计算机科学
COMPUTER SCIENCE

基于本地化差分隐私的频率特征提取

黄觉，周春来

引用本文

黄觉，周春来. 基于本地化差分隐私的频率特征提取[J]. 计算机科学, 2022, 49(7): 350-356.

HUANG Jue, ZHOU Chun-lai. Frequency Feature Extraction Based on Localized Differential Privacy [J]. Computer Science, 2022, 49(7): 350-356.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于本地化差分隐私的键值数据关联分析

Correlation Analysis for Key-Value Data with Local Differential Privacy

计算机科学, 2021, 48(8): 278-283. https://doi.org/10.11896/jsjkx.201200122

支持本地化差分隐私保护的 k-modes 聚类方法

k-modes Clustering Guaranteeing Local Differential Privacy

计算机科学, 2021, 48(2): 105-113. https://doi.org/10.11896/jsjkx.200700172

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900229

基于本地化差分隐私的频率特征提取

黄 觉 周春来
中国人民大学信息学院 北京 100872
(3287401165@qq.com)

摘 要 大数据时代信息技术不断发展,隐私问题越来越受到人们的关注。尤其是随着移动端的普及,如何在数据发布的同时保护用户个人的隐私信息是当前面临的重大挑战。此前学术界曾提出依赖于可信第三方的中心化差分隐私技术,但在实际应用中可信第三方的条件通常不成立;随后,在中心化差分隐私的基础上进一步提出了本地化差分隐私,它能够防止来自不可信第三方的隐私攻击,并且面对具有任意知识背景的隐私攻击者依然具有很强的防御效果。但是,市场通常不仅要迎合用户的需求,也要满足运营商的要求。为了对两者进行平衡,如何解决运营商的分析任务是亟待解决的问题。RAPPOR(Randomized Aggregatable Privacy-Preserving Ordinal Response)算法能够很好地完成这个任务,它通过使用两次随机响应机制对用户数据进行加密,保证了隐私保护的力度,并使用 Lasso 回归模型对加密数据进行解密,保证了频率特征提取的准确度。文中的贡献在于将 RAPPOR 算法应用于疫情信息采集,在保护受访者隐私信息的同时能获取真实的疫情资料,以美国各地新冠确诊人数的数据集进行实验,实验结果表明,所提方法较高程度地拟合了真实结果,完成了频率特征提取的分析任务。RAPPOR 算法实现了本地化差分隐私技术从理论走向应用,切实保障了个人的隐私问题。

关键词:本地化差分隐私;RAPPOR;频率特征;随机响应
中图法分类号 TP311

Frequency Feature Extraction Based on Localized Differential Privacy

HUANG Jue and ZHOU Chun-lai
Department of Information,Renmin University,Beijing 100872,China

Abstract With the continuous development of information technology in the era of big data,privacy problem has attracted more and more attention. Especially with the increasing popularity of mobile terminals,how to protect users' privacy information while releasing data is a major challenge at present. Previously,academic circle has proposed the center differential privacy technology that relies on a trusted third platform,but the condition that needs a trusted third platform is usually not valid in practical applications. On the basis of center differential privacy,localized differential privacy is further proposed. It can prevent privacy attacks from untrusted third platforms,and it still has a strong defensive effect against privacy attackers with abundant knowledge background. But markets often cater to the needs of service providers as well as users. In order to balance the contradiction between the two,how to accomplish the analysis tasks of service providers is a problem that must be solved. RAPPOR is a good mechanism to accomplish these tasks. It encrypts user data by using two random response mechanisms to ensure the strength of privacy protection. Lasso regression model is used to decrypt the encrypted data to ensure the accuracy of frequency feature extraction. In this paper,RAPPOR algorithm is applied to COVID-19 epidemic information collection,which can obtain real epidemic data while protecting the privacy of respondents. The dataset which includes people diagnosed with COVID-19 in the United States is used to simulate the RAPPOR mechanism and fits the real results to a high degree. RAPPOR algorithm realizes the localized differential privacy technology from theory to application,and effectively protects personal privacy.

Keywords Localized differential privacy,RAPPOR,Frequency characteristics,Random response

1 引言

随着信息技术的日渐发达,隐私问题成为了人们普遍关注的热点问题。大数据时代,信息技术给人们带来巨大便利的同时,也带来了威胁自身发展的数据安全与用户隐私问题。为了保证信息技术的可持续发展,保护个人数据隐私

到稿日期:2021-09-27 返修日期:2021-12-20
基金项目:国家自然科学基金重点项目(61732006);国家自然科学基金(61972404,12071478)
This work was supported by the Key Program of the National Natural Science Foundation of China(61732006) and National Natural Science Foundation of China(61972404,12071478).
通信作者:周春来(czhou@ruc.edu.cn)

成为了政府和企业的当务之急。但是,一味地加大隐私保护力度而忽视数据的可用性,会对社会科学等一些依赖数据分析的学科造成巨大冲击,从而抑制社会的发展,这与我们做学术的初衷相悖。因此,需要找到一个合适的平衡点,使得用户在使用产品时能够保护自己的隐私信息,同时又使研究者获取到一些总体的数据特征(如频率),用于科学研究。此前学术界提出了许多优秀隐私保护模型,如 k-匿名、l-多样性和 t-紧密性,但是这些模型具有一个共同的漏洞:当攻击者具有较多的背景知识时,这些模型的防御能力将大大减弱。2006年,美国电影公司 Netflix 举办了一个推荐系统算法竞赛,公布了一些经过匿名化处理的用户影评数据用作参赛的样例数据,这些数据的特点是只有每个用户对电影的评分和评分的时间戳。然而,德州大学奥斯汀分校的两位研究人员借助公开的互联网电影数据库(IMDB)的用户影评数据将 Netflix 上的电影浏览信息(其中包含涉及敏感题材的电影)成功破解80%,该次隐私泄漏事件充分说明了传统的隐私保护手段在强力的技术攻击面前十分脆弱。由此,本文引入了差分隐私的概念。

RAPPOR 算法是差分隐私技术在计算机中的应用,该算法在布隆过滤器上使用随机响应机制,从而达到给隐私信息加密的目的。本文的贡献在于将 RAPPOR 算法应用到疫情信息采集,在收集信息的过程中充分保证了个人的病史不被泄露,又能得到总体上的疫情信息(如各地患病的频数信息),以制定精准的政策。

本文第 2 节详细讲述了本地化差分隐私如何代替中心化差分隐私,成为更可靠的隐私保护机制;第 3 节讨论了差分隐私技术在计算机中实现的关键技术——随机响应机制;第 4 节详细讲述了本文的核心算法——RAPPOR 算法,实现了 RAPPOR 算法在疫情信息采集中的应用,通过实验充分说明了该算法的实用性和有效性;最后总结全文。

2　差分隐私

差分隐私(Differential Privacy)这一概念是近几年提出的崭新概念,它的主要思想是:在数据集中,无论是删除还是添加一条记录,都不会影响最终的查询结果。目前,差分隐私分为两大类,一类是本地化差分隐私(Local Differential Privacy),另一类是中心化差分隐私(Center Differential Privacy)。这两种差分隐私的定义十分接近,在形式上几乎没有差别,只在处理的对象方面有些许差别。而在应用中,这种差别表现为本地化差分隐私通常是在用户端实现,而中心化差分隐私通常是在服务器端或者说是数据库上实现。但是,中心化差分隐私技术的实现前提是有一个可信的存储数据的第三方平台,由于近几年不断曝出的隐私泄露丑闻,可信的第三方数据收集者已经成为奢求,因此各方都将策略调整为构思一种不需要依赖第三方平台的隐私保护技术[1-5]。本地化差分隐私技术正是这种不需要依赖第三方平台的隐私保护技术,它脱胎于中心化差分隐私,但是比中心化差分隐私具有更好的实用性。

本文只考虑本地化差分隐私,因此下面只给出本地化差分隐私的定义,而中心化差分隐私的定义可以参考文献[6]。

差分隐私机制一般通过隐私算法来表示,隐私算法是一种随机算法,即,如果 $M:X\to Y$ 是一种随机算法,那么对于任意 $x\in X$,$M(x)$是一个 Y 上的一个概率密度函数。

定义 1(本地化差分隐私)　给定一个含有 n 条用户记录的数据集,同时给定一种隐私算法 M 及其定义域 $D(M)$和值域 $R(M)$,如果算法 M 在任意两条记录 t 和 t'($t,t'\in D(M)$)上得到相同的输出结果 t^*($t^*\in R(M)$)满足下列不等式,则 M 满足 ε-本地化差分隐私。

$$Pr[M(t)=t^*]\leqslant e^{\varepsilon}\times Pr[M(t')=t^*]$$

定义 1 表明,对原始数据集中任何一条记录进行修改后,得到的输出集的概率分布变化相对较小,其概率比值不超过 e^{ε},也就是说两个分布处处相近。这就意味着随机算法 M 的输出不会因为某一条记录的存在或改变而受到太大影响,即便公开了算法的结果,攻击者通过观察随机算法 M 的输出,也很难分辨出原始数据集,从而实现了对原始数据集中数据的保护。隐私保护预算 ε 是非常重要的,算法需要的隐私保护程度越高,隐私保护预算取值就越小,但数据集的可用性也相对降低;相反,隐私保护预算取值越大,隐私保护程度越低,数据可用性就越高。现实中,应根据具体需求,合理地选取隐私预算 ε。对于本地化差分隐私技术而言,它不用依赖第三方的数据收集者,因为该项技术在每个用户的用户端就能实现,所以也就免除了第三方平台泄露隐私的可能性。

3　随机响应

Warner 于 1965 年提出了随机响应机制,以实现隐私保护[7]。其主要思想是,利用对敏感问题响应的不确定性对原始数据进行隐私保护。Warner 通过一个概率设备对原始数据添加不确定性。例如,如果受访者的真实数据是 1,通过一个概率设备使之以 p 概率上传 1,以 $1-p$ 概率上传 0;同理可推受访者真实值为 0 的情况。该随机响应机制根据伯努利分布进行随机化,而随机化后的输出结果通常只有两种情况,因此可以很容易通过 0 和 1 来表示。具体的工作机制如图 1 所示。

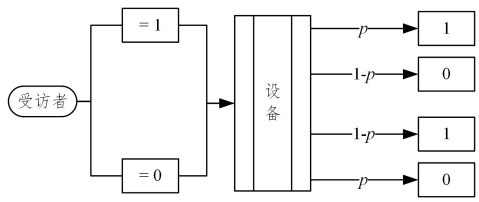

图 1　Warner 模型

Fig. 1　Model of Warner

但是,如果随机化后的结果不是二值的该如何处理呢?这是一个急需解决的问题,因为随着数据集越来越大、越来越复杂,实际问题往往不止一种回答结果,为了更适应实际问题,随机响应机制通常使用一种具有普遍适用性的编码方式——one-hot 编码[8-9]。

one-hot 编码方式将受访者的回答编码成一个长度为 Q 的向量,向量中每个元素只允许为 0 或者 1。这里的 Q 与问题的回答结果个数有关,普遍规定 Q 与回答结果的个数相同。除此之外,Q 个元素中只允许有一个元素为 1,而其他元素全为 0。这样做恰好对应受访者只能在 Q 个结果中选择

一种结果。接下来，我们将添加扰动因素，其思想的本质就是多次应用传统的随机响应机制。对于每个受访者 $\boldsymbol{n}_i$（$\boldsymbol{n}_i$ 是一个 Q 位的向量），对 n_i 的每个元素进行随机化，随机化规则为：以概率 p 使元素 n_{ij}（表示 $\boldsymbol{n}_i$ 的第 j 个位置，其中 j 属于 1 到 Q 之间）保持不变，而以 $1-p$ 的概率使其改变。这里“改变”的含义是：若某一位置为 1，那么就变为 0；若为 0，则变为 1。

经过这一操作后，一段 one-hot 编码的数据将会变成一系列添加了噪声的数据，问题是：如何校正这些数据使之成为可以获取正确统计数据的有用数据呢？以频数为例，本文将介绍一种简单的通过噪声数据得到较为精确频数的方法。例如，任务是需要统计受访者中第 j 个位置为 1 的频数。那么便可以通过构造原始数据和噪声数据之间的关系来进行估计，在随机响应机制中，这种关系通常是一个等式。用 N_j 表示噪声数据中第 j 位为 1 的个数，这个量是可以直接统计得到的。可以非常直观地判断出，第 j 位虽然为 1，但是受访者的原始数据可能为 1 也可能为 0。因此，噪声数据中第 j 个位置的 1 有两个来源，一个是由原始数据中的 1 提供，另一个是由原始数据中的 0 提供。那么我们可以很容易地构造两者之间的联系：

$$N_j = p \times F_j + (1-p)(n-F_j) \tag{1}$$

其中，F_j 是需要估计的频数，它表示原始数据中第 j 个位置为 1 的频数。那么根据式(1)可以得到：

$$F_j = \frac{1}{2p-1}[N_j - (1-p)n] \tag{2}$$

4 RAPPOR

4.1 RAPPOR 基本算法

RAPPOR 是一项从用户端处理众包数据(Crowdsourcing Data)[10]的技术。我们设想一个场景：一个用户打算向服务器传送一条隐私信息 v，那么该用户的终端就会启动 RAPPOR 算法，最终将该信息处理成一个长度为 k 的二进制串，该串就是通过随机响应对隐私信息 v 添加噪声之后的数据，服务器无法理解用户上传的隐私信息的含义，这样就减小了第三方出现隐私泄露的可能。在理解和讲述 RAPPOR 算法的过程中，文献[11]对 RAPPOR 算法的总结给本文提供了极大的帮助，在文献[11]的基础上本文进行了进一步的扩展和应用。RAPPOR 算法设计了两个随机响应机制，用户端的 RAPPOR 算法涉及多个参数(m, k, h, f, p, q)。为了更好地理解算法，本文将逐一解释以上参数的含义，各个参数的含义如表 1 所列。

表 1 实验参数释义

Table 1 Explanations of parameters

Parameter	Explanation
k	布隆过滤器的位数
h	每个数据映射到布隆过滤器上使用哈希函数的次数
m	将原始数据样本划分成的组数
f	进行永久随机响应时布隆过滤器每位以($1-1/2\times f$)的概率保持不变
p	进行短暂随机响应时布隆过滤器上每位由 0 变为 1 的概率
q	进行短暂随机响应时布隆过滤器上每位 1 保持不变的概率

在用户端本地，RAPPOR 算法的执行流程如下：

(1)编码。布隆过滤器是由一个固定大小的二进制向量或者位图(Bitmap)和一系列映射函数组成的，这些映射函数在应用中通常是哈希函数。在初始状态时，对于长度为 k 的位数组，它的 h 个映射函数将这个变量映射成位图中的 h 个点，把它们置为 1。假定用户的真实值为 v，v 通常是一个字符串，我们将它哈希到大小为 k 的布隆过滤器中。需要注意的是，通常情况下只进行一次哈希的效果往往不好，因此需要进行多次哈希。这里将哈希的次数记为 h，它等价于使用 h 个哈希函数进行映射，哈希之后布隆过滤器上将会有 h 个位变为 1。以 16 位布隆过滤器为例，在 $h=2$ 的条件下进行编码。布隆过滤器的工作机制如图 2 所示。

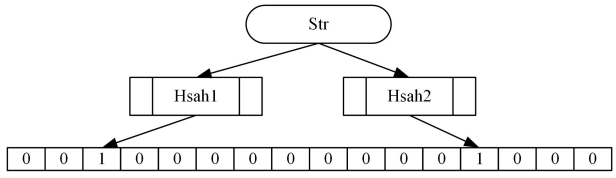

图 2 布隆过滤器的工作机制

Fig. 2 Bloom filter's mechanism

(2)永久随机响应(Permanent Randomized Response)。这是 RAPPOR 算法第一次进行随机响应。对于用户的真实值 v，我们对布隆过滤器 $\boldsymbol{B}$ 的每一位都进行相同的操作。以第 i 位为例，其中 $i \in [0,k]$，通过以下方法来进行随机化。

$$B_i' = \begin{cases} 1, & \text{概率为} \frac{1}{2}f \\ 0, & \text{概率为} \frac{1}{2}f \\ B_i, & \text{概率为} 1-f \end{cases}$$

(3)短暂随机响应(Instantaneous Randomized Response)。对 $\boldsymbol{B}'$ 的每一位进行随机响应，得到的结果是 k 位的二进制串 S。操作基本上与永久随机响应相同，只有概率规则不一样。根据以下方法进行随机化：

$$P(S_i=1) = \begin{cases} q, & \text{如果 } B_i'=1 \\ p, & \text{如果 } B_i'=0 \end{cases}$$

该方法中的 p 表示布隆过滤器上一位为 0 而变为 1 的概率，q 表示布隆过滤器上一位为 1 而保持不变的概率。

(4)上传。将经过一系列随机响应操作后产生的 S 传向服务器。这是数据聚合的一种方式。

上述过程主要集中在 RAPPOR 机制的加密和聚合阶段，是集中体现 RAPPOR 机制具有差分隐私性质的部分，但是应用中的 RAPPOR 机制还需要有解码和分析的过程，因此可以归纳为图 3 所示的流程。

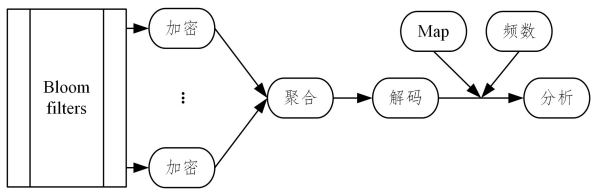

图 3 RAPPOR 机制的流程

Fig. 3 RAPPOR's mechanism

永久随机化和短暂随机化过程都是 RAPPOR 机制的核心之一。这两种随机化有各自不同的作用，它们在实际应用

中都有不同的使命，接下来将详细讲解两种随机响应的作用。

永久随机响应在真实数据的布隆过滤器 $\boldsymbol{B}$ 的基础上产生了带有噪声的数据 $\boldsymbol{B}'$。而 $\boldsymbol{B}'$可能含有也可能不含有关于 B 的任何信息，这就取决于布隆过滤器上每个为 1 的位是否以规定概率被 0 替换。永久随机响应可以保证用户的隐私不受攻击，因为攻击者无法区分布隆过滤器上的每一位是否添加了噪声。此外，需要理解永久随机响应中的“永久”一词，它表示通过这次随机响应产生的 $\boldsymbol{B}'$将会永久存储在用户端本地，这是与短暂随机响应的一个区别点。

短暂随机响应看似只是重复使用随机响应机制，但起到了十分重要的作用。我们不会直接上传永久随机响应产生的 $\boldsymbol{B}'$的原因是，若只进行一次随机化，攻击者通过窃听用户的窗口依然有较大概率复原用户的原始数据。多进行一次随机响应可以明显增大攻击者追踪用户数据的难度，因为即使攻击者可以进行无偏估计，但他估计所得的是经过一次加密后的数据，这就使得隐私保护的力度得到增强。

4.2　RAPPOR 的差分隐私特性

随着数据规模的逐渐增大，攻击者的攻击手段也逐渐变得复杂。任何一个安全可靠的系统都希望有一个严谨的而非经验性的隐私保护机制来抗衡各种各样的攻击手段。在分析 RAPPOR 的隐私保护力度方面，本文采用了一个严谨的隐私概念——差分隐私。上文已经介绍了差分隐私的定义，接下来利用其定义来证明 RAPPOR 算法是符合差分隐私的。从直观上来讲，永久随机响应部分确保真实数据在添加噪声后可以抵抗攻击，而短暂随机响应则保证攻击者在纵向上无法追踪数据。

4.2.1　永久随机响应与差分隐私

首先证明永久随机响应符合差分隐私。

定理 1　永久随机响应满足 ε-差分隐私，其中隐私预算的计算式如下：

$$\varepsilon=2h\ln\left(\frac{1-\frac{1}{2}f}{\frac{1}{2}f}\right) \tag{3}$$

假定 $S=s_1,\cdots,s_k$为使用 RAPPOR 算法后获得的上传数据，那么在给定用户真实值 $\boldsymbol{v}$ 并且 $\boldsymbol{B}'$ 已知的条件下观察到上传数据为 $\boldsymbol{S}$ 的概率为：

$$Pr(\boldsymbol{S}=\boldsymbol{s}|\boldsymbol{V}=\boldsymbol{v})=Pr(\boldsymbol{S}=\boldsymbol{s}|\boldsymbol{B}')Pr(\boldsymbol{B}'|\boldsymbol{B}) \tag{4}$$

相关概率为：

$$Pr(b_i'=1|b_i=1)=1-\frac{1}{2}f \tag{5}$$

$$Pr(b_i'=1|b_i=0)=\frac{1}{2}f \tag{6}$$

不失一般性，我们将布隆过滤器的第 $1\cdots h$ 位设置为 1，即 $\boldsymbol{b}^*=\{b_1=1,\cdots,b_h=1,b_{h+1}=0,\cdots,b_k=0\}$。那么我们可以得到：

$$Pr(\boldsymbol{B}'=\boldsymbol{b}'|\boldsymbol{B}=\boldsymbol{b}')=\left(\frac{1}{2}f\right)^{k-h+b_1'+\cdots+b_h'-b_{h+1}'-\cdots-b_k'}\times\left(1-\frac{1}{2}f\right)^{h-b_1'-\cdots-b_h'+b_{h+1}'+\cdots+b_k'} \tag{7}$$

根据差分隐私的定义，需要选取两个不同的真实值（输入值），记为 $\boldsymbol{B}_1,\boldsymbol{B}_2$，这里的 $\boldsymbol{B}_1,\boldsymbol{B}_2$ 都是 k 位长的布隆过滤器。那么我们需要计算这两个真实值转化为另一值的概率的比值，我们记为 RR。可得：

$$RR=\frac{Pr(\boldsymbol{B}'|\boldsymbol{B}=\boldsymbol{B}_1)}{Pr(\boldsymbol{B}'|\boldsymbol{B}=\boldsymbol{B}_2)} \tag{8}$$

这里需要对 $\boldsymbol{B}'$进行解释，它可以是布隆过滤器的任何一种情况。为了界定，我们还需要一个条件，即 $\boldsymbol{B}_1,\boldsymbol{B}_2$ 每一位的变化都是独立的，这一条件是导出 RAPPOR 与差分隐私联系的关键条件。根据以上给出的条件，我们可以通过以下等式进行推导。

$$RR=\left(\frac{1}{2}f\right)^{2(b_1'+\cdots+b_h'-b_{h+1}'-\cdots-b_{2h}')}\times\left(1-\frac{1}{2}f\right)^{2(b_{h+1}'+\cdots+b_{2h}'-b_1'-\cdots-b_h')} \tag{9}$$

如果要最大化概率比值，则需要使 $\boldsymbol{B}_1,\boldsymbol{B}_2$ 的海明距离最大化。而 $\boldsymbol{B}_1,\boldsymbol{B}_2$ 均有 h 个位为 1，那么两者的最大海明距离为 $2h$。因此：

$$\varepsilon=2h\ln\left(\frac{1-\frac{1}{2}f}{\frac{1}{2}f}\right)$$

证毕。

值得注意的是，ε 不是 k 的函数，但是我们清楚 k 越小，那么真实值在通过哈希函数映射到布隆过滤器的过程中发生碰撞的概率越大，因此若适当增大 k 值将会有益于提高数据分析的精度。这是因为 RAPPOR 算法没有专门处理碰撞的机制，我们只是尽可能地减小发生碰撞的概率，因为每发生一次碰撞，都会导致在数据估计时估计量产生偏差。为了尽可能减小这种偏差，我们必须尽可能减小发生碰撞的概率。

4.2.2　短暂随机化与差分隐私

如果对每个用户只进行一次数据收集，那么攻击者便可以直接从上传数据 $\boldsymbol{S}$ 中学习到关于原始值 $\boldsymbol{B}$ 的信息。因此，如果新增一个随机化过程将能为用户提供更高水平的隐私保护力度。因为短暂随机响应是第二步的随机化过程，上传数据中的 1 可由原始数据中的 0 转变而来，也可能由原始数据中的 1 转变而来。若由原始数据中的 1 转变而来，我们将这种转变的概率记为 q^*，这种转变可由两种变化组成，由 1 变为 1 再变为 1，或者由 1 变为 0 再变为 1。那么：

$$q^*=Pr(S_i=1|b_i=1)=\frac{1}{2}pf+\left(1-\frac{1}{2}f\right)q \tag{10}$$

若由原始数据中的 0 转变而来，则将这种转变概率记为 p^*。这种转变也可由两种变化组成，由 0 变为 0 再变为 1，或者由 0 变为 1 再变为 1。那么：

$$p*=Pr(S_i=1|b_i=1)=\frac{1}{2}qf+\left(1-\frac{1}{2}f\right)p \tag{11}$$

定理 2　短暂随机响应满足 ε-差分隐私，其中隐私预算的计算式如下：

$$\varepsilon=h\log\left(\frac{q^*(1-p^*)}{p^*(1-q^*)}\right) \tag{12}$$

为了证明这个定理，需要构造两种不同输入 $\boldsymbol{B}_1,\boldsymbol{B}_2$ 输出相同值的概率比值。我们将该比值记为 RR，那么可计算得：

$$RR=\frac{Pr(S|B=B_1)}{Pr(S|B=B_2)} \tag{13}$$

进而可以根据本地化差分隐私定义列出不等式：

$$RR \leqslant \left[\frac{q^*(1-p^*)}{p^*(1-q^*)}\right]^h \tag{14}$$

最后，计算出隐私预算的值：

$$\varepsilon = h\log\left(\frac{q^*(1-p^*)}{p^*(1-q^*)}\right)$$

证毕。

4.3 高效的响应解码

4.3.1 Lasso 回归

本文未对 Lasso(Least Absolute Shrinkage and Selection Operator)回归方法[12]进行详细的叙述，只是为了更好地理解响应解码的流程，把该回归方法当作一种知识储备进行阐述，因此我们一概省略 Lasso 回归一些数学上的偏差分析、方差分析，而只学习其定义和性质。

Lasso 回归方法是一种压缩估计，它通过构造一个惩罚函数来得到一个较为精炼的模型，使其压缩一些回归系数，即强制系数绝对值之和小于某个固定值，同时设定一些回归系数为 0。因此，该回归方法保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。

Lasso 回归的惩罚函数是在一般线性回归的惩罚函数的基础上添加 L1 正则化，具体表达式如下：

$$J = \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2 + \lambda \| \omega \|$$

Lasso 回归一开始主要是为了解决普通线性回归的过拟合问题，其由于容易使部分参数变为 0 而被用于特征提取，在响应解码过程中主要用于估计部分较大的频数，而那些频数较小的项目直接作归 0 处理。我们的主要思路是将每个项目的频数作为待估计的参数，然后通过 Lasso 回归模型进行求解，而求解的过程便是不断地调整参数使得惩罚函数变得更小。

4.3.2 布隆过滤器的位频数估计

首先解释位频数(Bit Frequency)的概念，上文提到过，布隆过滤器是一个长度为 k 的二进制数，将样本中第 i 位为 1 的频数记为 t_i。在真正实现 RAPPOR 算法时，为了减少碰撞，我们引入了队列的概念，将样本第 j 个队列中第 i 位为 1 的频数记为 t_{ij}。但是，由于 RAPPOR 算法中添加了随机响应机制，因此无法直接获取这类频数数据。这个问题可以利用第 2 节中的方法得到解决，我们需要构造 t_{ij} 与上传数据的关联，关联如下：

$$C_{ij} = t_{ij} \times q^* + (N_j - t_{ij}) \times p^* \tag{15}$$

先对式(15)中的变量进行阐释，C_{ij} 表示上传数据中的位频数，即第 j 个队列中第 i 位为 1 的频数，这个数据是可以直接获取的，因此它是一个已知变量。p^* 和 q^* 在 4.2.2 节中已有提及，此处不再赘述。N_j 表示第 j 个队列的数目，这个变量的值也可以从上传数据中通过计数得到，因此它也是一个已知变量。我们可以得到 t_{ij} 的无偏估计公式：

$$t_{ij} = \frac{C_{ij} - \left(p + \frac{1}{2}qf - \frac{1}{2}pf\right)N_j}{(1-f)(q-p)} \tag{16}$$

经过计算后我们可以得到一个 km 长的向量，我们将其记为 $\boldsymbol{Y}$。它由计算得到的无偏估计的频数构成。在 Lasso 的惩罚函数中，表现为 y_i。

4.3.3 MAP 矩阵

该矩阵是一个 $km \times M$ 的矩阵，将该矩阵记为 $\boldsymbol{X}$，其中 M 表示候选的项目数量。以集合$\{a,a,a,b,b,b,b,c\}$为例，在这个例子中 $M=3$，因为候选的项目只有$\{a,b,c\}$。该矩阵起到一个映射的作用，而该映射的定义域是项目集合，值域是布隆过滤器的集合。该矩阵的第一列表示第一个项目所有的布隆过滤器的情况，这是因为 *RAPPOR* 算法在实施时会给真实值分配一个队列数，那么每个真实值就有可能分到 m 个队列中的一个，并且分到每个队列的概率都是相同的。因此，某一候选项目的真实值通过哈希函数映射到布隆过滤器的过程就是在 m 个布隆过滤器中随机选择一个(因为队列数是随机分配的)。也就是说，每个候选项目都会对应 m 个布隆过滤器。因此，每个候选项目都会对应一个大小为 km 的向量，也就是 MAP 矩阵的一列。该矩阵的生成需要结合算法中的哈希规则，对每个 RAPPOR 算法都有一个固定的 MAP 矩阵。需要注意的是，此矩阵的生成一定要严格匹配算法使用的哈希规则，保证算法的一致性。

4.3.4 解码

本节主要介绍如何将获得的两个矩阵嵌入 Lasso 回归模型。本文以 $x_1,\cdots,x_M$ 表示待估计的频数。为了计算方便，将 $\boldsymbol{X},\boldsymbol{Y}$ 合并成一个增广矩阵 $\boldsymbol{Z}$，表达式如下：

$$\boldsymbol{Z} = [\boldsymbol{X},\boldsymbol{Y}]$$

$\boldsymbol{Z}$ 是一个 $km \times (M+1)$ 的矩阵。为了契合 Lasso 回归模型，需要将回归模型的惩罚函数嵌入到解码过程中。上文已提到过惩罚函数的 y_i 是矩阵 $\boldsymbol{Y}$ 的元素，可以直接使用。下面主要介绍如何构造 $f(x_i)$。

对于给定的参数 $x_1,\cdots,x_M$，我们可以计算出某一队列内某一位上为 1 的频数。我们以第 j 队列第 i 位为例，可得估计频数为：

$$\sum_{l=1}^{M}\frac{1}{m}(2^{\boldsymbol{X}[m\times j+i][l]}-1)(x_l)^{\boldsymbol{X}[m\times j+i][l]} \tag{17}$$

该式与 Lasso 回归模型惩罚函数的 $f(x_i)$ 对应。

接下来就能很自然地进行损失函数的计算以及参数的调整，当损失函数达到最小值时，此时对应的参数即为各项目的频数。

4.4 实验分析

4.4.1 数据说明

本文使用的数据集是 2021 年 4 月 18 日美国各州确诊新冠肺炎的人数，该数据来源于 johns Hopkins[1)]。在发起问卷统计确诊人数时，为了保障病人的隐私，需要在 RAPPOR 机制下对受访者上传的信息进行处理。受访者上传的信息格式是“州名”+“1”或者是“0”。1 表示患有新冠肺炎，0 表示不患新冠肺炎。因为 0 不是敏感信息，所以不用处理，我们只对敏感信息进行处理和分析。在进行实验时我们并没有对所有的州进行统计分析，而是抽样了 26 个州作为实验的对象，那么我们需要处理的项目数量就是 26，即我们需要估计 26 个州确诊患有新冠肺炎的人数。

1) https://coronavirus.jhu.edu/

我们处理的数据集大小为 200 000，这是按照美国人口的 1%进行抽样后得到的数据。根据隐私预算做了 3 组实验，其中隐私预算 ε 分别取 10，1，0.01。实验参数如表 2 所列。

表 2　实验所选参数

Table 2　Experimental parameters

k	h	m	p	q	f	ε
16	2	64	0.20	0.80	0.200	10
16	2	64	0.08	0.92	0	1
16	2	64	0.20	0.80	0.855	0.01

4.4.2　结果分析

本文分别对估计量的拟合程度和单个州的估计偏差做了分析，此外，模拟了攻击者获取信息的过程，并通过一个实验描述了永久随机响应在保障 RAPPOR 机制安全性方面的重要作用。

频率特征提取的效果具体如图 4—图 6 所示。

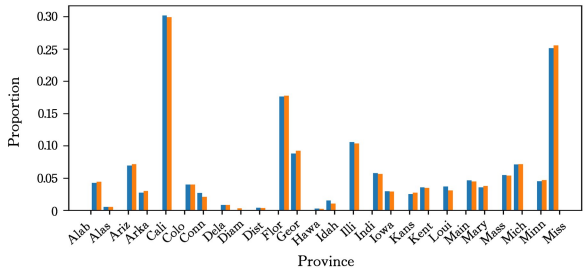

图 4　ε=10 时的实验结果(电子版为彩图)

Fig. 4　Experimental result when ε=10

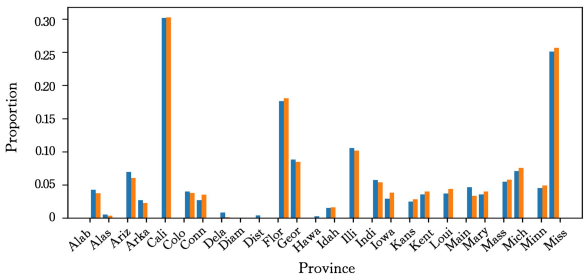

图 5　ε=1 时的实验结果(电子版为彩图)

Fig. 5　Experimental result when ε=1

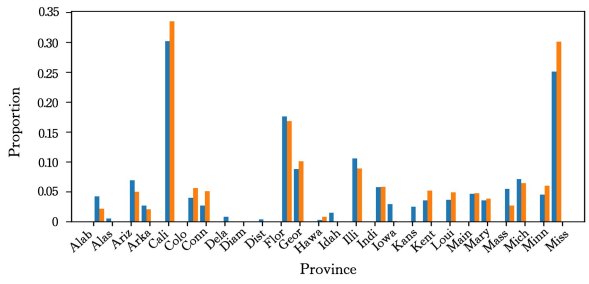

图 6　ε=0.01 时的实验结果(电子版为彩图)

Fig. 6　Experimental result when ε=0.01

图 4—图 6 中，纵坐标表示属性值占总数的比值，横坐标为美国各州的名称简写，简写规则是取州名的前四个字母；直方图中的蓝色代表原始数据中的真实频数，即通过普查统计所得的数据，将这个数据作为比较估计精确度的参考；黄色表示估计的频数，即通过算法处理后所得的估计值，从所绘制的直方图来看，估计的结果基本能够拟合真实的数据分布。除此之外还看到，对于不同的隐私预算 ε[13-14]，估计结果的拟合程度是不同的。ε 越小，拟合程度就越差，ε 越大，拟合程度就越好，这也正对应了差分隐私的定义，隐私预算越小，添加的扰动机制就越大，那么就会增加解码分析的难度，自然估计的结果也就越差。实际上，当 ε 过小时，差分隐私技术将会失去其统计上的意义，因为这样将得不到理想的可拟合的结果。一般情况下，会让 ε 有一个下确界，这个下确界通常是为了防止出现隐私预算太低导致拟合效果不理想的情况[15]。在分布上可以看到，RAPPOR 算法可以取得很好的结果，基本上可以获取数据的整体特征。接下来需要观察对于某一个属性值 RAPPOR 算法的估计效果，我们以 Georgia 州为例，收集 100 次 RAPPOR 算法后该州确诊人数的估计值，结果如图 7 所示。

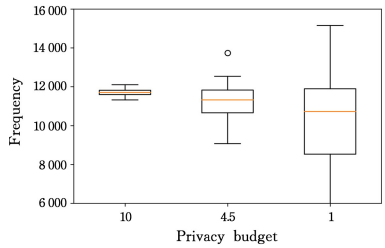

图 7　Georgia 州确诊人数估计箱型图(电子版为彩图)

Fig. 7　Estimation of people diagnosed in Georgia

图 7 给出了 3 种隐私预算下 Georgia 州确诊人数的估计结果的分布。其中，黄色水平线表示 100 次估计值的均值，横坐标表示隐私预算，纵坐标表示频数。其中，Georgia 州真实的确诊人数为 11 163，因此可以看出，在 3 种不同隐私预算的情况下，估计结果基本无偏，但是随着 ε 逐渐减小，估计结果的方差逐渐增大，这一点从箱型图纵向上的长度可以看出，箱型图纵向上长度越长方差就越大，估计精度越低；长度越短，估计精度就越高。

最后通过实验解释 RAPPOR 需要设定两个随机响应机制的原因。本文在分析过程中通过隐藏永久随机响应机制，来模拟攻击者无法获取用户在本地的操作信息，具体参数信息如表 3 所列。

表 3　模拟攻击时使用的实验参数

Table 3　Experimental parameters used in simulating attack

k	h	m	p	q	f
16	2	64	0.1	0.9	0.95

模拟攻击的结果如图 8 所示。

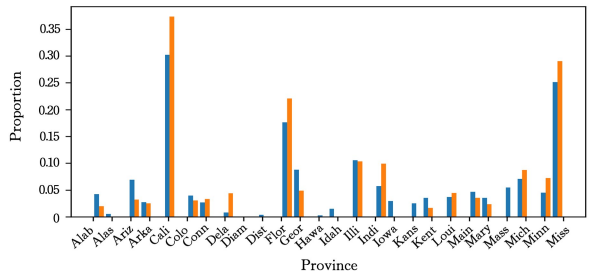

图 8　攻击者模拟分析的实验结果(电子版为彩图)

Fig. 8　Experimental results of attacker simulation analysis

图 8 中黄色柱表示攻击者重构估计的频率，蓝色柱表示真实的频率，很显然攻击者重构的效果不是十分理想，这就体现出了短暂随机响应在安全方面的重要作用。

结束语 差分隐私为个人的隐私保护添加了一道极强的保护机制，但是如何保护隐私并不困难，真正困难的是如何解决各种各样的分析任务。差分隐私是继中心化差分隐私后新兴的隐私保护模型，其打破了中心化差分隐私中关于可信第三方数据收集者的假设，在用户端对数据进行隐私化处理。目前本地化差分隐私保护技术是隐私保护领域的研究热点，本文主要研究了在这种隐私保护技术下获取数据各属性频数的方法，十分详细地讲解了编码过程和解码过程。在结果分析过程中，既看到了 RAPPOR 算法优秀的性能，也看到了其局限性。RAPPOR 算法的不足之处在于[16-18]：1)对于隐私预算的确定十分经验化，缺少一套系统的标准来确定最满足需求的隐私预算；2)解码的开销太大，对于超大型的数据采集可能需要极大的解码开销；3)在真实场景中，不可能只维护一个静态的 MAP 矩阵，因为在设计时不可能总能将所有候选项目都考虑到，在这种场景下如何使用 RAPPOR 算法又是一个值得讨论的问题。总之，本地化差分隐私保护技术还是一个新兴研究领域，仍有诸多关键问题需要进行深入细致的研究。

参 考 文 献

[1] GEORGINA E, GARY K, ADAM D S, et al. Differentially Private Survey Research[DB/OL]. (2021-03-21)[2021-06-18]. https://j.mp/3jAYXo3.

[2] SAMARATI P, SWEENEY L. Generalizing Data to Provide Anonymity when Disclosing Information[C]// Proceedings of the Seventeenth ACM-SIGACT-SIGMOD-SIGART Symposium on Principles Systems. New York: ACM, 1998: 98-188.

[3] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. l-Diversity: Privacy Beyond k-anonymity[C]// Proceedings of the 22nd International Conference on Data Engineering. Atlanta: IEEE Press, 2006: 24-24.

[4] LI N, LI T, VENKATASUBRAMANIAN S. t-Closeness: privacy Beyond k-Anonymity and l-diversity[C]// Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE). IEEE, 2007: 106-115.

[5] GEORGINA E, GARY K, MARGARET S, et al. Statistically Valid Inferences from Privacy Protected Data [DB/OL]. https://j.mp/2qkWjfj.

[6] DWORK C. Differential Privacy[C]// Automata, Languages and Programming. Venice: Springer, 2006: 1-12.

[7] WARNER S L. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias[J]. Journal of the American Statistical Association, 1965, 60(309): 63-69.

[8] YOSHUA B, REJEAN D, PASCAL V, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research(JMLR), 2003, 3: 1137-1155.

[9] WANG N, XIAO X K, YANG Y, et al. Collecting and Analyzing Multidimensional Data with Local Differential Privacy[C]// IEEE 35th International Conference on Data Enginee Ring(ICDE). Macao, China, 2019: 638-649.

[10] WANG J N, KRASKA T, FRANKLIN M J, et al. CrowdER: Crowdsourcing Entity Resolution[C]// Proceedings of the VLDB Endowment, Istanbul: VLDB Endowment, 2012: 1483-1494.

[11] ULFAR E, VASYL P, ALEKSANDRA K. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response[C]// Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM 2014: 1054-1067.

[12] ROBERT T. Regression Shrinkage and Selection Via the Lasso [J]. Journal of the Royal Statistical Society: series B, 1994, 58(1): 267-288.

[13] JOHN C D, MICHAEL I J. Local Privacy and Statistical Minimax Rates[C]// Proceedings of the IEEE 54th Annual Symposium on Foundations of Computer Science. New York: IEEE Press, 2013: 1592-1592.

[14] DING B, WINSLETT M, HAN J, et al. Differentially Private Data Cubes: Optimizing Noise Sources and Consistency[C]// Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. NEW YORK: ACM, 2011: 217-228.

[15] HARDT M, ROTHBLUM G N. A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis[C]// Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science. New York: IEEE Press, 2010: 61-70.

[16] OBERSKI D L, KREUTERM F. Differential Privacy and Social Science: An Urgent Puzzle[EB/OL]. https://doi.org/10.1162/99608f92.63a22079.

[17] HARDT M, LIGETT K, MCSHERRY F. A Simple and Practical Algorithm for Differentially Private Data Release[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2012: 2339-2347.

[18] YE Q Q, MENG X F, ZHU M J, et al. Survey on Local Differential Privacy[J]. Journal of Software, 2018, 29(7): 1981-2005.

HUANG Jue, born in 1998, postgraduate. His main research interests include artificial intelligence uncertainty.

ZHOU Chun-lai, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include uncertainty in AI and privacy in data science.

(责任编辑：喻藜)