分布 $\pi(\Delta fc/\max\{r_i^{\max}\}^n\varepsilon)$.

下面结合矩阵加运算、SNP 连锁不平衡下的差分隐私定义和模余运算,给出矩阵差分隐私的定义.

定义 2(矩阵差分隐私). 给定$\varepsilon\geqslant 0$,任意两个邻近矩阵 $(x_{ij})^1_{n\times m}$ 与 $(x_{ij})^2_{n\times m}$, 对于具有全背景知识的攻击者,M 的任意输出 $S=(s_{ij})_{n\times m}\subseteq Range(M)$,使得 $\Pr[M((x_{ij})^1_{n\times m})\in S]\leqslant e^{\max\{r_i^{\max}\}^n\varepsilon}\Pr[M((x_{ij})^2_{n\times m})\in S]+\delta$. 那么,随机机制

$$M=[(x_{ij})_{n\times m}+round((y_{ij})_{n\times m})]\ \bmod\ 3 \tag{3}$$

是 $(\max\{r_i^{\max}\}^n\varepsilon,\delta)$-差分隐私.

另外,由于个体 i 的 SNP 序列值表示为向量 $X_i=(x_{i1},x_{i2},\ldots,x_{im})$.类似地,下面我们来定义向量差分隐私.

定义 3(向量差分隐私). 给定$\varepsilon\geqslant 0$,任意两个邻近向量 $(x_{ij})^1_{1\times m}$ 与 $(x_{ij})^2_{1\times m}$, 对于具有全背景知识的攻击者,M 的任意输出 $S_i=(s_{ij})_{1\times m}\subseteq Range(M)$,使得 $\Pr[M((x_{ij})^1_{1\times m})\in S_i]\leqslant e^{r_i^{\max}\varepsilon}\Pr[M((x_{ij})^2_{1\times m})\in S_i]+\delta$. 那么,随机机制

$$M=[(x_{ij})_{1\times m}+round((y_{ij})_{1\times m})]\ \bmod\ 3 \tag{4}$$

是 $(r_i^{\max}\varepsilon,\delta)$-差分隐私.

因此,向量差分隐私是矩阵差分隐私的特例.下面给出矩阵差分隐私的通用算法 1.其中,概率分布$\pi(\Delta fc/\max\{r_i^{\max}\}^n\varepsilon)$ 可以是拉普拉斯分布和高斯分布,即噪音矩阵$(y_{ij})_{n\times m}$是由拉普拉斯机制(Laplace mechanism,简称 LM)和高斯机制(Gaussian mechanism,简称 GM)[20]产生的.相应的常数 c 分别为 1 和 $\sqrt{2\ln(1.25/\delta)}$.由于 SNP 二倍体基因型矩阵存储$(x_{ij})_{n\times m}$中元素 $x_{ij}\in\{0,1,2\}$,这里暂且将 x_{ij} 看作字符型,简单地定义基因型 x_{ij} 的效用函数为 $u:x_{ij}\to x_{ij}$,也就是说,$u(x_{ij}=0)=0$,$u(x_{ij}=1)=1$ 和 $u(x_{ij}=2)=2$,那么效用函数的敏感度为$\Delta u=2$,因此在指数机制下选取基因型值 0、1 和 2 的概率分别正比于 1、$e^{\varepsilon/4}$ 和 $e^{\varepsilon/2}$.因为 SNP 基因型矩阵及其对应的效用矩阵的元素都是 0、1 和 2,所以通过指数机制选择基因型值 0、1 和 2 的随机性较差,那么在 SNP 基因型数据的这种效用函数定义方式下,使用指数机制将导致基因型数据及其相关的敏感信息泄露,因此本文没有考虑指数机制(exponential mechanism,简称 EM)[20].

算法 1. 在 SNP 连锁不平衡下的矩阵差分隐私.

输入:SNP 二倍体基因型矩阵$(x_{ij})_{n\times m}$,且 $x_{ij}\in\{0,1,2\}$.初始化ε,δ和Δf;

输出:随机扰动和置换的 SNP 二倍体基因型矩阵$(s_{ij})_{n\times m}$.

1: 计算 SNP 连锁不平衡的相关系数 r_{ij}

2: 生成噪音矩阵$(y_{ij})_{n\times m}$,且 $y_{ij}\sim\pi(\Delta fc/\max\{r_i^{\max}\}^n\varepsilon)$

3: $(s_{ij})_{n\times m}=[(x_{ij})_{n\times m}+round((y_{ij})_{n\times m})]\ \bmod\ 3$

3 矩阵差分隐私的分析

下面从理论上分析矩阵差分隐私的性质.

定理 1. 矩阵差分隐私是 $(\max\{r_i^{\max}\}^n\varepsilon,\delta)$-差分隐私.

证明:让 $(x_{ij})^1_{n\times m}$ 与 $(x_{ij})^2_{n\times m}$ 是邻近矩阵,因此有 $d((x_{ij})^1_{n\times m},(x_{ij})^2_{n\times m})=1$. 噪音矩阵 $round((y_{ij})_{n\times m})$ 中元素 y_{ij} 服从尺度参数为 $\Delta fc/\max\{r_i^{\max}\}^n\varepsilon$ 的概率分布 $\pi(\Delta fc/\max\{r_i^{\max}\}^n\varepsilon)$. 在矩阵差分隐私中,噪音矩阵是由拉普拉斯机制和高斯机制产生的.所以,根据性质 2,对于个体 i,使用噪音向量$(y_{ij})_{1\times m}$ 扰动基因数据$(x_{ij})_{1\times m}$ 是满足 $(r_i^{\max}\varepsilon,\ \delta)$-差分隐私的.这里,对应于邻近矩阵 $(x_{ij})^1_{n\times m}$ 与 $(x_{ij})^2_{n\times m}$, 添加到 $(x_{ij})^1_{n\times m}$ 与 $(x_{ij})^2_{n\times m}$ 的噪音矩阵$(y_{ij})_{n\times m}$ 服从期望为 0 的概率分布 $\pi(\Delta fc/\max\{r_i^{\max}\}^n\varepsilon)$, 则有

$$\Pr[(x_{ij})^1_{1\times m}+(y_{ij})_{1\times m}]\leqslant e^{r_i^{\max}\varepsilon}\Pr[(x_{ij})^2_{1\times m}+(y_{ij})_{1\times m}]+\delta \tag{5}$$

由性质 2 可知,不等式 $\Pr[(x_{ij})^1_{n\times m}+(y_{ij})_{n\times m}]\leqslant e^{\max\{r_i^{\max}\}^n\varepsilon}\Pr[(x_{ij})^2_{n\times m}+(y_{ij})_{n\times m}]+\delta$ 成立.由性质 1,下面两个不等式成立:

$$\Pr[(x_{ij})^1_{n\times m}+round((y_{ij})_{n\times m})]\leqslant e^{\max\{r_i^{\max}\}^n\varepsilon}\Pr[(x_{ij})^2_{n\times m}+round((y_{ij})_{n\times m})]+\delta \tag{6}$$

$$\Pr[((x_{ij})^1_{n\times m}+round((y_{ij})_{n\times m}))\bmod 3]\leqslant e^{\max\{r_i^{\max}\}^n\varepsilon}\Pr[((x_{ij})^2_{n\times m}+round((y_{ij})_{n\times m}))\bmod 3]+\delta \tag{7}$$

所以不等式 $\Pr[M((x_{ij})^1_{n\times m})\in S]\leqslant e^{\max\{r_i^{\max}\}^n\varepsilon}\Pr[M((x_{ij})^2_{n\times m})\in S]+\delta$ 成立.因此,矩阵差分隐私机制 M 满足 $(\max\{r_i^{\max}\}^n\varepsilon,\delta)$-差分隐私. □

为了分析矩阵差分隐私的效用,因为 $S=(s_{ij})_{n\times m}\subseteq Range(M)$, 本文使用 $U=\left|(x_{ij})_{n\times m}\cap(s_{ij})_{n\times m}\right|\Big/\left|(x_{ij})_{n\times m}\right|$ 度量矩阵差分隐私机制的效用[18].

定理 2. 矩阵差分隐私的效用在[R_0,1]区间,R_0 表示隐私预算ε最小时矩阵差分隐私下噪音矩阵中模 3 余 0 元素数量的百分比值.

证明:首先考虑 3 种极端的情况.

(1) 当噪音矩阵 $Y=(y_{ij})_{n\times m}$ 的所有元素满足 $round(y_{ij})\bmod 3=0$ 时,$round((y_{ij})_{n\times m})$的所有元素都模 3 同余 0.因此,在$(x_{ij})_{n\times m}$ 与$(s_{ij})_{n\times m}\subseteq Range(M)$之间的所有 SNP 二倍体基因型数据相同.因此,矩阵差分隐私机制的最大效用为 1.

(2) 当噪音矩阵 $Y=(y_{ij})_{n\times m}$ 的所有元素满足 $round(y_{ij})\bmod 3\equiv 1$ 时,$(0+1)\bmod 3\equiv 1$,$(1+1)\bmod 3\equiv 2$ 和$(2+1)\bmod 3\equiv 0$.因此,$(x_{ij})_{n\times m}$ 与$(s_{ij})_{n\times m}\subseteq Range(M)$之间的所有 SNP 二倍体基因型取值都不相同,此时矩阵差分隐私机制的效用是 0.

(3) 当噪音矩阵 $Y=(y_{ij})_{n\times m}$ 的所有元素满足 $round(y_{ij})\bmod 3\equiv 2$ 时,$(0+2)\bmod 3\equiv 2$,$(1+2)\bmod 3\equiv 0$ 和$(2+2)\bmod 3\equiv 1$.因此,$(x_{ij})_{n\times m}$ 与$(s_{ij})_{n\times m}\subseteq Range(M)$之间的所有 SNP 二倍体基因型取值也都不相同,此时矩阵差分隐私机制的效用是 0.

上述证明中考虑(2)和(3)两种极端情况,使矩阵差分隐私下基因数据的效用为 0.然而,由于噪音的随机性,矩阵差分隐私下基因数据的最小效用是大于 0 的,详见第 4.4 节基因数据的效用分析.下面考虑第 4 种情况.

(4) 在矩阵差分隐私中,由于隐私预算ε越小,邻近基因数据矩阵 $(x_{ij})^1_{n\times m}$ 与 $(x_{ij})^2_{n\times m}$ 的不可区分性越好,进而矩阵差分隐私保护越强,那么基因数据的效用达到最低.在矩阵差分隐私中基因数据的效用与模 3 余 0 的噪音数量的百分比值是一致的.也就是说,如果隐私预算ε最小,矩阵差分隐私产生模 3 余 0 的噪音数量百分比值为 $R_0(1>R_0>0)$,那么基因数据的最小效用为 R_0.反之,隐私预算ε越大,基因数据效用可达到百分比值 1.

综上,由于噪音的随机性,矩阵差分隐私机制的效用属于区间[R_0,1]. □

定理 3. 考虑连锁不平衡、矩阵加运算和模余运算的计算复杂度分别为 $O(n\times m^2)$、$O(n\times m)$和 $O(n\times m)$.矩阵差分隐私的计算复杂度如下:(1) 当 $n=m$ 时,矩阵差分隐私的计算复杂度为 $O(n^3)$;(2) 当 $n>m$ 时,矩阵差分隐私的计算复杂度为 $O(nm^2)$;(3) 当 $n<m$ 时,矩阵差分隐私的计算复杂度为 $O(nm^2)$.

证明:在矩阵差分隐私中,产生随机噪音是有效的,忽略其计算复杂度,而计算连锁不平衡、矩阵加运算和模余运算分别需要 $8n\times(m^2-m)$、$n\times m$ 和 $n\times m$ 次运算,考虑 3 种情况.

(1) 当 $n=m$ 时,矩阵差分隐私的计算复杂度为 $O(n^3)$.

(2) 当 $n>m$ 时,矩阵差分隐私的计算复杂度为 $O(nm^2)$.

(3) 当 $n<m$ 时,矩阵差分隐私的计算复杂度为 $O(nm^2)$. □

总之,矩阵差分隐私满足差分隐私的定义,同时具有效用属于区间[R_0,1],其中,R_0 是矩阵差分隐私下隐私预算最小时噪音矩阵中模 3 余 0 元素数量的百分比值,并且矩阵差分隐私的计算复杂度是多项式时间的.

4 实验分析

本文在矩阵差分隐私下选择拉普拉斯分布和高斯分布来进行实验分析.首先进行噪音分析,然后与拉普拉斯机制和高斯机制比较分析矩阵差分隐私保护模型的隐私和效用.在所有的实验分析中,考虑 SNP 二倍体基因型数据的特点,初始化 SNP 连锁不平衡的相关系数为 $r_{ij}=1$ 和敏感度$\Delta f=2$.另外,分别初始化隐私预算$\varepsilon=0.001$ 和

概率值δ=0.01.

4.1 数据集

国际人类基因组单体型图计划(Int'l Hapmap Project)的数据是公开可用的[21],本文使用 2010 年 5 月发布的阶段 III 的 165 个 CEU(utah residents with northern and Western European ancestry from the CEPH collection)群体的 22 号染色体的基因型和频率数据集.在实验分析之前,基于频率数据集预处理基因型数据集,将 SNP 二倍体基因型数据编码为 0、1 和 2.在 CEU 基因型数据集中,将丢失的数据'NN'用 0 代替.本文分别选择 500、1 000 和 1 500 个 SNP 位点进行实验分析.

4.2 噪音分析

在矩阵差分隐私中,尺度参数为 $\Delta fc/\max\{r_i^{\max}\}^n\varepsilon$ 的拉普拉斯机制(LM)和高斯机制(GM)产生的噪音矩阵为 $(y_{ij})_{n\times m}$.在两种机制下,图 3 所示分别计算矩阵 $round((y_{ij})_{165\times500})$、$round((y_{ij})_{165\times1000})$和 $round((y_{ij})_{165\times1500})$模 3 余 0 的噪音数量的百分比值 R.可以观察到,模 3 余 0 的噪音数量百分比值随着隐私预算的增加而增加,而不随噪音数量的大小而发生变化.这个结果为解释隐私和基因数据效用的实验结果奠定了基础.随着隐私预算的增加,拉普拉斯机制与高斯机制相比,所有模 3 余 0 的噪音数量的百分比值明显更快地增加.当隐私预算ε=7 时,拉普拉斯机制的 R 值将达到 80%,而高斯机制的 R 值才达到 40%.这是因为,在相同的隐私预算下,拉普拉斯分布与高斯分布相比,基于拉普拉斯机制的矩阵差分隐私产生的噪音矩阵中模 3 余 0 的元素更多.

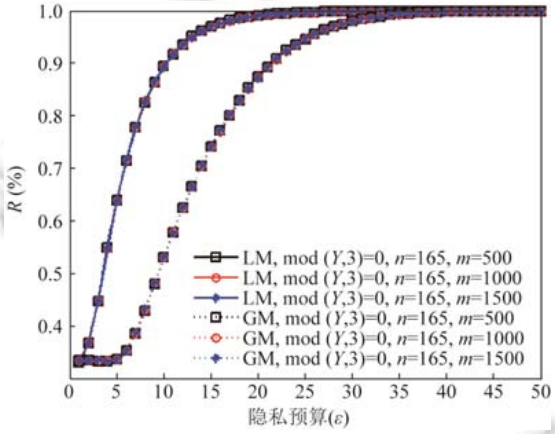

Fig.3 The percentage of noises matrix entries module 3 satisfying the residue to be 0 for matrix differential privacy

图 3 矩阵差分隐私下噪音矩阵模 3 余 0 的元素数量的百分比值

4.3 隐私分析

为了评估基因隐私保护模型的隐私,对于拥有全背景知识的攻击者,本文定义标准化期望估计误差作为隐私度量.因为元素 x_{ij} 在矩阵差分隐私下的随机扰动元素为 s_{ij},因此,定义基因数据的隐私度量为

$$E=\frac{\sum_{1\leqslant i\leqslant n}\sum_{1\leqslant j\leqslant m}P(s_{ij})\left\|s_{ij}-x_{ij}\right\|_1}{mn} \tag{8}$$

通过比较,我们来分析矩阵差分隐私与拉普拉斯机制、高斯机制的标准化期望估计误差.如图 4 和图 5 所示,矩阵差分隐私、拉普拉斯机制和高斯机制的标准化期望估计误差都随隐私预算的增大而减小.主要原因是,隐私预算越大,拉普拉斯分布和高斯分布的方差越小,矩阵差分隐私产生模 3 余 0 的噪音越多.因此拉普拉斯机制和高斯机制直接添加噪音到 SNP 基因型数据会导致效用灾难,而矩阵差分隐私通过噪音模余运算提高了 SNP 基因型数据的效用,见第 4.4 节矩阵差分隐私的效用分析.由此,矩阵差分隐私实现了基因数据的隐私保护,不过,隐私保护强度显然低于拉普拉斯机制和高斯机制.另外,由图 4 和图 5 可知,随着隐私预算的增加,高斯机制

的标准化期望误差较拉普拉斯机制要大,为了更好地权衡隐私和效用,可以选择拉普拉斯机制实现矩阵差分隐私.

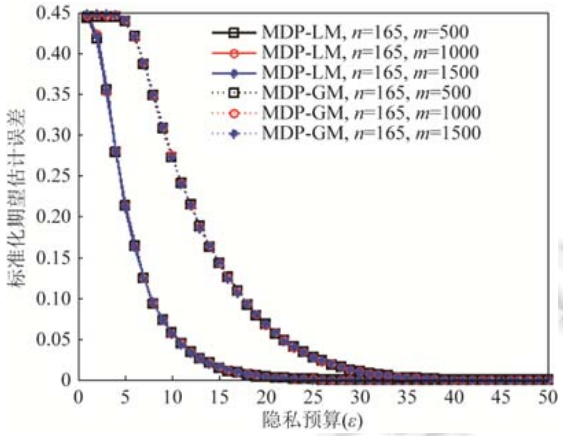

Fig.4 The normalized expected estimation error for matrix differential privacy

图 4 矩阵差分隐私下的标准化期望估计误差

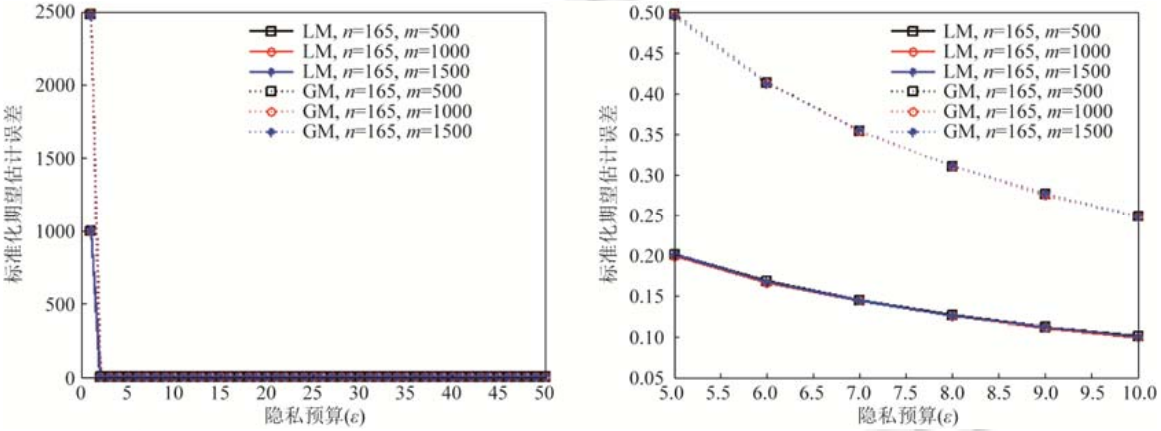

Fig.5 The normalized expected estimation error for Laplace mechanism and Gaussian mechanism

图 5 拉普拉斯机制和高斯机制下的标准化期望估计误差

因此,根据 SNP 连锁不平衡下差分隐私的不可区分性,矩阵差分隐私实现了 SNP 基因型数据和 SNP 连锁不平衡的隐私保护.

4.4 效用分析

尽管矩阵差分隐私可以实现 SNP 基因型数据的隐私保护,考虑到 SNP 基因型数据的分析,因此还需要分析 SNP 基因型数据的效用.在矩阵差分隐私中,对于原始的 SNP 基因型数据$(x_{ij})_{n\times m}$和扰动后的 SNP 基因型数据$(s_{ij})_{n\times m}$,根据$U=\left|(x_{ij})_{n\times m}\cap(s_{ij})_{n\times m}\right|/\left|(x_{ij})_{n\times m}\right|$作为效用度量方法实验分析基因数据的效用.

如图 6 所示,随着隐私预算的增加,矩阵差分隐私保护模型下的基因数据效用递增,并且增长到 100%保持不变.这是因为,随着隐私预算增大,拉普拉斯分布和高斯分布的方差变小,矩阵差分隐私产生模 3 余 0 的噪音就更多.当隐私预算较小时,基于拉普拉斯机制的矩阵差分隐私可以实现更好的基因数据效用,以此保证较好的计算不可区分性,进而实现更好的差分隐私保护.例如,当ε=7 时,基于拉普拉斯机制的基因数据效用可以达到 80%,而基于高斯机制的基因数据效用为 40%,这与图 3 中拉普拉斯机制和高斯机制产生噪音矩阵的四舍五入近似值模

3 余 0 的噪音数量的百分比值是一致的.而图 7 中随着隐私预算的增加,基因组数据的效用保持 0 不变.这是因为,拉普拉斯机制和高斯机制直接添加噪音到基因数据,破坏了基因数据效用,导致基因数据效用灾难.由此可知,矩阵差分隐私比拉普拉斯机制和高斯机制更适合于基因数据的隐私保护.

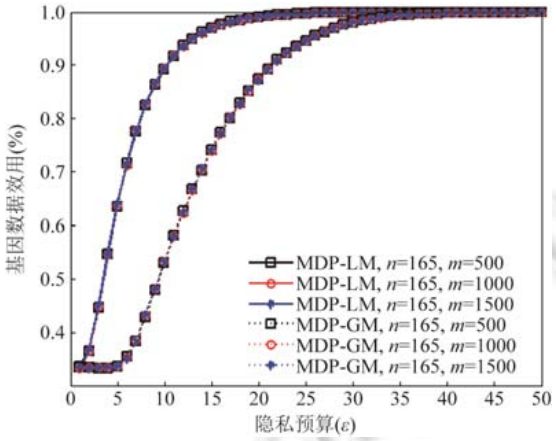

Fig.6 The genome data utility for matrix differential privacy

图 6 矩阵差分隐私下的基因数据效用

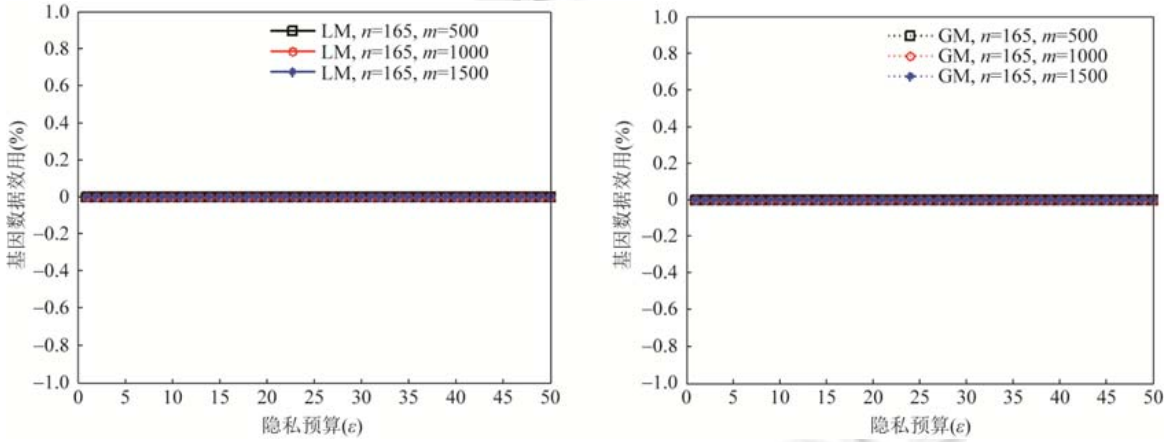

Fig.7 The genome data utility for Laplace mechanism and Gaussian mechanism

图 7 拉普拉斯机制和高斯机制下的基因数据效用

因此,矩阵差分隐私相比于拉普拉斯机制和高斯机制更适合于基因数据的隐私保护,保证了基因数据和 SNP 连锁不平衡的隐私保护与基因数据效用之间的权衡.在表 1 中,通过比较分析,总结矩阵差分隐私与拉普拉斯机制、高斯机制的相关性质.其中,最小效用 R_0 表示矩阵差分隐私在最小隐私预算下所有模 3 余 0 的噪音数量的百分比值.

Table 1 The comparison among matrix differential privacy, Laplace mechansim and Gaussian mechanism

表 1 矩阵差分隐私与拉普拉斯机制、高斯机制的比较

机制	理论基础	扰动过程	置换过程	是否满足差分隐私	基因数据效用
矩阵差分隐私	概率分布不可区分性	添加四舍五入取整噪音	模余运算	是	$[R_0,1]$
拉普拉斯机制	概率分布不可区分性	直接添加噪音	–	是	0
高斯机制	概率分布不可区分性	直接添加噪音	–	是	0

5 结 论

为了保护 SNP 连锁不平衡下基因关联的敏感信息,本文提出了矩阵差分隐私保护模型.该模型满足差分隐私,同时保证基因数据效用在$[R_0,1]$区间,其中,R_0是矩阵差分隐私在隐私预算最小时噪音矩阵中模 3 余 0 的噪音数量的百分比值,并且矩阵差分隐私是多项式时间计算有效的.

对于基因数据,基因隐私保护模型在连锁不平衡下保证隐私是可行的.通过结合矩阵加运算、SNP 连锁不平衡下差分隐私的定义和模余运算,提出了向量差分隐私和矩阵差分隐私,并且向量差分隐私是矩阵差分隐私的特例.根据矩阵差分隐私的性质,为了疾病标记发现,基因隐私保护模型可以用于 DNA 数据集的差分隐私选择[15];在 GWAS 中,矩阵差分隐私也可以对基于隐私编辑距离相似患者查询提供隐私保护[12];矩阵差分隐私阻止从 GWAS 统计值中识别特定的个体[16];并且,矩阵差分隐私可以实现隐私保护罕见疾病变异分析[8];矩阵差分隐私在基因组串搜索中是有效的隐私保护方法[11].更进一步说,在矩阵差分隐私下可以实现宏基因组分析[13].因此,矩阵差分隐私可以推广到基因数据收集、搜索和序列配对等应用的隐私保护中.

在矩阵差分隐私中,可以通过行划分、列划分或者其他快速矩阵计算方法[22]降低其计算复杂度,进而提高计算效率.另外,考虑高阶的 SNP 连锁不平衡,Samani 等人[23]表明了对隐藏 SNP 的个体基因数据具有更强的推断攻击.Tramèr 等人[17]考虑有界先验知识的差分隐私,并应用于 GWAS.通过孟德尔定律、基因变异之间的统计关系和基因与表型之间的统计关系,在个体的基因组或表型被观察到的情况下,Humbert 等人[4]详述了重构攻击推断该个体的亲戚的基因组.相比较考虑攻击者的背景知识,本文仅考虑了 SNP 连锁不平衡下基因隐私保护.在下一步的工作中,研究 SNP 连锁不平衡下具有先验知识的基因隐私保护模型,除了考虑成对 SNP 连锁不平衡外,还需要考虑高阶的 SNP 连锁不平衡,并考虑攻击者更多的先验知识,包括可利用的基因数据、个体的血缘关系以及重组规则等.

References:

[1] Li Y, Chen L. Big biological data: Challenges and opportunities. Genomics, Proteomics & Bioinformatics, 2014,12(5):187–189. [doi: 10.1016/j.gpb.2014.10.001]

[2] Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux JP, Malin BA, Wang X. Privacy in the genomic era. ACM Computing Surveys (CSUR), 2015,48(1):6:1–44. [doi: 10.1145/2767007]

[3] Wagner I. Evaluating the strength of genomic privacy metrics. ACM Trans. on Privacy and Security (TOPS), 2017,20(1):2:1–34. [doi: 10.1145/3020003]

[4] Humbert M, Ayday E, Hubaux JP, Telenti A. Quantifying interdependent risks in genomic privacy. ACM Trans. on Privacy and Security (TOPS), 2017,20(1):3:1–31. [doi: 10.1145/3035538]

[5] Lin Z, Owen AB, Altman RB. Genomic research and human subject privacy. Science, 2004,305(5681):183. [doi: 10.1126/science.1095019]

[6] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics, 2008,4(8):1–9. [doi: 10.1371/journal.pgen.1000167]

[7] Gottlieb S. US employer agrees to stop genetic testing. British Medical Journal, 2001,322(7284):449. [doi: 10.1136/bmj.322.7284.449/a]

[8] Chen F, Wang S, Jiang X, Ding S, Lu Y, Kim J, Sahinalp SC, Shimizu C, Burns JC, Wright VJ, Png E, Hibberd ML, Lloyd DD, Yang H, Telenti A, Bloss CS, Fox D, Lauter K, Ohno-Machado L. PRINCESS: Privacy-protecting rare disease international network collaboration via encryption through software guard extensions. Bioinformatics, 2017,33(6):871–878. [doi: 10.1093/bioinformatics/btw758]

[9] Ayday E. Cryptographic solutions for genomic privacy. In: Proc. of the Int'l Conf. on Financial Cryptography and Data Security. Berlin, Heidelberg: Springer-Verlag, 2016. 328–341. [doi: 10.1007/978-3-662-53357-422]

[10] Wang S, Zhang Y, Dai W, Lauter K, Kim M, Tang Y, Xiong H, Jiang X. HEALER: Homomorphic computation of exact logistic regression for secure rare disease variants analysis in GWAS. Bioinformatics, 2016,32(2):211–218. [doi: 10.1093/bioinformatics/btv563]

[11] Shimizu K, Nuida K, Rätsch G. Efficient privacy-preserving string search and an application in genomics. Bioinformatics, 2016, 32(11):1652–1661. [doi: 10.1093/bioinformatics/btw050]

[12] Wang XS, Huang Y, Zhao Y, Tang H, Wang X, Bu D. Efficient genome-wide, privacy-preserving similar patient query based on private edit distance. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. New York: ACM, 2015. 492–503. [doi: 10.1145/2810103.2813725]

[13] Wagner J, Paulson JN, Wang X, Bhattacharjee B, Bravo HC. Privacy-preserving microbiome analysis using secure computation. Bioinformatics, 2016,32(12):1873–1879. [doi: 10.1093/bioinformatics/btw073]

[14] Dwork C, Pottenger R. Toward practicing privacy. Journal of the American Medical Informatics Association, 2013,20(1):102–108. [doi: 10.1136/amiajnl-2012-001047]

[15] Zhao Y, Wang X, Jiang X, Ohno-Machado L, Tang H. Choosing blindly but wisely: Differentially private solicitation of DNA datasets for disease marker discovery. Journal of the American Medical Informatics Association, 2015,22(1):100–108. [doi: 10.1136/amiajnl-2014-003043]

[16] Cai R, Hao Z, Winslett M, Xiao X, Yang Y, Zhang Z, Zhou S. Deterministic identification of specific individuals from GWAS results. Bioinformatics, 2015,31(11):1701–1707. [doi: 10.1093/bioinformatics/btv018]

[17] Tramèr F, Huang Z, Ayday E. Differential privacy with bounded priors: Reconciling utility and privacy in genome-wide association studies. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. New York: ACM, 2015. 1286–1297. [doi: 10.1145/2810103.2813610]

[18] Simmons S, Berger B. Realizing privacy preserving genome-wide association studies. Bioinformatics, 2016,32(9):1293–1300. [doi: 10.1093/bioinformatics/btw009]

[19] McSherry FD. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In: Proc. of the 2009 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 2009. 19–30. [doi: 10.1145/1559845.1559850]

[20] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 2014,9(3-4):211–407. [doi: 10.1561/0400000042]

[21] NCBI retiring HapMap Resource. https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/

[22] Golub GH, Van Loan CF. Matrix Computations. 4th ed., Baltimore: The Johns Hopkins University Press, 2012. 1–104.

[23] Samani SS, Huang Z, Ayday E, Elliot M, Fellay J, Hubaux JP, Kutalik Z. Quantifying genomic privacy via inference attack with high-order SNV correlations. In: Proc. of the 2015 IEEE Security and Privacy Workshops. IEEE, 2015. 32–40. [doi: 10.1109/SPW.2015.21]

刘海(1989－),男,贵州遵义人,博士生,主要研究领域为生物医学大数据隐私保护.

彭长根(1963－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为密码学,信息安全,大数据隐私保护.

吴振强(1968－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络与信息安全,分布式计算,数据隐私保护.

雷秀娟(1975－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为生物信息计算,智能计算.