DPSQL 差分隐私算法

名词解释

- 隐私数据 private data
  - 隐私保护的目标，防止被 re-identification，一般保护实体为用户、设备、公司等
- 隐私机制 privacy mechanism
  - 一个能够处理隐私数据并输出具有隐私保护结果的程序
  - 这种机制一般都是相对标准的，如 laplace、gaussian 等
  - 这种机制可以处理聚合结果比如 count、sum，也有的能处理一个单独的数据对象，比如梯度
  - 隐私机制每次执行会产生一定的 privacy budget cost
- epsilon
  - epsilon 是衡量隐私泄漏的标准
  - epsilon 越小表示隐私泄漏越少，同时会引入更多的噪音。
- delta
  - 表示结果不能受到 epsilon 保护的概率，一般设置的非常小(如$\frac{1}{n^2}$，其中n代表数据集的条目数)
- 隐私预算 privacy budget
  - 当允许多次查询时，可以将多次查询的 noisy 结果结合起来进行 re-identify，因此除了衡量每次的隐私损失外，还需要衡量多次查询后的总的隐私损失
  - 隐私预算是指在隐私数据不能再被访问之前，所有查询可能发生的隐私损失

总量，是一个预期的上限值

- 组合 composition
  - 描述多次查询的隐私消耗如何组合在一起得到总的消耗情况
- 敏感度 sensitivity
  - 在数据集中任意删除一条数据，对特定查询f结果的最大改变。
  - 敏感度是决定加入噪音量大小的关键参数

DP 机制

机制	Laplace	Gaussian
定义	$Lap(b)$	$\mathcal{N}(0,\sigma^2)$
参数	$b=\frac{\Delta f}{\varepsilon}$	$\sigma=\frac{\Delta f\sqrt{2log(1.25/\delta)}}{\varepsilon}$
DP	ε-DP	(ε,δ)-DP
方差	$\frac{2\Delta f^2}{\varepsilon^2}=O\left(\frac{1}{\varepsilon^2}\right)$	$\frac{2\Delta f^2}{\varepsilon^2}\log\frac{1.25}{\delta}=O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$
(α, β)	$\frac{\Delta f}{\varepsilon}\log\frac{1}{\beta}$	$\frac{2\Delta f}{\varepsilon}\sqrt{log\frac{1.25}{\delta}}erf^{-1}(1-\beta)\leq\frac{2\Delta f}{\varepsilon}\sqrt{\log\frac{1}{\beta}\log\frac{1.25}{\delta}}$

1. Laplace 机制介绍

- Laplace 分布
  - 概率密度函数$Lap(x\mid b)=\frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$
- Laplace 机制
  - 加噪服从分布$Lap(b)$，scale 参数$b=\Delta f/\varepsilon$，满足ϵ-DP
- 可用性

- 方差,$2b^2 = \frac{2\Delta f^2}{\varepsilon^2} = O\left(\frac{1}{\varepsilon^2}\right)$
- (α, β)-accuracy, $\alpha = \frac{\Delta f}{\varepsilon}\log\frac{1}{\beta}$

可用性分析：

Laplace 机制添加噪音N_L服从分布$Lap\left(b = \frac{\Delta f}{\varepsilon}\right)$

Laplace 分布满足 $P[|N_L| \ge t * b] = \exp(-t)$，将$\exp(-t)$替换成 β，则有

$P\left[|N_L| \ge \frac{\Delta f}{\varepsilon}\log\frac{1}{\beta}\right] = \beta$.

Gauss 机制介绍

- Gauss 分布
  - 概率密度函数 $\mathcal{N}(x|\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)$
- Gauss 机制
  - 加噪服从分布$\mathcal{N}(0, \sigma^2)$，标准差参数$\sigma = \frac{\Delta f\sqrt{2log(1.25/\delta)}}{\varepsilon}$，满足$(\varepsilon, \delta)$-DP.
- 可用性
  - 方差，$\sigma^2 = \frac{2\Delta f^2\log(1.25/\delta)}{\varepsilon^2} = O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$
  - (α, β)-accuracy, $\alpha = \frac{2\Delta f}{\varepsilon}\sqrt{log\frac{1.25}{\delta}}erf^{-1}(1-\beta) \le \frac{2\Delta f}{\varepsilon}\sqrt{\log\frac{1}{\beta}\log\frac{1.25}{\delta}}$

可用性分析：

Gauss 机制添加噪音 N_G 服从分布 $\mathcal{N}(0, \sigma^2)$。

Gauss 分布的逆累积分布函数 $F(x) = \frac{1}{2} + \frac{1}{2}erf\left(\frac{x}{\sigma\sqrt{2}}\right)$，其中 erf 是高斯分布的误差函数 error function: $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$.

令$F(\alpha) = 1 - \frac{\beta}{2}, F(\alpha) - F(-\alpha) = 1 - \beta$，则

$P[|N_G| \le \alpha] = 1 - \beta$，*其中*$\alpha = \frac{2\Delta f}{\varepsilon}\sqrt{log\frac{1.25}{\delta}}erf^{-1}(1-\beta) \le \frac{2\Delta f}{\varepsilon}\sqrt{\log\frac{1}{\beta}\log\frac{1.25}{\delta}}$

组合定理(k adaptive queries，查询之间存在关联)

- 若把 privacy loss 看作是一个随机变量：对于一个输出o，机制M在相邻数据集X, X'上的 privacy loss 是$L^{o}_{M(X),M(X')} = \ln \frac{Pr[M(X)=o]}{Pr[M(X')=o]}$
- (ε, δ)-DP 表示对于任意的相邻数据集，最多有δ的概率，这个随机变量的值大于ε.

Basic composition

- 一组k个满足$(\varepsilon_i, \delta_i)$-DP 的机制，对于同一数据集的进行$k$次查询，$k$-fold adaptive composition 满足$(\sum\varepsilon_i, \sum\delta_i)$-DP.
- Basic composition 是将 privacy loss 这个随机变量的上界累加（基于 KL 散度）

Advanced composition 1

- 一组k个满足(ε, δ)-DP 的机制，对于同一数据集的进行k次查询，k-fold adaptive composition 满足$(\varepsilon', k\delta + \delta')$-DP，其中$\varepsilon' = \varepsilon\sqrt{2k\ln(1/\delta')} + k\varepsilon(e^{\varepsilon} - 1)$
- 当$\varepsilon = 10^{-5}$的时候，后面一项 $k\varepsilon(e^{\varepsilon} - 1)$近似于 0，则$\varepsilon' = \varepsilon\sqrt{2k\ln 1/\delta'} = O(\sqrt{k})$,相比于 basic composition 的$O(k)$节省了更多隐私预算，因此 Advanced composition 1 在k较大时显著好于 basic composition。
- 该方法将 privacy loss 期望的上界累加（基于 max 散度）

Advanced composition 2

- 一组k个满足(ε, δ)-DP 的机制，对于同一数据集的进行k次查询，k-fold adaptive composition 满足$\left((k-2i)\varepsilon, 1-(1-\delta)^k(1-\delta_i)\right)$-DP
  - 其中对于任意的$i = \{0,1,\dots,\lfloor k/2 \rfloor\}$，$\delta_i = \frac{\sum_{l=0}^{i-1}\binom{k}{l}\left(e^{(k-l)\varepsilon} - e^{(k-2i+l)\varepsilon}\right)}{(1+e^{\varepsilon})^k}$
- 给出了一个 tighter bound，依赖于 operational interpretation of the privacy as hypothesis testing，证明了这个结果的最优性 optimality

自适应选择：

- 一组k个满足(ε, δ)-DP 的机制，对于同一数据集的进行k次查询，k-fold adaptive composition 满足$\left(\varepsilon', 1-(1-\delta)^k(1-\delta_i)\right)$-DP，其中

$$\varepsilon' = min\left\{k\varepsilon, \frac{\varepsilon k(e^{\varepsilon}-1)}{e^{\varepsilon}+1} + \varepsilon\sqrt{2k\ln(1/\delta')}, \frac{\varepsilon k(e^{\varepsilon}-1)}{e^{\varepsilon}+1} + \varepsilon\sqrt{2k\ln(e+\frac{\sqrt{k\varepsilon^2}}{\delta'})}\right\}$$

- 当 k 个机制分别满足$(\varepsilon_i, \delta_i)$-DP，则该组机制满足$\left(\varepsilon', 1-(1-\delta')\prod_{i=1}^{k}(1-\delta_i)\right)$-DP，其中$\varepsilon' =$

$$min\left\{\sum_{i=1}^{k}\varepsilon_i, \sqrt{\sum_{i=1}^{k}2\varepsilon_i^2\ln(1/\delta')} + \sum_{i=1}^{k}\frac{\varepsilon_i(e^{\varepsilon_i}-1)}{(e^{\varepsilon_i}+1)}, \sqrt{\sum_{i=1}^{k}2\varepsilon_i^2\ln\left(e+\frac{\sqrt{\sum_{i=1}^{k}\varepsilon^2}}{\delta'}\right)} + \sum_{i=1}^{k}\frac{\varepsilon_i(e^{\varepsilon_i}-1)}{(e^{\varepsilon_i}+1)}\right\}$$

参考文献：

The Composition Theorem for Differential Privacy

敏感度

全局敏感度 Global sensitivity

函数$f: D^n \to \mathbb{R}^d$，输入为一个数据集，输出为d维向量，对于任意相邻数据集$x, y \in D^n$，函数f的全局敏感度为

$$GS_f = \max_{x,y:d(x,y)=1}\|f(x)-f(y)\|_1$$

- 函数的全局敏感度由函数f本身决定

局部敏感度 Local sensitivity

函数$f: D^n \to \mathbb{R}^d$，输入一个数据集$x \in D^n$，输出d维向量，对于给定数据集x及其相邻数据集y，函数f在x上的局部敏感度为

$$LS_f(x) = \max_{y:d(x,y)=1}\|f(x)-f(y)\|_1$$

- 局部敏感度由函数f和给定数据集x中的具体数据共同决定。由于利用了数据集的数据分布特征， 局部敏感度通常要比全局敏感度小的多
- 与全局敏感度的关系：

$$GS_f = \max_x(LS_f(x))$$

- 局部敏感度在一定程度上体现了数据集的分布特征，如果直接应用局部敏感度来计算噪音量则会泄露数据集的敏感信息，因此局部敏感度的平滑上界被用来与局部敏

感度一起确定噪音大小

平滑敏感度 Smooth sensitivity

参考文献：smooth sensitivity

- LS 的平滑上界
  - 给定数据集$x \in D^n$及任意相邻数据集y: $d(x, y) = 1$，函数f局部敏感度为$LS_f(x)$

  对于$\beta > 0$，若函数$S: D^n \to \mathbb{R}^d$满足$S_f(x) \geq LS_f(x), S_f(x) \leq e^{\beta} S_f(y)$,则称$S$为$f$的局部敏感度的$\beta$-平滑上界
- 平滑敏感度
  - 给定数据集x, y,

$$S_{f,\beta}(x) = \max_{y \in D^n}\left(e^{-\beta d(x,y)} LS_f(y)\right)$$

- 平滑敏感度的计算

  定义在距离k上的敏感度

$$A_f^{(k)}(x) = \max_{y \in D^n : d(x,y) \leq k} LS_f(y)$$

  - $S_{f,\beta}(x) = \max_{k=0,1,\ldots,n} e^{-\beta k}\left(\max_{y : d(x,y)=k} LS_f(y)\right) = \max_{k=0,1,\ldots,n} e^{-\beta k} A_f^{(k)}(x)$

弹性敏感度 Elastic sensitivity

参考文献：elastic sensitivity

- Elastic Sensitivity is an Upper Bound on Local Sensitivity
- 定义
  - Elastic Sensitivity: 查询q在与真实数据库x距离k上的弹性敏感度$\hat{S}^{(k)}(q, x)$
  - Elastic stability: relational transformation 关系变换r的弹性稳定度 $\hat{S}_R^{(k)}(r, x)$
  - Maximum frequency: 数据库x中关系r中属性a的频率最高值的频率$mf(a, r, x)$
- Maximum frequency at distance k: 与真实数据库x距离k，关系r中属性a的频率

最高值的频率$mf_k(a,r,x)$，其中

$$mf_k(a,r,y) \geq \max_{y:d(x,y)\leq k} mf(a,r,y)$$

- 计算弹性敏感度

Maximum frequency	1. $mf_k(a,t,x) = mf(a,t,x) + k$ 2. $mf_k\left(a_1, r_1 \underset{a_2=a_3}{\bowtie} r_2, x\right) = mf_k(a_1,r_1,x)mf_k(a_3,r_2,x) \quad a_1 \in r_1$ $mf_k\left(a_1, r_1 \underset{a_2=a_3}{\bowtie} r_2, x\right) = mf_k(a_1,r_2,x)mf_k(a_2,r_1,x) \quad a_1 \in r_2$ 3. $mf_k\left(a, \Pi_{a_1,\dots,a_n} r, x\right) = mf_k(a,r,x)$ 4. $mf_k\left(a, \sigma_\varphi r, x\right) = mf_k(a,r,x)$ 5. $mf_k(a, Count(r), x) = \perp$
Elastic stability	1. $\hat{S}_R^{(k)}(t,x) = 1$ 2. $\hat{S}_R^{(k)}\left(r_1 \underset{a=b}{\bowtie} r_2, x\right) = \max\left(mf_k(a,r_1,x)\hat{S}_R^{(k)}(r_2,x), mf_k(b,r_2,x)\hat{S}_R^{(k)}(r_1,x)\right)$, $\lvert\mathcal{A}(r_1) \cap \mathcal{A}(r_2)\rvert = 0$ $\hat{S}_R^{(k)}\left(r_1 \underset{a=b}{\bowtie} r_2, x\right) = mf_k(a,r_1,x)\hat{S}_R^{(k)}(r_2,x) + mf_k(b,r_2,x)\hat{S}_R^{(k)}(r_1,x) + \hat{S}_R^{(k)}(r_1,x)\hat{S}_R^{(k)}(r_2,x)$, $\lvert\mathcal{A}(r_1) \cap \mathcal{A}(r_2)\rvert > 0$ 3. $\hat{S}_R^{(k)}\left(\Pi_{a_1,\dots,a_n} r, x\right) = \hat{S}_R^{(k)}(r,x)$ 4. $\hat{S}_R^{(k)}(\sigma_\varphi r, x) = \hat{S}_R^{(k)}(r,x)$ 5. $\hat{S}_R^{(k)}\left(Count(r)\right) = 1$
Elastic sensitivity	1. $\hat{S}^{(k)}(Count(r), x) = \hat{S}_R^{(k)}(r,x)$ 2. $\hat{S}^{(k)}\left(\underset{G_1,\dots,G_n}{Count}(r), x\right) = 2\hat{S}_R^{(k)}(r,x)$

- Smooth the elastic sensitivity
  - $S = \max_{k=0,1,\dots,n} e^{-\beta k}\hat{S}^{(K)}, \beta = \frac{\varepsilon}{2\ln(2/\delta)}, \quad \delta = 10^{-8}$
  - Laplace 机制的 scale 参数$b = \frac{2S}{\varepsilon}$