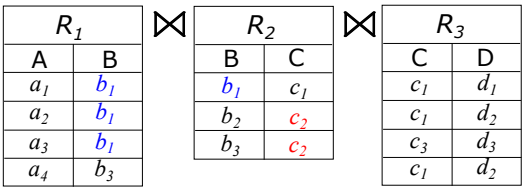


图 1：删除一个元组对连接大小的影响。

示例 1. 图 *1* 展示了一个连接查询 $q{=}Count(R_1(A,B)\ \bowtie\ R_2(B,C)\ \bowtie\ R_3(C,D))$。删除或插入关系中的一个元组时，*ES* 和 *RS* 使用不同的方法计算局部敏感度的上界。

(1) ES 根据每个连接属性的最大频数计算上界

，如下所示。

$$max(mf(R_1.B) \cdot mf(R_2.C)$$
$$mf(R_1.B) \cdot mf(R_3.C)$$
$$mf(R_2.B) \cdot mf(R_3.C))$$

$mf(X)$ 表示属性 X 的最频繁值的频数。$mf(R_1.B) \cdot mf(R_2.C)$ 表示属性 C 在 $R_1 \bowtie R_2$中的频数上界。最坏情况敏感度出现在以下情形： R_2 中的每个元组都具有$(R_2.C)$ 的最频繁值，且同时包含属性值$(R_2.B)$ ，该值与$(R_1.B)$的最频繁值相匹配。但现实并非总是最坏情况。如图 *1* 所示，$(R_1.B)$ 的最频繁值是 b_1，与$(R_2.C)$的最频繁值 c_2 并不匹配。因此，从 R_3中插入或删除一个元组所产生的实际影响远小于该上界。

(2) RS 根据一组分组查询计算局部敏感度的上界。例如，为了计算从R_3 添加一个元组对查询结果的影响，*RS* 会计算残差查询 $R_1 \bowtie R_2$ （属于 q）的一个称为“最大边界”的统计量，它表示属性$R_2.C$ 在 $R_1 \bowtie R_2$中的最大频数。由于 $R_2.C$ 可以与 R_3连接，$R_1 \bowtie R_2$ 的“最大边界”决定了向 R_3添加一个元组所产生的最大影响。可以通过以下查询计算这个“最大边界”。

Q_1: SELECT MAX(count) from (SELECT R2.C, COUNT(*) from R1,R2 where R1.B=R2.B GROUP BY R2.C) as T;

由于考虑了子查询 $R_1 \bowtie R_2$ 的真实连接结果，因此能够更准确地估计删除或插入一个元组对 R_3 的影响。然而，这样的查询仍然开销很大，因为它包含“连接”和“分组”运算符。计算成本更高，因为我们需要计算在连接查询 q 中从每个表插入或删除一个元组所产生的影响。也就是说，计算 *RS* 需要执行许多形式为 Q_1 的分组查询。

通常，ES 可以根据连接值的频率高效计算，但误差较大。RS 更准确但效率较低，因为它

基于某些子查询的真实结果。如 [2], 所示，根据 RS 计算带噪多路连接结果所需的查询处理时间几乎增加了 10 倍，效率过低，无法支持快速响应。目前尚无现有工作能够在减少误差的同时高效计算多路连接查询的敏感度。我们观察到，查询敏感度通常基于数据库的某些统计量进行计算，而更准确的敏感度估计依赖于更复杂且耗时更长的统计量。为解决这一问题，我们提出使用近似查询处理（AQP）方法来近似估计查询敏感度。我们提出了一个敏感度估计框架，如图 2 所示。本文重点研究多路连接查询的敏感度估计。但近似估计查询敏感度的基本思想也可用于多种查询。

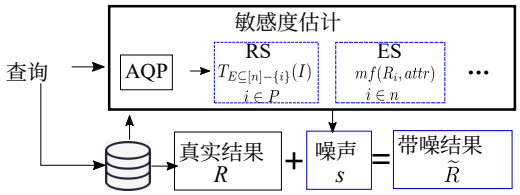


图2：框架

估计查询敏感度仍非易事。一方面，直接使用采样方法来估计分组查询会产生不必要的计算成本。分组查询通常通过估计每个组的聚合值来执行；然而，如图 1 所示，RS 的此类分组查询的聚合函数为最大计数，这意味着估计最大组以外的其他组并不会对结果产生贡献。因此，我们需要专注于估计最大组。另一方面，RS 需要计算许多仍包含连接运算符的子查询，即使对每个子查询采用采样，计算仍然耗时。因此，我们提出了另一种算法，以使用更少的样本来估计所有所需的统计量。

即使采样方法能够降低敏感度计算的成本，在线抽取样本仍然开销很大。

为解决这一问题，我们提出了另一种基于草图的敏感度估计方法。草图适用于概括数据流，可以通过单遍扫描数据来构建。我们提出了基于草图的敏感度估计方法：离线为每个关系构建一个草图，并利用这些草图估计 $O(1)$ 中在线查询的敏感度。

本文的贡献如下：

- 我们提出了一种基于采样的敏感度估计方法，根据随机游走估计多连接查询的残差查询中最大分组的频率。该方法在保持与 RS 相近准确性的同时提高了效率。

- 我们提出了一种基于草图的敏感度估计方法。根据 AGMS 草图定义草图敏感度。我们提出的基于草图的方法与 ES 效率相当，但准确性远高于 ES。

- 实验结果表明，在保持准确性相当的同时，这两种方法的效率均高于计算 RS 和 ES。

本文其余部分安排如下。第 2 节回顾相关工作。第 3 节介绍差分隐私的预备知识以及现有的敏感度定义。第 4 节介绍两种基于采样和草图的多连接查询敏感度估计方法。第 5 节对我们的方法与现有方法进行实验比较。第 6 节总结全文。

2. 相关工作

差分隐私 [3]是一种用于量化和管理隐私风险的数学框架。它广泛应用于隐私保护数据发布 [4, 5]、数据挖掘 [6, 7]、、机器学习 [8, 9] 和社交网络分析 [10]领域。通过添加与查询结果敏感度相关的噪声，可以轻松地应用差分隐私来保护

查询处理的隐私 [11] 。

计算连接查询的敏感度颇具挑战。已有多种支持连接查询的机制，例如 PINQ [12] 和 wPINQ [13]，但它们都使用全局敏感度，这意味着敏感度可能非常大，从而导致效用较低。Nissim 等人 [14]提出了局部敏感度：将任意两个相邻数据集中的一个固定为当前待查询的数据集，并考虑它的所有邻居。但这种方法不满足差分隐私的要求。平滑敏感度 [14] 是局部敏感度最紧致的平滑上界，能够防止局部敏感度导致的隐私泄漏，但其计算复杂度为 NP-困难。弹性敏感度（ES） [1] 和残差敏感度（RS） [2] 都基于寻找局部敏感度平滑上界的思路。ES 根据每个关系中连接值的最大频数计算，计算速度快，但引入的噪声比查询答案高出数个数量级。RS 根据多连接查询中的一些残差查询计算，而残差查询的最大边界通过计算将某个元组与其他私有关系连接的上界，来衡量向某个关系添加该元组所产生的影响。

近似查询处理（AQP）是一种可用于高效估计连接大小的技术 [15]。对连接结果进行采样是性能较高的 AQP 方法之一。Ripple join [16] 和 Wander Join [17] 是可用于包含连接运算符的查询的在线聚合方法。Zhao 等人 [18]重新审视了这一问题，将先前的方法整合到一个通用框架中。该框架包含两个主要步骤：计算连接大小权重的上界，然后从连接中采样。草图是用于流摘要任务的概率数据结构；人们提出了大量草图，例如 AGMS [19]、Count-sketch [20], 和 Count-Min 草图 [21] ，用于频率估计、重频繁项挖掘和连接大小估计等。AGMS 草图专为自连接估计而设计，也可用于估计包含条件过滤器 [22]的多连接查询的连接大小。

尽管人们为连接查询提出了多种敏感度定义，但这些定义都存在一些缺陷。对于多连接查询，全局敏感度没有界限。局部敏感度不满足差分隐私。平滑敏感度的计算是 NP-困难的。ES 易于计算，但准确性较低。RS 更准确，但耗时更长。现有研究尚未尝试采用 AQP 方法进行敏感度估计。然而，采样和草图等 AQP 方法无法直接用于降低计算成本。原因在于，多连接查询的敏感度取决于许多统计量，如 RS 所示；而逐一估计这些统计量仍然代价高昂。因此，我们提出一种基于采样的敏感度估计方法，着重估计那些有助于计算敏感度的统计量。我们还提出一种基于草图的方法，离线为每个关系构建草图，并在线高效地估计敏感度。

3. 预备知识

本节将介绍差分隐私和不同类型敏感度的相关定义。本节中使用的符号如表 1 所示。

表 1：符号表。

符号	含义
ϵ	隐私预算
δ	纯差分隐私不成立的概率。
$LS_q^k(I)$	数据库上距离 k 于 I 的 q 的局部敏感度。
$\overline{LS_q^k(I)}$	通过 ES 计算得到的 $LS_q^k(I)$ 的上界。
$\underset{(I)}{\check{}LS_q^k}$	由 RS 计算的 $LS_q^k(I)$ 的上界。
$mf(A)$	属性 A 的最频繁值的频数。
$T_E(I)$	残差查询 q_E 的最大边界。

3.1. 差分隐私

差分隐私（DP） [3] 最初由 Dwork 于 2006 年提出，用于解决统计数据库中的隐私泄漏问题，确保数据集中的任意

个人是否在数据集中对最终发布的查询结果影响很小。

定义 1. (ϵ, δ)-差分隐私。 若随机化隐私算法 A对于满足 $d(I, I') = 1$的任意一对输入数据集 I 和 I'，以及所有可能输出的集合 S，均满足 ϵ-*DP*。

$$\Pr[A(I) \in S] \leq exp(\epsilon) \cdot \Pr[A(I') \in S] + \delta \tag{1}$$

参数 ϵ 集是人为设定的，也称为隐私预算，用于表示隐私保护的程度。数值越小，保护程度越高。当 $delta = 0$时，我们称该算法满足 ϵ-差分隐私。

定理 1. 拉普拉斯机制 *[11]* 给定一个数据集 I，函数 $f{:}I \to R^d$。因此，随机算法 A： $A(I) = f(I) + Y$ 可提供 ϵ-差分隐私保护，而 $Y \sim Lap(\triangle f/\epsilon)$ 表示随机噪声，其服从尺度参数为 $\triangle f/\epsilon$ 的拉普拉斯分布。显然，称为"敏感度"的参数 $\triangle f$在噪声计算中起着重要作用。

3.2. 敏感度

敏感度用于衡量在结果中插入或删除一个元组时，查询结果的最大变化。本节将介绍五种不同的敏感度定义。

全局敏感度定义如下，它计算距离为 1 的两个数据库上的查询结果之间的最大差异，即 $d(I, I') = 1$。全局敏感度仅与 q有关。例如，简单 COUNT 查询的全局敏感度为 1，无论在哪个数据库上进行查询都是如此。但根据此定义，连接查询的敏感度是无界的。

定义 2. 全局敏感度。 ***[11]*** 对于 q： $D^n \to R^d$ 以及所有 $I, I' \in D^n$， q 的全局敏感度为

$$GS_q = \max_{I, I': d(I, I')=1} ||q(I) - q(I')|| \tag{2}$$

局部敏感度的定义方式类似，唯一的区别是它是针对固定数据库计算的。

定义 3. 局部敏感度。 *[14]* 对于 $q: \ D^n \to R^d$ 和 $I' \in D^n$， q 关于 I 的局部敏感度为

$$LS_q(I) = \max_{I':d(I,I')=1} ||q(I) - q(I')|| \tag{3}$$

局部敏感度引入的噪声较少，但根据局部敏感度发布查询结果并不满足差分隐私的定义。为解决这一问题，人们提出了平滑敏感度，用于计算局部敏感度的平滑上界。

定义 4. 平滑敏感度。 *[14]* 平滑敏感度基于对 f 在距离 k处的局部敏感度进行泛化来定义，

$$LS_q^k(I) = \max_{I' \in D^n:d(I,I')=k} LS_q(I') \tag{4}$$

q 关于 I 的平滑敏感度为

$$SS_q(I) = \max_{0 \le k \le n} e^{-\beta k} LS_q^k(I), \beta = \frac{\epsilon}{2 \ln 2/\delta} \tag{5}$$

计算平滑敏感度需要指数时间，这对于大数据上的查询而言代价很高。为降低成本，人们提出了弹性敏感度和残差敏感度来计算多路连接的敏感度；二者都是局部敏感度的平滑上界。Nissim 等人 [14]证明，任何局部敏感度的平滑上界都能满足差分隐私，从而保护隐私。

定义 5。 弹性敏感度。 *[1]* 数据库的弹性敏感度 I基于局部敏感度的平滑上界定义如下。

$$ES_q(I) = \max_{0 \le k \le n} e^{-\beta k} \widetilde{LS}_q^k(I), \beta = \frac{\epsilon}{2 \ln 2/\delta} \tag{6}$$

其中 $\widetilde{LS}_q^k(I)$ 是距离 k处局部敏感度的上界，

$$\begin{aligned} \widetilde{LS}_q^k(I) = \max_{i \in P}(& \prod_{j \in P - \{i\}} (mf(x_j \cap x_{p(j,i)}, I_j) + k) \\ & \cdot \prod_{j \in [n] - P} mf(x_j \cap x_{p(j,i)}, I_j)) \end{aligned} \tag{7}$$

$mf(x, I_j)$ 是属性 x 在 I_j中的最大频数，且 P是私有属性集。

弹性敏感度可以基于连接属性的最大频数进行计算。但它基于这样的假设来计算多路连接的最大频数：所有频数最高的连接属性值都可以彼此连接。因此，弹性敏感度会产生较大的误差。

定义6。 残差敏感度。 *[2]* 数据库的残差敏感度I 也基于距离k处局部敏感度的平滑上界进行定义，

$$RS_q(I) = \max_{0 \le k \le n} e^{-\beta k} \min(\hat{GS}_q, \hat{LS}_q^k(I)), \beta = \frac{\epsilon}{2 \ln 2/\delta} \tag{8}$$

其中，$\hat{}LS_q^k\ (I))$ 是距离 k处局部敏感度的上界，

$$\hat{LS}_q^k(I) = \max_{s \in S^k} \max_{i \in P} \hat{T}_{[n]-\{i\},s}(I) \tag{9}$$

$$\hat{T}_{E,s}(I) = \sum_{E' \subseteq E} \left(T_{E-E'}(I) \prod_{i \in E'} s_i \right) \tag{10}$$

$\hat{}T_{E,s}(I)$ 计算 I'： $d(I, I') = k$, s 的最大边界，这是将 k 划分为 E中不同关系的一种方式。

我们使用以下示例来演示残差敏感度的计算。

示例2。 考虑图 *1* 中的查询 $q = R_1 \bowtie R_2 \bowtie R_3$ ，假设R_1、R_2和 R_3 都是私有关系。根据残差敏感度的定义，

$$\hat{LS}_q^k(I) = \max_{s \in S^k} \max\{\hat{T}_{\{1,2\},s}(I), \hat{T}_{\{2,3\},s}(I), \hat{T}_{\{1,3\},s}(I)\} \tag{11}$$

其中 $s = s_1$、 s_2、 s_3 和 $s_1 + s_2 + s_3 = k$。$\hat{}T_{\{1,2\},\ s}\ (I)$ 是以下查询的上界：

*SELECT MAX(*组*) from(SELECT* $\hat{}R2$*.C,COUNT(*) from* $\hat{}R1$*,* $\hat{}R2$ *where*$\hat{}\hat{}$ $\hat{}R1$*.B=*$R2.B$ *GROUP BY* $R2$*.C) as T;*

ˆ ˆ这里， $R1$ 是插入 s_1 个元组后的关系，而 $R2$ 是插入 s_2 个元组后的关系。

根据公式*10*，

$$\hat{T}_{\{1,2\},s}(I) = T_{\{1,2\}}(I) + s_1 \cdot T_{\{2\}}(I) + s_2 \cdot T_{\{1\}}(I) + s_1 \cdot s_2 \tag{12}$$

我们可以用同样的方法计算ˆ$T_{\{2,3\},\ s}(I)$和ˆ$T_{\{1,3\},\ s}(I)$。最后，可以根据公式*8*计算残差敏感度。

如上例所示，残差敏感度需要计算每个T_E，其中对于每个私有关系 $R_i \in \{R_1, ..., R_n\}$，$E \subseteq [n] - \{i\}$。每个 T_E都可以通过类似示例中形式的聚合查询来计算；当查询包含连接运算符时，计算仍然开销较大。

4. 连接查询的敏感度估计

本节分别提出两种基于采样和草图的敏感度估计方法。本节中的符号见表2。

表2：敏感度估计的符号。

符号	含义
q_E	多连接查询 q 在子集 E 上的残差查询。
$m_{E,i}$	残差查询 q_E 第 i 组的样本量。
$\tau_{E,i}$	q_E 第 i 组置信区间的半宽度。
g	残差查询的组数。
J	连接大小
η	置信区间不成立的概率
$sk(R)$	关系 R的 AGMS 草图

4.1. 基于采样的敏感度估计

计算多路连接查询的残差敏感度开销较大，因为它需要对一系列残差查询计算分组聚合。降低开销的一个基本思路是使用采样方法来估算结果。

示例 1 中以 Q_1 形式表示的每个残差查询。受“Wander- Join”[23], 启发，我们在第 4.1.1 节介绍一种基于随机游走快速估计此类残差查询的方法。由于残差查询结果仅与连接结果中的最大组相关，因此我们重点关注最大组的连接路径。为进一步降低成本，我们提出一种使用一组连接路径估计所有残差查询的方法，详见第 4.1.2 节。

4.1.1. 单个残差查询的估计

计算每个残差查询最大边界的真实结果代价高昂。为此，我们提出一种基于采样的敏感度估计方法（Sampling-SE）。受“Wander-Join”[23], 的启发，我们通过以下步骤启动算法：

算法 1 Sampling-SE

输入： 多连接查询 q

输出： q 的敏感度

1: **对于**每个残差查询 q_E （属于 q ）**执行**

2: $T_E \leftarrow RQE(q_E)$

3: **结束循环**

4: 根据定义6，基于 T_E 为每个 q_E 计算敏感度。

以轮转方式对每个组的连接路径进行采样。为了确保所有组都得到充分估计，“Wander-Join”会迭代地选择置信区间最大的组，以启动下一次随机游走。与此不同，我们并不关注所有组，而只关注最大组。为此，我们采用类似于“iFOCUS”的思路 [24] ，对我们关注的组进行更多采样。区别在于，“iFOCUS”为了保证排序，会不断移除置信区间与其他组不重叠的组；而我们则不断移除置信区间与最大组不重叠的组。

Sampling-SE算法的伪代码如算法 1 所示。由于计算多连接查询中每个残差查询的 T_E 代价高昂，我们采用基于采样的方法 RQE 来估计每个查询的最大边界 T_E

残差查询 q_E。

算法 2RQE

输入： 残差查询 $q_E = \bowtie_{i \in E} R_i$，误差界 τ_0

输出： 估计值 T_E

1: 初始化计数估计 $C_1, C_2, ..., C_g$ 对于每个不同值 $v_1, v_2, ..., v_g$ 使用 m 随机游走。 $G = \{1, 2, ..., g\}$

2: 根据第 1 步中的随机游走，初始化连接大小 J 和误差界 τ_J，用于 q_E 游走。

3: $n \leftarrow m \cdot g$

4: **当** $\tau > \tau_0$ **时**

5: $m \leftarrow m + 1$

6: **对于**每个 $i \in G$，**执行**

7: $n \leftarrow n + 1$

8: 从 $t(v_i)$ 开始进行随机游走 p。

9: $C_i, \quad \tau_i \leftarrow$ 估计(p, $t(v_i) \bowtie q_E$, m, C)

10: $J, \quad \tau_J \leftarrow$ 估计(p, q_E, n, J)

11: $\tau \leftarrow \tau_i \cdot (J + \tau_J)$

12: **结束循环**

13: **对于** 每个 $i \in G$，**执行**

14: **如果** $C_i + \tau < \underset{j \in G}{\text{MAX}} (C_j - \tau)$ **然后**

15: $G \leftarrow G - \{i\}$

16: **结束如果**

17: **结束循环**

18: **结束while循环**

19: **返回：** $T_E = max_{i \in G}(C_i + \tau)$

残差查询估计（RQE）方法的细节如算法 2 所示。该算法首先根据随机游走为每个组执行 m 连接路径，并初始化 $G = \{1, 2, ..., g\}$ 中每个组的估计值 C_1，C_2，..... C_g（第 1 行）。然后，我们估计查询 q_E的连接大小 J以及置信区间的半宽度 τ_J。该算法的主要部分

算法3估计(p, q, m, C)

1: $x \leftarrow$ 根据 WanderJoin 估计 $|q|$，路径为 p。

2: $C \leftarrow \frac{m-1}{m}C + \frac{1}{m}x$

3: $\tau = \sqrt{\frac{}{}}$

4:**返回：** C ，τ.

（第3-18行）会迭代增加每个候选大组的样本量，其置信区间与当前最大组的置信区间存在重叠。对于 G 中的候选大组，算法添加随机游走路径以更新估计值（第6-12行）。算法 $Estimate$根据连接路径 p （第9行）计算第 i 组的新估计值 C_i和置信区间的半宽度 τ_i，然后以相同方式估计 J 的连接大小和置信区间（第10行）。对于置信区间与当前最大组没有重叠的小组，我们将其从 G 中移除（第13-17行）。当置信区间的半宽度 τ 小于 τ_0 时，算法停止，这意味着估计已经足够准确。

算法3根据连接路径 p 估计查询 q 的结果大小 C，并根据霍夫丁不等式 [25] 计算置信区间的半宽度 τ。

我们通过两个步骤证明，对于每个 T_E，算法的输出都是足够准确的估计：(1) 算法不会漏掉最大组；(2) 真实最大组的大小大于算法输出的概率小于 η。

第一步。 我们使用以下定理2证明算法不会漏掉最大组。

定理2。 如果对于每个组 i，对于每个 $1 \leq m \leq N$都有 $|C_i - \mu_i| \leq \tau$，那么在终止时刻，最大组 $j \in G$ 。

证明。 假设最大组 $j \notin G_{terminate}$，则存在一个组 k，其下界高于组 j 的上界，即根据算法 2（第 14 行）， $C_j + \tau < C_k - \tau$。由于对于每个

$1 \leq m \leq N$,

$$\mu_j \leq C_j + \tau < C_k - \tau \leq \mu_k \tag{13}$$

由于 $\mu_j < \mu_k$, j 不是最大组，这与假设相矛盾，因此该假设不成立。

□

第二步。我们使用定理 3 证明，真实结果大于算法输出的概率是有限的。

引理 1。霍夫丁不等式 *[25]*。设 $Y = y_1, y_2...,y_N$ 是 *[0,1]* 中的一组 N 值，其平均值为 $\frac{1}{N}\sum_{i=1}^{N} y_i = \mu$。设 $X_1,...,X_m$ 是从 Y 中不放回抽取的一组随机变量。对于每个 $1 \leq g \leq N$ 和 $\epsilon > 0$,

$$\Pr[\max_{g \leq m \leq N-1}(\frac{\sum_{i=1}^{m} X_i}{m} - \mu) \geq \tau] \leq \exp(-2g\tau^2) \tag{14}$$

假设 J 是 $\{\bowtie_{i \in E} R_i\}$ 的总连接大小，我们将每个组 C_i 的连接大小除以 J，使每个 $\frac{C_i}{J} \in [0,1]$。因此，我们可以使用上述不等式得到每个组的误差界。

定理 3. 对于 G 连接大小中的所有组，以及算法 *1* 中的所有轮次，我们有 $\Pr[\exists i, m, 1 \leq i \leq g, 1 \leq m \leq m_i : (est_{i,m} - \mu_i) \geq J_{up} \cdot \tau] \leq \eta$，其中 $\tau = \sqrt{\frac{2\log\log(m) + \log((g+1)\pi^2/6\eta)}{2m}}$，且 $J_{up} = J_{est} + \sqrt{\frac{2\log\log(n) + \log((g+1)\pi^2/6\eta)}{2n}}$。这里，$m$ 是用于估计每个组大小的样本数，n 是用于估计总连接大小的样本数。

证明。我们以与 IFO-CUS [24]，中定理 3.2 类似的方式证明该定理。不同之处在于，我们只需证明估计值超过置信区间上界的概率受到限制。我们使用上述引理计算每个组的上界，如下所示。

$$\tau_i = \sqrt{\frac{(2\log\log(m) + \log(\pi^2/6\eta_i))}{2m}} \tag{15}$$

于是，$\Pr[\exists m, 1 \le m \le N\text{：}\ (\frac{\sum_{i=1}^{m} X_i}{m} - \mu) > \tau] \le \eta$。

$$\begin{aligned}
&\Pr[\exists m, 1 \le m \le N : (\frac{\sum_{i=1}^{m} X_i}{m} - \mu) > \tau_m] \\
&\le \sum_{r=1} \Pr[\exists m, \kappa^{r-1} \le m \le \kappa^r : (\frac{\sum_{i=1}^{m} X_i}{m} - \mu) > \tau_m] \\
&\le \sum_{r=1} \Pr[\max_{\kappa^{r-1} \le m \le N-1} (\frac{\sum_{i=1}^{m} X_i}{m} - \mu) > \tau_{\kappa^r}]
\end{aligned} \tag{16}$$

根据引理 1，

$$\Pr[\max_{\kappa^{r-1} \le m \le N-1} (\frac{\sum_{i=1}^{m} X_i}{m} - \mu) > \tau_{\kappa^r}] \le \frac{6\eta}{\pi^2 r^2}. \tag{17}$$

由于 $\sum_{r \ge 1} \frac{1}{r^2} = \frac{\pi^2}{6}$，

$$\sum_{r=1} \Pr[\max_{\kappa^{r-1} \le m \le N-1} (\frac{\sum_{i=1}^{m} X_i}{m} - \mu) > \tau_{\kappa^r}] \le \eta \tag{18}$$

公式 15 计算了用于估计$\frac{C_i}{N}$的置信区间半宽度 τ_i。我们可以将 τ_i 乘以总连接大小 $N = J$，轻松计算出 C_i的置信区间半宽度。连接大小 J事先未知，但也可以根据每轮选取的随机游走进行估计。假设我们获得了 n 条路径来估计 J，

$$\tau_J = \sqrt{\frac{(2\log\log(n) + \log(\pi^2/6\eta))}{2n}} \tag{19}$$

那么，$\Pr[\exists n, 1 \le n \le N\text{：}\ |\frac{\sum_{i=1}^{n} X_i}{n} - \mu| > \tau_n] \le \eta$。

将 J_{up} 视为包含所有连接结果的大组的上界，我们得到 $g+1$ 组。假设对于每一对 $i,\ j \in \{0, 1, ..., g+1\}$，我们有 $\eta_i = \eta_j\ \eta_i = \eta/(g+1)$。在公式 15 和公式 19 中将 η 设为 $\eta/(g+1)$，即可得到算法 2 第 11 行中的 τ。

□

由于许多组在其估计的上界低于最大组的下界时就会被移除，因此每组的样本复杂度各不相同。我们使用以下定理来证明算法的样本复杂度。

定理 4. 样本复杂度。 以至少 $1-\eta$ 的概率，算法 *2* 输出最大组的上界，并抽取 $O\left(J_{up}^2 \sum_{i=1}^{g} \log(\frac{g}{\eta}) + \log\log(\frac{1}{\alpha_i}) \right)$ 个样本。这里， $\alpha_i = \max\{\frac{|\mu_{max}-\mu_i|}{4}, \tau_0\}$

证明。 我们首先证明，对于每个组，一旦 i 达到 τ ≤最大值$\{\frac{|\mu_{max}-\mu_i|}{4}, \tau_0\}$，采样就会停止。

如果组 i 在 τ 达到 τ_0之前就从 G 中移除，即 $\tau > \tau_0$，那么 $C_i + \tau < C_{max} - \tau$。由于置信区间始终包含真实结果，因此 $\mu_{max} \in [C_{max} - \tau, C_{max} + \tau]$，并且 $\mu_i \in [C_i - \tau, C_i + \tau]$。为确保 $C_i + \tau < C_{max} - \tau$在最坏情况下成立，我们得到 $\tau \leq \frac{|\mu_{max}-\mu_i|}{4}$ 。

如果组 i 直到 τ 达到 τ_0才从 G 中移除，那么 $\tau = \tau_0$。

因此，当停止添加样本时， α_i =最大值$\{\frac{|\mu_{max}-\mu_i|}{4}, \tau_0\}$ 是组 i 的置信区间半宽度。令 $\epsilon = \alpha_i$，我们得到

$$m_i = O\left(J_{up}^2 \frac{\log(\frac{g}{\eta}) + \log\log(\frac{1}{\alpha_i})}{\alpha_i^2}\right) \tag{20}$$

所有组的样本复杂度为

$$O\left(J_{up}^2 \sum_{i=1}^{g} \frac{\log(\frac{g}{\eta}) + \log\log(\frac{1}{\alpha_i})}{\alpha_i^2}\right). \tag{21}$$

□

我们可以从上述定理推知，样本量复杂度与最大组和其他组的大小差异密切相关。差异越大，所需的样本就越少。

4.1.2. 改进的基于采样的敏感度估计

在上一节中，我们介绍了用于多路连接查询中每个残差查询的估计算法。然而，要计算残差敏感度，需要计算 $\{q_E | E \subseteq nr - \{i\}, i \in P\}$ 中的所有残差结果。我们可以直接使用算法 2 来估计每个残差查询，但这会涉及冗余采样。如图3所示，对于连接路径 $a_1 \to b_2 \to c_2$，我们可以估计 $|t(a_1) \ltimes R_1 \ltimes R_2 \ltimes R_3| = \frac{1}{\frac{1}{3}\cdot\frac{1}{2}} = 6$和 $|t(b_1) \ltimes R_2 \ltimes R_3| = \frac{1}{\frac{1}{2}} = 2$。

算法 4改进版 RQE

输入： 多路查询 q，残差查询 $RQS = \{q_E | E \subseteq [n] - \{i\}, i \in P\}$，

终端误差界 τ_0

输出： 估计值 $\{T_E | q_E \in RQS\}$

$ARQS \leftarrow RQS$

当 $ARQS \neq \emptyset$ **时执行**

  $q_E \leftarrow$参数 $\max_{q_E \in ARQS} |E|$

  对于每个组 $i \in G_E$ **执行**

    执行随机游走 p，从 $t(v_i)$ 开始。

    对于每个 $E' \subseteq E$ 在 p **执行**

      $m_{E',i} \leftarrow m_{E',i} + 1$

      $n_{E'} \leftarrow n_{E'} + 1$

      $C_{E',\ i}, \quad \tau_{E',\ i} \leftarrow$ 估计$(p, t(v_i) \bowtie q'_E, m, C_{E',\ i})$

      $J_{E'}, \quad \tau_{J_{E'}} \leftarrow$ 估计$(p, q_{E'}, n, J)$
      $\tau_{E',i} \leftarrow \tau_{E',i} \cdot (J_{E'} + \tau_{J_{E'}})$

    结束循环

  结束循环

  对于每个 $q_E \in ARQS$ **执行**

    对于每个 $i \in G_E$ **执行**

      如果$C_{E,\ i} + \tau_{E,\ i} <$ MAX $_{j \in G_E} (C_{E,j} - \tau_{E,j})$ **然后**

        $G_E \leftarrow G_E - \{i\}$

      结束条件

    结束循环。

    如果$\tau_{E,\ i} <= \tau_0$ 对于每个组 $i \in G_E$ **那么**

      $ARQS \leftarrow ARQS - \{q_E\}$

    结束 if

  结束循环

结束循环

返回： $\{T_E = max_{i \in G_E} \ (C_{E,\ i} + \tau_{E,\ i}) \ | q_E \in RQS\}$

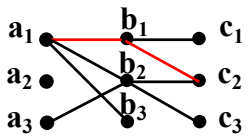


图3：一条连接路径上各值的估计值。

因此，我们提出一种算法，通过利用每条路径来估计该路径上的所有值的候选项，从而降低样本复杂度。

伪代码见算法4。由于基本思路是估计路径上每个值的连接大小以降低样本复杂度，因此我们首先运行算法，从具有最多关系的残差查询中抽取样本。因为较长的路径能够为更多值提供估计值。随后，算法为 q_E 中的组抽取新样本，以更新每个 $E' \subseteq E$ 的估计值和置信区间（第4–13行）。如果某个组不是候选的最大组，算法便迭代地将其从每个 $q_E \in ARQS$ 的组集 G_E 中移除（第15–19行）；一旦 q_E 中所有剩余组都得到充分估计，算法就会将 q_E 从 $ARQS$ 中移除（第20–22行）。当活动残差查询集（$ARQS$）为空时，算法停止。

在该算法中，可以使用一条连接路径来估计路径上每个值的连接大小。因此，可以降低样本复杂度。

4.2. 基于草图的敏感度估计

草图是有用的数据流摘要，广泛用于频数估计、重频项查找和连接大小估计。在本节中，我们提出一种基于 AGMS 草图的敏感度估计方法。

4.2.1. 基于草图的多重连接大小估计

AGMS 的基本思想是将关系 R 中连接属性 A 的值 v_1、v_2、……、$v_{|dom(A)|}$ 映射为四两两独立的随机变量 $\xi(v_1)$、$\xi(v_2)$、……、$\xi(v_{|dom(A)|})$，其中每个 $\xi(v_i) \in \{-1, +1\}$，且 $\Pr[\xi(v_i) = +1] = \Pr[\xi(v_i) =$

$-1] = 1$。关系 R 的 AGMS 草图为

$$sk(R) = \sum_{i \in dom(A)} f(i)\xi(v_i), \tag{22}$$

其中， $f(i)$ 是 v_i 的频数。 $sk(R)$ 可以通过对 R 进行单遍扫描来计算。对于两个关系 R_1 和 R_2 ，其草图 $sk(R_1)$ 与 $sk(R_2)$ 的乘积是 $R_1 \bowtie R_2$ 的连接大小的无偏估计： $\mathbb{E}[sk(R_1) \cdot sk(R_2)] = |R_1 \bowtie R_2|$。

多连接查询中的每个关系都可能包含多个连接属性。对于每一对等值连接属性，定义一个不同的随机族 $\xi_1, \xi_2, ..., \xi_n$ ，即可使用 AGMS 估计多连接大小。每个关系 R 的草图可写为:

$$sk(R) = \sum_{t \in R} \prod_{i \in JA} \xi_i(t[i]), \tag{23}$$

其中， JA 是所有连接属性对的集合， $t[i]$ 是元组 t 的属性 i 的值。我们用以下示例说明如何计算图 1 中各关系的草图。

示例 3. 考虑图 *1* 中的示例，我们为连接属性 B 和 C 定义两组四重独立随机变量，分别为 ξ_1 和 ξ_2。针对 $R_1(A, B)$、 $R_2(B, C)$ 和 $R_3(C, D)$ 构造了三个独立的草图:

$sk(R_1) = \sum_{t \in R_1} \xi_1(t[B])$,
$sk(R_2) = \sum_{t \in R_2} \xi_1(t[B]) \cdot \xi_2(t[C])$,
$sk(R_3) = \sum_{t \in R_3} \xi_2(t[C])$.
$X = sk(R_1) \cdot sk(R_2) \cdot sk(R_3)$ 的值给出了 $R_1 \bowtie R_2 \bowtie R3$ 的无偏估计。

尽管单个估计的准确性不足，但增强技术可以通过对多个独立估计进行平均和中位数选择，进一步提高准确性。最终增强后的估计是 s_2变量 Y_1,..., Y_{s2} 的中位数，其中每个 Y_i 都是 s_1 个独立估计的均值

X_1,..., X_{s1}。为简化表达，以下部分中，我们仅用 $s1$ 个独立估计的均值来表示基于 AGMS 草图的连接大小估计。

4.2.2. 多连接查询的草图敏感度

我们将多连接查询的草图敏感度 定义为 $sk^{(k)}(R_i)$，即与数据库相距 k 的关系 R_i 的草图。并建立局部敏感度与草图敏感度之间的联系。

我们首先考虑如何根据草图估计局部敏感度，如下所示。

定理 5。 局部敏感度可计算如下：

$$LS_q(I) = \max_{i \in P} \max_{v \in dom(JA(R_i))} (v \ltimes_{j \in [r]-\{i\}} R_{j,I}) \tag{24}$$

其中 $R_{j,\ I}$ 是实例 I中的第 j个关系。

根据以下在参考文献 [19], 中证明的定理，基于草图的连接大小估计值的相对误差可以得到限制。

定理 6。 设 Q 是定义在关系 R_1、.....R_r上的无环多连接查询，满足 $Count(Q) \geq L$ 且 $Self\text{-}Join(sk_k) \leq U_k$。那么，使用大小为$O(\frac{2^{2n}(\prod_{k=1}^{2} U_k) log(1/\eta)}{L^2\tau^2} \sum_{j=1}^{n} \log |dom(A_j)|)$的草图，可以对 $Count(Q)$进行近似，使得估计值的相对误差至多为 τ，且概率至少为$1-\eta$。

我们通过将估计值除以 $(1-\tau)$ 来计算连接大小的上界，因此，以至少 $1-\eta$ 的概率，真实连接大小 $J < \frac{est}{1-\tau}$。

公式 24 中的每个 $v \ltimes_{j \in [n]-\{i\}} R_{j,\ I}$ 都可以根据每个关系的草图进行估计：

$$\begin{aligned} & (v \ltimes_{j \in [n]-\{i\}} R_j) \\ & \leq \frac{\underset{s \in [1,s1]}{mean} \left(\prod_{l \in JA(R_{i,I})} \xi_{l,s}(v) \cdot \prod_{j \in [n]-\{i\}} sk_s(R_{j,I}) \right)}{1-\tau} \\ & \leq \frac{\underset{s \in [1,s1]}{mean} \left| \prod_{j \in [n]-\{i\}} sk_s(R_{j,I}) \right|}{1-\tau} \end{aligned} \tag{25}$$

其概率至少为 $1-\eta$。公式 25 中的第二个不等式成立，因为 $\prod_{l\in JA(R_i)}$ $\xi_{l,s}(v) \in \{-1,1\}$。

由于局部敏感度的平滑上界是根据距离为 k 时的局部敏感度定义的，我们也考虑根据草图对其进行估计。

定理 7。假设 $\mathcal{I}^k = \{I' : d(I,I') = k\}$ 是与 I 的距离为 k 的实例集合，则距离为 k 时的局部敏感度可计算如下：

$$LS_q^{(k)}(I) = \max_{I'\in\mathcal{I}^k} \max_{i\in P} \max_{v\in dom(JA(R_i))} (v \ltimes_{j\in[r]-\{i\}} R_{j,I'}) \tag{26}$$

令 $S^k = \{(k_1, k_2, ..., k_r) | \sum k_i = k, k_{x\notin P} = 0)\}$ 为 k 个元组的所有划分的集合，则 $LS_q^{(k)}(I)$ 可改写如下。

$$LS_q^{(k)}(I) = \max_{(k_1,k_2,...k_r)\in S^k} \max_{i\in P} \max_{v\in dom(JA(R_i))} (v \ltimes_{j\in[r]-\{i\}} R_j^{(k_j)}) \tag{27}$$

每个 $v \ltimes_{j\in[n]-\{i\}} R_j^{(k_j)}$ 都可以根据各个关系的草图进行估计：

$$\begin{aligned} &(v \ltimes_{j\in[n]-\{i\}} R_j) \\ &<= \frac{\underset{s\in[1,s1]}{mean}\left(\prod_{l\in JA(R_i)} \xi_{l,s}(v)\cdot \prod_{j\in[n]-\{i\}} sk_s^{(k_j)}(R_j)\right)}{1-\tau} \\ &<= \frac{\underset{s\in[1,s1]}{mean}\left|\prod_{j\in[n]-\{i\}} sk_s^{(k_j)}(R_j)\right|}{1-\tau} \end{aligned} \tag{28}$$

因此，

$$LS_q^{(k)}(I) \le \max_{I'\in\mathcal{I}^k} \max_{i\in P} \frac{\underset{s\in[1,s1]}{mean}\left|\prod_{j\in[n]-\{i\}} sk_s^{(k_j)}(R_j)\right|}{1-\tau} \tag{29}$$

我们可以将上述公式的右侧视为$LS_q^{(k)}(I)$的上界，并将草图敏感度定义如下。

与弹性敏感度和残差敏感度类似，我们的**草图敏感度（SKS）** 必须先使用平滑敏感度进行平滑，之后才能与拉普拉斯机制一起使用。

$$SKS = \max_{k>0} e^{-\beta k} \hat{LS}_q^{(k)}, \tag{30}$$

其中

$$\hat{LS}_q^{(k)} = \max_{I' \in \mathcal{I}^k} \max_{i \in P} \frac{\underset{s \in [1,s1]}{mean} \left| \prod_{j \in [n]-\{i\}} sk_s^{(k_j)}(R_j) \right|}{1-\tau} \tag{31}$$

我们用以下示例说明如何计算草图敏感度。

示例 4. 考虑图 *1* 中的示例数据库 I，我们为连接属性 B 和C 定义两组四两两独立随机变量，分别为 ξ_1 和 ξ_2。我们可以按如下方式为 R_1、 R_2和 R_3 构造草图。$sk(R_1) = \sum_{t \in R_1} \xi_1(t[B])$， $sk(R_2) = \sum_{t \in R_2} \xi_1(t[B]) \cdot \xi_2(t[C])$，以及 $sk(R_3) = \sum_{t \in R_3} \xi_2(t[C])$。

向数据库 I' 添加一个元组的影响可以通过以下方式估计，其中该数据库与原始数据库的距离为 k：

$$\begin{aligned} \hat{LS}_q^{(k)} = \max_{\sum_{j=1}^{3} |k_j| = k} \max \{ & \underset{s \in [1,s1]}{mean} \left| sk_s^{(k_1)}(R_1) \cdot sk_s^{(k_2)}(R_2) \right|, \\ & \underset{s \in [1,s1]}{mean} \left| sk_s^{(k_1)}(R_1) \cdot sk_s^{(k_3)}(R_3) \right|, \\ & \underset{s \in [1,s1]}{mean} \left| sk_s^{(k_2)}(R_2) \cdot sk_s^{(k_3)}(R_3) \right| \} \cdot \frac{1}{1-\tau} \end{aligned} \tag{32}$$

其中 $sk^{(k_j)_s}(R_j) = sk_s(R_j) + k_j$。然后，可以根据公式 *30* 计算 $R_1 \bowtie R_2 \bowtie R_3$ 的草图敏感度。

5. 实验

我们设计实验来验证所提方法的有效性和效率。在第 5.1 节中，我们介绍实验设置，包括实验涉及的硬件、数据集、查询、对比方法、误差指标和一些参数。在第 5.2 节中，我们将所提方法与 RS 和 ES 的准确性进行比较。此外，我们还在第 5.3 节中验证所提方法实现差分隐私保护的效率。在第 5.4 节中，我们测试不同参数（如隐私预算和采样率）的影响。最后，在第 5.5 节中，我们简要总结实验结果。

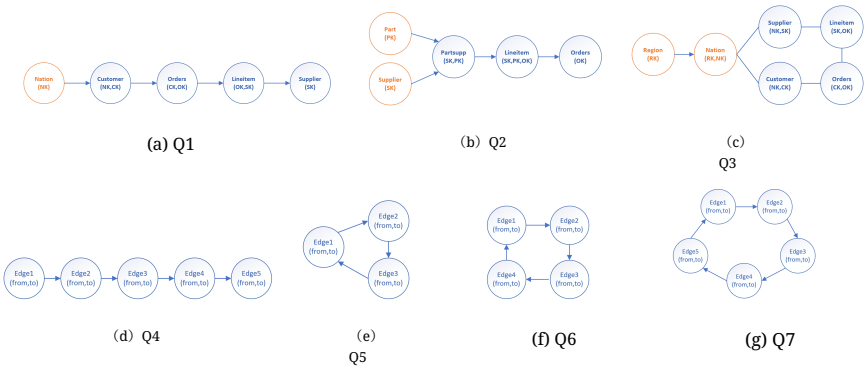


图4：查询。

5.1. 实验设置

硬件和软件库

所有这些实验均在一台配备 256 GB内存的机器上运行，操作系统为 Ubuntu(20.04.1)，并使用 PostgreSQL(14.5) 和 Python 3.9。

数据集

我们基于两个数据集测试我们的方法，包括 TPC-H 和 Facebook自我网络。

TPC-H 数据集[1] TPC-H 数据集提供了一组模式，用于支持 TPC 基准测试 H（TPC-H）。它用于衡量高度复杂的决策支持数据库的性能。该数据集包含许多仅在数据量上有所不同的模式。TPC-H 模式由八个表组成，每个表的列数和行数各不相同。这些表分别命名为 nation、region、part、customer、lineitem、orders、partsupp 和 supplier。该模式旨在代表现实世界中的决策支持系统。根据每个表所含数据及其尺度，我们将前三个表视为公共关系，其余五个表视为私有关系。此外，为方便起见，我们从数据集中提取连接条件所需的关键属性，并据此创建了对应的八个表。

[1]https://www.tpc.org/tpch/

Facebook自我网络数据集。[2] 由于 TPC-H 数据集的分布均匀，因此我们引入了另一个具有偏斜性的数据集。Facebook自我网络数据集包含 4,039 个节点和 176,467 条边。如[2], 所示，我们采用类似的方法将该数据集合并为 5 个关系，每个关系都包含两个属性，例如 from 和 to。

查询

实验所使用的查询如图4所示。我们对 TPC-H 数据集使用 Q1-Q3，对 Facebook 数据集使用 Q4-Q7。这些查询包括链式查询、无环查询和有环查询。橙色圆圈表示非私有关系，蓝色圆圈表示私有关系。

对比方法

在以下实验中，我们的方法称为“Sampling-SE”和“Sketch-SE”，分别代表基于采样的敏感度估计和基于草图的敏感度估计。我们方法的对比方法包括 ES 和 RS。

（1）ES：弹性敏感度是局部敏感度的平滑上界。ES 根据每个关系中最频繁连接属性的频数计算。

（2）RS：残差敏感度是比 ES 更紧的平滑上界。它根据残差查询的最大边界计算。

误差指标

偏差(DE)：

$DE = |TrueResult - EstimateResult|$。

参数

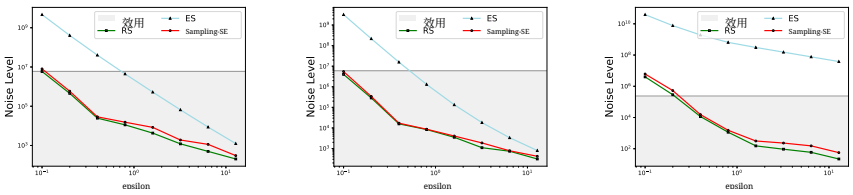
ϵ：隐私预算，表示隐私保护的级别。

δ：(ϵ， δ)-DP 中的一个新隐私参数，表示满足纯差分隐私要求的失败概率。

r：采样率。

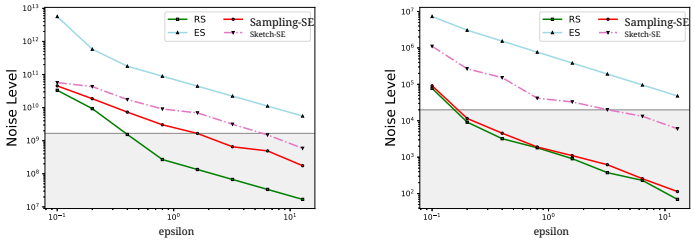
[2]https://snap.stanford.edu/data/ego-Facebook.html

5.2. 准确性



(a) Q1 (b) Q2 (c) Q3

图 5： ϵ 对 TPC-H 数据集的影响。



(a) Q4 (b) Q5

(c) Q6 (d) Q7

图 6： ϵ 对 Facebook 数据集的影响。

在接下来的实验中，我们的方法分别称为“Sampling-SE”和“Sketch-SE”，即基于采样的敏感度估计和基于草图的敏感度估计。我们在 Facebook数据集上比较 Sampling-SE 和 Sketch-SE 与 ES、RS 的准确性。在 TPC-H 数据集上，我们仅比较 Sampling-SE 与 ES、RS。因为

TPC-H 的结果涉及过多的不同值，因此基于草图的方法不适用于域较大的连接结果。我们测试了不同方法在 ϵin $[0.1、0.2、0.4、0.8、1.6、3.2、6.4、12.8]$ 下的噪声水平。较小的 ϵ 可带来更强的隐私保护和更多噪声。我们将 Sampling-SE 的采样率设为 0.0001，并将 Sketch-SE 中每个关系的估计器数量设为 100,000。

TPC-H（1GB）和 Facebook 上的实验结果分别如图 5 和图 6 所示。误差以相对于真实查询结果的偏差来衡量，即所添加的噪声。在这些图中，我们还绘出了实际查询结果，即图中的阴影区域。显然，低于查询结果的噪声水平值被视为具有效用，而高于该值的噪声水平则相应地被视为不具有效用。

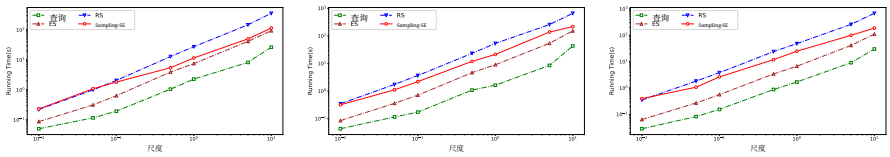
从图中可以看出，我们的方法 Sampling-SE 的噪声水平较低，与 RS 相近，且二者都远低于按照 ES 计算出的噪声水平，这进一步凸显了我们基于采样的估计策略的优势。图 6 展示了 Facebook数据集上的噪声水平。Sketch-SE 的噪声水平高于 Sampling-SE，但低于 ES。随着查询中涉及的关系数量增加，连接结果的域也随之扩大，因此要提供足够准确的结果，需要大量估计器。因此，当连接查询涉及许多关系且不同的连接值数量较大时，Sketch-SE 的准确性不如 Sampling-SE。不过，如图 6 所示，Sketch-SE 的准确性仍高于 ES。

5.3. 效率

在本 part 中，我们将本方法与 RS 和 ES 进行比较，以验证其效率。我们通过提交查询到返回加噪查询结果之间所用的时间来衡量效率，该时间主要由两部分组成：回答查询所需的时间，以及根据差分隐私机制计算敏感度并添加噪声所需的时间。计算 ES 中每个连接属性的最大频数所需的时间，以及

为每个关系构建草图所需的时间未计入图中，因为它们是在离线状态下计算的。

图7展示了在 TPC-H 数据集上，不同查询和数据规模下 RS 和 ES 的运行时间，其中 ϵ=0.8 适用于所有三个查询，而 $\delta = 10^{-7}$，2×10^{-8}，10^{-8}，2×10^{-9}，10^{-9}，2×10^{-10}，10^{-10} 分别表示各查询在 0.01G、0.05G、0.1G、0.5G、1G、5G 和 10G 等不同数据规模下的情况。显然，Sampling-SE 的时间开销始终介于 RS 和 ES 之间。这种效果在数据规模较大时尤为明显。这种差异主要源于两个原因：一方面，Sampling-SE 通过采样降低了计算大量残差查询的成本；另一方面，一些包含主外键连接的复杂残差查询可以简化为查找某个关系的最大频次值，从而进一步降低成本。



(a) Q1 (b) Q2 (c) Q3

图7：TPC-H 上查询的运行时间。

我们还比较了 Q4 到 Q7 上不同方法的效率，结果如图8所示。我们设置了参数 $\epsilon = 0.8$ 和 $\delta = 10^{-7}$。对于我们的 Sampling-SE 方法，我们将采样率固定为 0.0001。

可以看出，在大多数情况下，Sampling-SE 在 Facebook数据集 上也表现良好。由于该查询仅涉及三个较小的关系，而计算 Q5 的 RS 本身耗时很短，因此 Sampling-SE 在 Q5 上的效率略低于 RS。但对于更大的数据规模和更复杂的连接条件，我们的 Sampling-SE 方法表现要好得多，效用提升也更加显著。由于 ES 的统计量和 Sketch-SE 的草图都是离线计算的，因此它们的在线效率远高于其他方法。

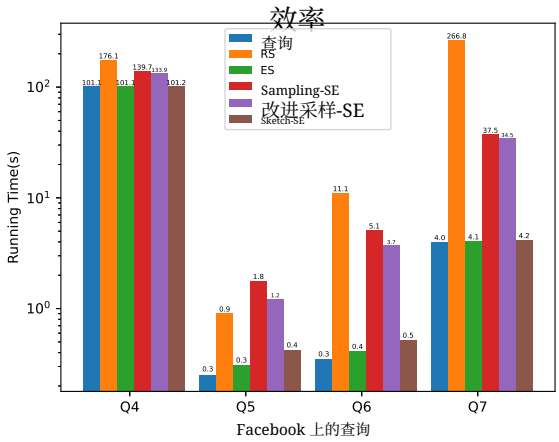


图 8：Facebook 上查询的运行时间。

5.4. 参数的影响

样本量的影响 我们的 Sampling-SE 方法的关键因素是

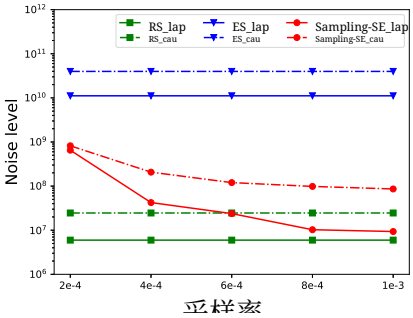


图 9：不同采样率下的噪声水平。

采样率。我们进行了实验，以观察采样率 r 对噪声结果的影响。为简化实验，我们选取两个数据集以及两个链式查询 Q1 和 Q4。对于参数，我们将 TPC-H 的尺度设为=1GB、 ϵ =6.4 和 $\delta = 2 \times 10^{-10}$，Facebook 的尺度设为 $\delta = 10^{-7}$。此外，我们还实现了包括 $LaplaceMechanism$ [11] 和 $GeneralCauchyMechanism$ [14] 在内的不同噪声机制，以验证结果。显

显著结果如图9和图10所示。

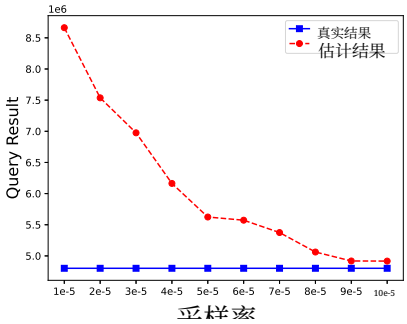


图 10：不同采样率下的估计结果。

如图9所示，采用不同差分隐私机制的 Sampling-SE 比 ES 准确得多。随着采样大小增加，Sampling-ES 的噪声水平越来越接近准确结果。此外，对于这两种基线方法和我们的方法，拉普拉斯机制下的结果始终小于GeneralCauchy机制下的结果。

为进一步验证该采样方法的有效性，我们考虑一个残差查询：“select max(count) from (select edge2 from, count(*) from edge2, edge3, edge4, edge5 where edge2 to=edge3 from and edge3 to=edge4 from and edge4 to=edge5 from group by edge2 from) as t;”，这也是 Sampling-SE 针对的典型查询结构。图10展示了采样率从 10^{-5} 到 10^{-4}变化时查询估计结果的变化。可以观察到，即使采样率相对较低，我们也能获得准确性较高的结果。

由于差分隐私的噪声与隐私预算 ϵ 相关，且预算值越小，噪声变化越大。因此，我们测试了采样率对敏感度的影响，以消除添加到数据中的差分隐私噪声所带来的干扰。我们采用了前面 Q4 实验中所示的相同条件。这里，我们也使用拉普拉斯机制，在 Q4 上将隐私预算设为 ϵ =6.4 和 $\delta = 10^{-7}$ 。实验结果如图11所示。为处理包含过多

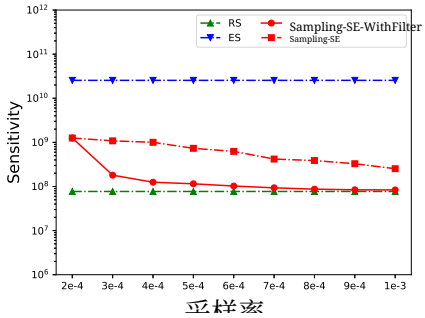
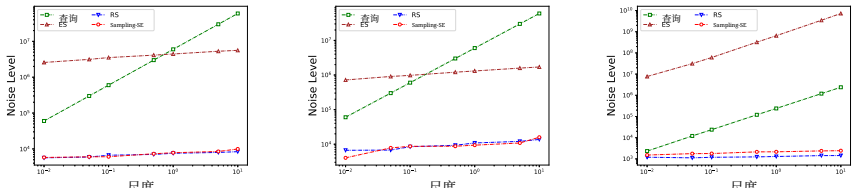


图11：不同采样率下的敏感度。

不同值的连接结果，并进一步提高 Sampling-SE算法的效率，我们采用了一种策略：估计查询结果中若干随机样本所包含的各组中的最大组大小，而不是估计每组的大小。如图11所示，我们将这一策略称为 Sampling-SE-WithFilter。可以观察到，随着采样率提高，Sampling-SE-WithFilter 比 Sampling-SE 收敛得更快。

数据规模的影响



（a）
Q1

（b）Q2

（c）
Q3

图 12：数据规模对噪声水平的影响。

在这一部分，我们评估了 TPC-H 数据规模对噪声水平的影响。我们使用规模因子从 0.01 到 10 的 TPC-H 数据集，并通过 Q1、Q2 和 Q3 测试不同方法的性能。对于我们的方法，我们将采样率固定为 0.0001。从图 12 可以看出，随着规模增大，我们的 Sample-SE 和 RS 产生的波动都较小，这意味着我们的方法适用于大型数据集。值得注意的是，最大

Q1 和 Q2 的边界在使用 Sampling-SE 方法后并未改变，因此这两个查询的噪声水平与原始 RS 结果几乎相同。我们将这些噪声水平与查询答案进行比较，以验证添加噪声后的效用；噪声水平高于原始查询结果时，原始查询结果的信息会被完全掩盖。特别是，RS 和 Sampling-SE 的噪声水平始终低于查询答案。

5.5. 实验结果总结

实验结果总结如下：

- 对于规模较大的数据集上的连接查询，Sampling-SE 和 Sketch-SE 的效率均高于残差敏感度。

- 在采样率适当时，Sampling-SE 的准确性与残差敏感度相当，同时时间开销更低。

- Sketch-SE 的效率与弹性敏感度相同，但其敏感度值相对较低，因此准确性更高。

6. 结论

本文提出了两种分别基于采样和草图的多重连接查询敏感度估计方法。基于采样的方法能够以更高的效率估计出与残差敏感度相近的敏感度；基于草图的方法则能以与弹性敏感度相近的效率，获得更准确的敏感度估计。未来，我们将把敏感度估计方法的思路应用于更多类型的复杂查询，包括包含复杂谓词和用户定义函数的查询。

致谢

本研究得到国家自然科学基金项目（62202113）、SDGC2229、SL2022A04J01306 以及 PCL 重大重点项目（项目编号：PCL2021A09、PCL2021A02、PCL2022A03）的资助。

参考文献

[1] N. M. Johnson、J. P. Near、D. X. Song，《面向 SQL 查询的实用差分隐私》，VLDB基金会会刊 11（2017年），526–539。[2] W. Dong、K. Yi，《差分隐私多路连接的残差敏感度》，2021年数据管理国际会议论文集（2021）。[3] C. Dwork，《差分隐私》，载于《密码学与安全百科全书》，2006。[4]S. Ayd¨ore、W. Brown、M. Kearns、K. Kenthapadi、L. Melis、A. Roth、A. Siva，《通过自适应投影实现差分隐私查询发布》，载于国际机器学习会议，2021。[5] T. Wang、J. Q. Chen、Z. Zhang、D. Su、Y. Cheng、Z. Li、N. Li、S. Jha，《在集中式和本地差分隐私下持续发布数据流》，2021年 ACM SIGSAC 计算机与通信安全会议论文集（2020）。[6] M. Maruseac、G. Ghinita，《高置信度关联规则的精度增强型差分隐私数据挖掘》，IEEE 可靠与安全计算汇刊 17（2020），1297–1309。[7] T. Wang、N. Li、S. Jha，《本地差分隐私频繁项集数据挖掘》，2018年 IEEE 安全与隐私研讨会（SP）（2018），127–143。[8] A. Triastcyn、B. Faltings，《面向机器学习的贝叶斯差分隐私》，载于国际机器学习会议，2019。

[9] H. Zheng、Q. Ye、H. Hu、C. Fang、J. Shi，利用差分隐私扰动保护机器学习模型的决策边界，《IEEE Transactions on Dependable and Secure Computing》19（2020）2007–2022。

[10] H. Jiang、J. Pei、D. Yu、J. Yu、B. Gong、X. Cheng，社交网络分析中的差分隐私应用：综述，《IEEE Transactions on Knowledge and Data Engineering》35（2023）108–127。[11] C. Dwork、F. McSherry、K. Nissim、A. D. Smith，在隐私数据分析中根据敏感度校准噪声，载于：《密码学理论会议》，2006。[12] F. McSherry，隐私集成查询：用于隐私保护数据分析的可扩展平台，《2009年ACM SIGMOD国际数据管理会议论文集》（2009）。[13] D. Proserpio、S. Goldberg、F. McSherry，在隐私数据分析中根据敏感度校准数据，《VLDB基金会会刊》7（2012）637–648。[14] K. Nissim、S. Raskhodnikova、A. D. Smith，隐私数据分析中的平滑敏感度与采样，载于：《计算理论研讨会》，2007年。[15] S. Chaudhuri、B. Ding、S. Kandula，近似查询处理：没有万能方案，《2017年ACM数据管理国际会议论文集》（2017年）。[16] P. J. Haas、J. M. Hellerstein，用于在线聚合的Ripple join，载于：ACM SIGMOD会议，1999年。[17] F. Li、B. Wu、K. Yi、Z. Zhao，Wander Join：通过随机游走实现在线聚合，《2016年国际数据管理会议论文集》（2016年）。[18] Z. Zhao、R. Christensen、F. Li、X. Hu、K. Yi，重新审视连接上的随机抽样，《2018年国际数据管理会议论文集》（2018年）。[19] A. Dobra、M. N. Garofalakis、J. Gehrke、R. Rastogi，处理数据流上的复杂聚合查询，载于：M. J. Franklin、B. Moon、A. Ailamaki

（编），《2002年ACM SIGMOD数据管理国际会议论文集》，美国威斯康星州麦迪逊，2002年6月3日至6日，ACM，2002年，第61–72页。doi:10.1145/564691.564699。URL https://doi.org/10.1145/564691.564699

[20] M. Charikar、K. Chen、M. Farach-Colton，《在数据流中查找高频项》，载于：《自动机、语言与编程国际会议论文集》，Springer，2002年，第693–703页。

[21] G. Cormode、S. Muthukrishnan，《改进的数据流摘要：计数最小草图及其应用》，《算法学报》55（1）（2005）58–75。

[22] D. Vengerov、A. C. Menck、M. Za¨ıt、S. Chakkappen，《受过滤条件约束的连接大小估计》，《VLDB基金会会刊》8（12）（2015）1530–1541。doi:10.14778/2824032.2824051。URL http://www.vldb.org/pvldb/vol8/p1530-vengerov.pdf

[23] F. Li、B. Wu、K. Yi、Z. Zhao，Wander Join：通过随机¨游走实现在线聚合，载于：F. Ozcan、G. Koutrika、S. Madden（编），《2016年国际数据管理会议论文集》，SIGMOD 2016年会议，旧金山，加利福尼亚州，美国，2016年6月26日至7月1日，ACM，2016年，第615–629页。doi:10.1145/2882903.2915235。URL https://doi.org/10.1145/2882903.2915235

[24] A. Kim、E. Blais、A. G. Parameswaran、P. Indyk、S. Madden、R. Rubinfeld，具有排序保证的可视化快速采样，《VLDB基金会会刊》8（5）（2015）521–532。doi:10.14778/2735479.2735485。URL http://www.vldb.org/pvldb/vol8/p521-kim.pdf

[25] W. Hoeffding，《有界随机变量和的概率不等式》作者：来源：《美国统计学会会刊》，第58卷，第301期（1963年3月），第13页起。出版方：美国统计学会。稳定链接，见《统计科学百科全书》。