

基于差分隐私保护技术的多方求和查询方法

何贤芒

(东莞理工学院网络空间安全学院, 广东 东莞 523808)

摘要: 差分隐私保护技术因其不需要攻击者先验知识的假设, 而被认为是一种非常可靠的保护机制。然而, 差分隐私保护技术很少在多方环境下使用。鉴于此, 将差分隐私保护技术用于多方环境下数据求和查询问题, 详细讨论了如何通过加入噪声的方法来实现数据的保护, 并证明该方法安全性。

关键词: 多方求和; 隐私保护; 差分隐私; 数据查询

中图分类号: TP301

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2020032

Multi-party summation query method based on differential privacy

HE Xianmang

School of Cyberspace Science, Dongguan University of Technology, Dongguan 523808, China

Abstract: Differential privacy is considered to be a very reliable protection mechanism because it does not require the a prior knowledge for the attacker. However, differential privacy is rarely used in a multi-party environment. In view of this, the differential privacy is applied to the data summation query in multi-party environment. This method was described in detail and proved the security of the method.

Key words: multi-party summation, privacy preservation, differential privacy, data query

1 引言

网络技术的飞速发展, 促使信息量正以超乎人们想象的速度增长。不同的组织间或个人间的合作计算变得越来越重要。不同的数据所有者需要通过合作交流计算信息得到更全面、更有价值

的计算结果。然而, 数据安全性与隐私保护问题制约着合作计算的进行, 甚至在某些情形下参与各方不得不舍弃合作来确保数据的私有与安全。在商业应用上, 多个商家的竞争关系往往为了决策需要合作进行数据挖掘来了解整个市场的情况。例如, 不同的手机运营商需要通过合作计算来了

收稿日期: 2019-10-08; 修回日期: 2020-02-22

通信作者: 何贤芒, x.m.he@163.com

基金项目: 国家自然科学基金(61672303); 广东省普通高校特色创新项目(2018KTSCX221)

Foundation Items: The National Natural Science Foundation of China (61672303), Characteristic Innovation Projects of Universities in Guangdong Province, China (2018KTSCX221)

论文引用格式: 何贤芒. 基于差分隐私保护技术的多方求和查询方法[J]. 网络与信息安全学报, 2020, 6(3): 14-18.

HE X M. Multi-party summation query method based on differential privacy[J]. Chinese Journal of Network and Information Security, 2020, 6(3): 14-18.

解某个地区某些手机品牌的使用情况。各个商家拥有的数据是商家的私有信息,不希望对方知道,也可能商家拥有的数据不宜对外公开(如使用者的通话记录)。在这种情况下,需要多方在合作进行数据挖掘的同时保证各方的私有数据不被泄露。为解决该问题,越来越多的学者将努力方向聚焦在安全多方计算(SMC, secure multi-party computation)^[1-2]的研究中,研究与设计在多方参与不泄露各方私有信息的条件下完成合作计算。安全多方计算使不公开私有信息的合作计算成为可能,其研究促进了各个组织或个人之间的信息流通并且在各个领域拥有广阔的应用前景。目前,安全多方计算研究已经取得了较大的进展,并且日益成为现代密码学的重要研究课题。

然而,由于基础理论研究的复杂性和应用问题的多样性,基于传统的安全多方计算协议设计过于复杂,不易操作。差分隐私保护是一种数据失真的隐私保护技术,采用添加噪声的技术使敏感数据失真但同时保持某些数据或数据属性不变,要求保证处理后的数据仍然可以保持某些统计方面的性质,以便进行数据挖掘等操作。本文拟将其应用在多方环境下数据求和查询问题,在保持查询结果精确性的前提下同时保证其安全性。

2 差分隐私保护技术

差分隐私保护技术建立在坚实的数学基础之上,对隐私保护进行了严格的定义并提供了量化评估方法,使不同参数处理下的数据集所提供的隐私保护水平具有可比较性。因此,差分隐私理论迅速被业界认可。

定义1 差分隐私^[3-4]。一个随机算法 A 满足 ϵ -差分隐私保护,当且仅当对于任何相差仅一个元组的两个集合 D_1 、 D_2 和任何输出 S , 满足如下条件

$$\frac{\Pr(A(D_1))}{\Pr(A(D_2))} \leq e$$

这里的 ϵ 是使用者指定的常数, D_1 和 D_2 至多相差一个元组, e 是自然对数常数。从数学上看,只要这个参数 ϵ 足够小,攻击者很难区分出对同样的输出 S , 查询函数到底是作用在 D_1 还是

在 D_2 上。当参数等于 0 时,输出的仅是噪声才能满足上述要求,所以参数 ϵ 只有大于 0 才有实际的意义。同样条件下,参数 ϵ 越小私密性越好。

一般来说,差分隐私保护的实现技术是基于拉普拉斯分布来完成的,对于任何查询函数 q 作用于数据集 D 上,返回值是 $q(D)+x$, 其中的 $q(D)$ 是查询的真实值, x 是满足拉普拉斯分布 $f(\mu, b)$ 的采样值: $x \propto f(\mu, b)$, 这里 $b = \frac{\Delta q}{\epsilon}$, Δq 定义参考

定义2。容易证明,这个方法实现了差分隐私保护技术的要求。可以注意到, ϵ 越大, b 就越小。

定义2 敏感度。 Δq 是查询函数 q 的敏感度,其定义如下。

$$\Delta q = \max_{D_1, D_2} |q(D_1) - q(D_2)|$$

数据集 D_1 和 D_2 之间至多相差一个元素。

3 目前方案

多方安全求和协议是一类重要基础的协议。目前已经有一系列文献^[5-14]对安全求和协议进行了研究。这些方法各有优劣。文献[5]提出从多个参与者中选择一个主站点,让其产生一个随机数加上自己的数据发给下一个参与者,后面的参与者加上自己的真实数据依次传下去,最终传回主站点再减去随机数之后进行广播。此方法通信性能好,但安全性太差,相邻的两个参与者共谋便可得到中间参与者的数据。

为了解决合谋问题,文献[6]将每位参与者自己输入的数据随机划分成 n 份,然后每位参与者把分成的 n 份分别发给所有的参与者,每位参与者分别计算自己收集到的数据的和,然后广播给其他参与者,最后每位用户再次分别计算自己收集到的数据,计算出总和。这种算法固然相对于文献[5]安全很多,但其通信需要 n^2 次,通信复杂度较高。文献[7]提出基于博弈论的安全多方求和协议,主要考虑了用户不诚实提供数据的情况,每个用户数据分成 k 份,每次计算做 k -轮迭代,做二次前后结果是一致的,发布求和结果,如果不一致,则说明有站点提供不真实数据,主站点重新发起求和运算,直到连续两次运算得到的和一样。文献[8]为了解决通信度高的问题,采用了公钥加密与随机函数,提出了一种能提高安全性

又能降低通信复杂度的协议。张恩等^[9]结合博弈论和密码算法,提出了一种基于电路计算的理性安全多方求和协议。Mehnaz 等^[10]提出了一种基于诚实模型的安全求和协议,并且应用在机器学习中。Ashouri-Talouki 等^[11]提出了无须安全信道的3个安全求和协议。仲红^[12]提出了基于安全多方求和的多候选人电子选举方案。文献[13]讨论了在通用可组合框架下研究安全多方计算的公平性问题。Liu 等^[14]提出了一种基于双粒子 Bell 态的量子安全多方求和协议。Jung 等^[15]提出了一种没有安全通信信道或没有可信密钥分发者的安全求和协议。

4 多方环境下求和协议

目前为止,差分隐私保护技术难以做到准确的数据查询,而本文可以用差分隐私这种方法实现在多方参与情况下数据准确求和查询,本文的安全性是基于差分隐私保护这种机制,具有安全性保证;从查询速度来说,与目前的其他数据交换方法相同,因此通信量是一致的。

本节提出一个在多方环境下的数据库统计查询方法,设参与方的数量 n 至少是3,显而易见,当 $n=2$ 时,求和查询是无法做到数据安全的,步骤如下。

步骤 1 参与的各方 $(P_1, P_2, \dots, P_n)$ 各有一个值 x_i , 各自选择一个比较小的正数 ε_i 和一个实数 u 。

步骤 2 各方 P_i 生成了 $n-1$ 个数 $x_{i,1}, x_{i,2}, \dots, x_{i,n-1}$, 满足拉普拉斯分布 $f(u, b)$, 这里的 $b = \frac{1}{\varepsilon}$ 。

步骤 3 任意的参与方 $P_i (1 \leq i \leq n)$, 通过安全通道或者通过加密方法, 发送 $x_{i,j}$ 到参与方 $P_j (1 \leq j \leq n, i \neq j)$ 。

步骤 4 任意的参与方 P_i , 计算 $Y_i = x_i - (x_{i,1} + x_{i,2} + \dots + x_{i,n-1}) + (x_{1,i} + x_{2,i} + \dots + x_{n,i})$, 并公布 Y_i 的值; 其中, $x_{i,j}$ 为所述各方中的参与方 P_i 中发送给 P_j 的数值, $x_{j,i}$ 为参与方 P_i 接收参与方 P_j 发送的数值。

步骤 5 参与的各方 P_i , 根据各方的公布值 Y_i 计算总和。

例 1 演示了上述求和全过程, 定理 1 证明上述方法满足隐私保护技术的要求。

定理 1 上述多方求和计算方法满足 $\max\{\varepsilon_i\}$ -差分隐私。

证明 设 D_1, D_2 是任意两个兄弟集合, 仅差一个数据元组, o 是算法 A_2 的一个输出, T 是算法 A_1 可能输出集合。

情形 1 A_1, A_2 算法输出是离散类型

$$\Pr(A_2(A_1(D_1) = o) = \Pr\left(\sum_{t \in T} \Pr(A_1(D_1) = t)\right) \cdot$$

$$\Pr(A_2(t) = o)$$

根据定义 1

$$\Pr\left(\sum_{t \in T} \Pr(A_1(D_1) = o)\right) \leq e^\varepsilon \Pr\left(\sum_{t \in T} \Pr(A_1(D_2) = o)\right)$$

$$\Pr(A_2(t) = o)$$

由上面的式子可以得出

$$\Pr(A_2(A_1(D_1) = o) \leq e^\varepsilon \Pr\left(\sum_{t \in T} \Pr(A_1(D_2) = t)\right) \cdot$$

$$\Pr(A_2(s) = o) = e^\varepsilon \Pr(A_2(A_1(D_2)) = o)$$

$$\Pr(A_2(t) = o)$$

情形 2 A_1, A_2 算法输出是连续类型

$$\Pr(A_2(A_1(D_2)) = o) =$$

$$\int_{t \in T} \Pr(A_1(D_1) = t) \cdot \Pr(A_2(t) = o) dt$$

类似地有

$$\Pr(A_2(A_1(D_1)) = o) \leq e^\varepsilon \int_{t \in T} \Pr(A_1(D_2) = t) \cdot$$

$$\Pr(A_2(t) = o) dt = e^\varepsilon \Pr(A_1(D_2) = o)$$

上述的证明表明任意参与方 P_i , 在其拥有的值 x_i 基础上增减多个满足拉普拉斯分布的数 $x_{i,j}$, 其依然满足 $\max\{\varepsilon_i\}$ -差分隐私的要求。

从过程来说, 这个协议通信次数需要 $O(n^2)$ 。结合博弈论等方法, 通信量可以降低到 $O(kN)$, $k \geq 2$ 是可能合谋的参与方的数量。在通信量降低的同时, 其安全性依然满足定理 1 的要求。

从协议的执行过程来看, 参数 ε 的选择与求和查询结果无关, 只跟加入的噪声大小有关。考虑到噪声的加入最后全部抵消, 因此, 定理 2 是显而易见的, 参数 ε 的选择不影响查询结果。

定理 2 参数 ε 的选择对求和结果没有影响。

例 假设有 4 个参与方: P_1, P_2, P_3, P_4

步骤 1 设 P_1, P_2, P_3, P_4 这 4 个参与方各

一个数, $x_1 = 3$, $x_2 = 4$, $x_3 = 6$, $x_4 = 7$, 同时假设各方选定的 $\epsilon = 0.01, u = 0$ 。

步骤 2 设参与方 P_1, P_2, P_3, P_4 各生成 3 个满足拉普拉斯 $f(100, 0)$, 此处的 $b=100$, 由 $b = \frac{1}{\epsilon} = \frac{1}{0.01}$ 获得的参数为 $x_{1,2} = 3.5$, $x_{1,3} = 121.2$, $x_{1,4} = -129.2$, $x_{2,1} = -2.5$, $x_{2,3} = 87.5$, $x_{2,4} = -12.5$, $x_{3,1} = -21.4$, $x_{3,2} = 176.4$, $x_{3,4} = 44.5$, $x_{4,1} = -12.3$, $x_{4,2} = 20.4$, $x_{4,3} = 78.6$ 。

步骤 3 如图 1 所示, 进行数据交换, 具体如下。

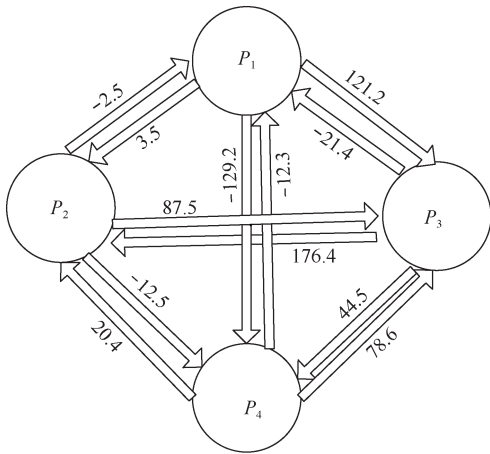


图 1 数据交换示例
Fig 1 Example of data exchange

P_1 给 P_2 的数字是 3.5, P_1 给 P_3 的数字是 121.2, P_1 给 P_4 的数字是 -129.2。

P_2 给 P_1 的数字是 -2.5, P_2 给 P_3 的数字是 87.5, P_2 给 P_4 的数字是 -12.5。

P_3 给 P_1 的数字是 -21.4, P_3 给 P_2 的数字是 176.4, P_3 给 P_4 的数字是 44.5。

P_4 给 P_1 的数字是 -12.3, P_4 给 P_2 的数字是 20.4, P_4 给 P_3 的数字是 78.6。

步骤 4 根据公式 $Y_i = x_i - (x_{i,1} + x_{i,2} + \dots + x_{i,n-1}) + (x_{1,i} + x_{2,i} + \dots + x_{n,i})$ 。

P_1 计算如下。

$$Y_1 = 3 - (3.5 + 121.2 - 129.2) + (-2.5 - 21.4 - 12.3) = -28.7$$

P_2 计算如下。

$$Y_2 = 4 - (-2.5 + 87.5 - 12.5) + (3.5 + 176.4 + 20.4) = 131.8$$

P_3 计算如下。

$$Y_3 = 6 - (-21.4 + 176.4 + 44.5) + (121.2 + 87.5 + 78.6) = 93.8$$

P_4 计算如下。

$$Y_4 = 7 - (-12.3 + 20.4 + 78.6) + (-129.2 - 12.5 + 44.5) = -176.9$$

步骤 5 计算 $Y_1 + Y_2 + Y_3 + Y_4 = 20$, 即总和。

根据定理 1, 上述的多方求和计算方法满足 0.01-差分隐私保护技术。

5 结束语

虽然差分隐私保护技术得到了学者的广泛关注, 但如何在实际中、在多方环境下应用差分隐私保护技术依然是个开放问题。鉴于此, 本文提出了一个基于差分隐私保护模型多方环境下求和查询方法, 下一步将其推广到 MAX 与 MIN 查询, 并且在这个方面取得进展。

参考文献:

- [1] BARVE R, SHRIVER E, GIBBONS P B. Modeling and optimizing I/O throughput of multiple disks on a bus[J]. ACM Sigmetrics Performance Evaluation Review, 1999, 27(1): 83-92.
- [2] LU Y P, DAVID H C D. Performance study of iSCSI-based storage subsystems[J]. IEEE Communications Magazine, 2003, 41(8): 76-82.
- [3] DWORK C. Differential privacy[C]//International Colloquium on Automata, Languages, & Programming. 2006: 1-12.
- [4] PROSERPIO D, GOLDBERG S, MCSHERRY F. Calibrating data to sensitivity in private data analysis[C]//The Third Conference on Theory of Cryptography. 2006: 265-284.
- [5] CLIFTON C, KANTARCIOGLU M, VAIDYA J, et al. Tools for privacy preserving distributed data mining [J]. Sigkdd Explorations, 2002, 4(2): 28-34.
- [6] 罗永龙, 徐致云, 黄刘生. 安全多方的统计分析问题及其应用[J]. 计算机工程与应用, 2005, 41(24): 141-143.
LUO Y L, XU Z Y, HUANG L S. Multivariate statistical analysis and its application[J]. Computer Engineering and Applications, 2005, 41 (24): 141-143.
- [7] 张国荣, 印鉴. 基于博弈论的安全多方求和方法[J]. 计算机应用研究, 2009, 26(4): 1497-1499.
ZHANG G R, YIN J. Multi-party secure sum computation based on game theory[J]. Journal of Computer Applications, 2009, 26 (4): 1497-1499.
- [8] 王峥, 郝林, 刘义成. 基于公钥加密的安全多方求和协议[J]. 计算机应用研究, 2017(4): 179-182.
WANG Z, HAO L, LIU Y C. Secure sum protocol based on public key encryption [J]. Journal of Computer Applications, 2017 (4): 179-182.
- [9] 张恩, 朱君哲, 范海菊, 等. 基于电路计算的理性安全多方求和协议[J]. 密码学报, 2019, 6(1): 126-135.
ZHANG E, ZHU J Z, FAN H J, et al. Rational secure multiparty sum protocol based on circuit computing [J]. Chinese Journal of Cryptography, 2019, 6 (1): 126-135.

- [10] MEHNAZ S, BELLALA G, BERTINO E. A secure sum protocol and its application to privacy-preserving multiparty analytics[C]//The 22nd ACM on Symposium on Access Control Models and Technologies. 2017: 219-230.
- [11] ASHOURI-TALOUKI M, BARAANI-DASTJERDI A. Cryptographic collusion-resistant protocols for secure sum[J]. International Journal of Electronic Security and Digital Forensics, 2017, 9(1): 19-34.
- [12] 仲红, 黄刘生, 罗永龙. 基于安全多方求和的多候选人电子选举方案[J]. 计算机研究与发展, 2006, 43(8): 1405-1410.
ZHONG H, HUANG L S, LUO Y L. A multi-candidate electronic voting scheme based on secure sum protocol[J]. Journal of Computer Research and Development, 2006, 43(8): 1405-1410.
- [13] 田有亮, 彭长根, 马建峰, 等. 通用可组合公平安全多方计算协议[J]. 通信学报, 2014(2):54-62.
TIAN Y L, PENG C G, MA J F, et al. Universally composable secure multiparty computation protocol with fairness[J]. Journal on Communications, 2014(2): 54-62.
- [14] LIU W, WANG Y B, FAN W Q. An novel protocol for the quantum secure multi-party summation based on two-particle bell states[J]. International Journal of Theoretical Physics, 2017, 56(9): 2783-2791.
- [15] JUNG T, LI X Y, WAN M. Collusion-tolerable privacy-preserving sum and product calculation without secure channel[J]. IEEE Transactions on Dependable and Secure Computing, 2015, 12(1): 45-57.

[作者简介]



何贤芒(1981-), 男, 浙江三门人, 东莞理工学院副教授, 主要研究方向为数据安全与隐私保护。