第41卷 第8期 西 安 交 通 大 学 学 报 Vol.41 №8
2007年8月 JOURNAL OF XI'AN JIAOTONG UNIVERSITY Aug. 2007

一种新型的分布式隐私保护计算模型及其应用

余智欣[1,2]，黄天戍[1]，杨乃扩[2]，汪阳[3]
(1.武汉大学电子与信息学院，430079，武汉；2.曼彻斯特大学计算机学院，M139PL，英国曼彻斯特；
3.武汉理工大学信息工程学院，430070，武汉)

摘要：针对分布式数据共享及计算中的隐私保护问题，提出了一种适用于大规模分布式环境的隐私保护计算模型(PPCMLS)，该模型的核心为隐私安全模块，其将计算划分为本地计算和全局计算.通过综合运用同态加密、安全点积协议、数据随机扰乱算法等多种安全技术，在实现了多个节点在一个互不信任的分布式环境下合作计算的同时，任何节点无法获取其他节点的隐私信息及敏感中间计算结果.据此，又给出了基于该模型的分布式隐私保护方差计算、分布式隐私保护数据聚类算法.安全及动态性分析结果表明，该模型及其应用算法既可保证隐私数据的安全性，又避免了繁琐的一对多的交互加密过程，并在节点变化时，恢复计算仅涉及到变化的节点和构成隐私安全模块的3个节点，从而满足了大规模分布式环境所要求的高效性和良好的动态适应性.

关键词：隐私保护计算；同态加密；安全点积协议；随机扰乱
中图分类号：TM73；TP393 **文献标识码**：A **文章编号**：0253-987X(2007)08-0954-05

Novel Privacy-Protecting Distributed Computation Model and Its Applications

Yu Zhixin[1,2], Huang Tianshu[1], Yang Naikuo[2], Wang Yang[3]
(1. School of Electronics and Information, Wuhan University, Wuhan 430079, China; 2. School of Computer, The University of Manchester, Manchester M139PL, UK; 3. School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China)

Abstract: For protecting the privacy in the large-scale distributed database sharing and computation, a generic privacy-protecting computation model supporting the privacy-protecting data computation and analysis is proposed. The key of this model is the 'privacy security module' which divides the distributed computation into 'local computation' and 'global computation'. By using homomorphic encryption, secure product protocol, random permutation and some other secure technologies comprehensively, the model achieves distributed data computations without compromising the privacy of both data record and sensitive intermediate computation result. Two privacy-protecting distributed algorithms based on this model, namely, the privacy-protecting distributed variance algorithm and privacy-preserving distributed k-means clustering algorithm, have also been introduced. The security and dynamic behavior analysis show that the proposed model and algorithms doesn't require any one-to-all interactions encryption, and the recovery computation only involves the changed nodes and the three nodes composing privacy security module when nodes change. So it's suitable for large-scale and dynamic distributed environments.

Keywords: privacy-preserving computation; homomorphic encryption; secure product protocol; random permutation

---

收稿日期：2006-12-22. 作者简介：余智欣(1981～)，男，博士生；黄天戍(联系人)，男，教授，博士生导师. 基金项目：公安部科研基金资助项目(200342-823-01)；武汉市信息产业基金资助项目(2004544222)；曼彻斯特大学计算机学院网络安全研究组科研项目.

数据的隐私性保护是大规模分布式数据共享及计算得以广泛应用的主要障碍，数据是分布在不同的地点，分属于不同组织，在进行合作运算的同时相互之间并不希望其他参与方知道自己的原始数据及一些能推测出有用信息的敏感中间计算结果，这就迫切需要发展一种能适用于在大规模分布式环境下(数据网格)具有隐私保护特性的通用分布式计算模型，使得各参与节点在无法获取其他节点信息及敏感中间计算结果的条件下协作计算，同时满足大规模分布式计算所要求的高效性和动态性.

本文通过综合运用同态加密算法[1-6]、安全点积协议、数据随机扰乱算法[7]等多种安全手段，提出了一种面向大规模分布式环境的隐私保护计算模型(PPCMLS). 基于该模型，本文又提出了基于安全点积协议的两方隐私保护计算协议，以确保中间敏感计算结果，同时还给出了分布式隐私保护方差算法和分布式隐私数据聚类算法. 由于加密是每个节点在本地独立完成的，所以当节点发生变化时，无繁琐的、涉及全局的交互式加密，恢复计算仅涉及到隐私安全模块和变化节点.

1　两方隐私保护计算协议

PPCMLS 模型将计算划分为本地计算和全局计算，通过本文自行提出的两方隐私保护计算协议，可对中间敏感计算结果进行隐私保护. 该协议基于安全点积协议[8]，描述如下.

输入：X 的 2 个私有数据(x_1,x_2)，Y 的 2 个私有数据(y_1,y_2).

输出：(x_1y_1/x_2y_2)、$x_1y_1+x_2y_2$、$(x_1+y_1)(x_2+y_2)$和$(x_1+y_1)/(x_2+y_2)$，但 X 不可能知道(y_1,y_2)，同时 Y 不可能得知(x_1,x_2).

(1)对于(x_1y_1/x_2y_2)：隐私保护计算比较简单，因为$(x_1y_1/x_2y_2)=(x_1/x_2)(y_1/y_2)$，所以可以通过 Y 仅把(y_1/y_2)的值发送给 X，从而计算出(x_1y_1/x_2y_2)，而不泄漏(y_1,y_2)的信息.

(2)对于 $x_1y_1+x_2y_2$：①把(x_1,x_2)和(y_1,y_2)看成 2 个二维矢量，$x_1y_1+x_2y_2$ 则是这 2 个矢量的点积，通过安全点积协议[8]可以实现隐私计算，并让 Y 得到 $XY=\sum_{k=1}^{2}x_ky_k+r=(x_1y_1+x_2y_2)+r$($r$ 仅被 X 所知道)；②Y 将 $x_1y_1+x_2y_2+r$ 发送给 X，X 将 r 从 $x_1y_1+x_2y_2+r$ 中减去，从而得到 $x_1y_1+x_2y_2$；③X 将最后的结果 $x_1y_1+x_2y_2$ 发送给 Y.

(3)对于$(x_1+y_1)(x_2+y_2)$：①Y 产生 3 个随机数 k、α 和 β，Y 将 $\lambda=k/\alpha\beta$ 发送给 X；②X 和 Y 用3.1节中的计算 $x_1y_1+x_2y_2$ 的方法计算$(x_1,1)$与$(\beta,\beta y_1)$及$(x_2,1)$与$(\alpha,\alpha y_2)$的点积，X 得到$\beta(x_1+y_1)$、$\alpha(x_2+y_2)$而不泄漏给 Y，计算 $\beta(x_1+y_1)\alpha(x_2+y_2)\lambda$ 得到 $k(x_1+y_1)(x_2+y_2)$；③X 将 $k(x_1+y_1)(x_2+y_2)$发送给 Y，Y 计算出 $k(x_1+y_1)(x_2+y_2)$并将其发送给 X.

(4)对于$(x_1+y_1)/(x_2+y_2)$：如同 $k(x_1+y_1)(x_2+y_2)$，但不同之处是 Y 仅产生 2 个随机数 α、β，将 $\lambda=\alpha/\beta$ 发送给 X[8].

2　大规模分布式环境下的隐私保护计算模型

PPCMLS 模型如图 1 所示，其核心是一个隐私安全模块，由密钥生成子模块、全局计算子模块、中间结果隐私保护子模块构成，它们分别部署于 3 个互不勾结的、遵循半诚实安全模型的节点上. 这 3 个节点既可以是参与计算的 3 个节点，也可以是 3 个可信的外部节点.

为了便于阐述，假定文中提到的这 3 个节点同时也都是 3 个数据提供及运算的节点. 模型详细的执行步骤描述如下.

(1)密钥生成子模块将产生一对用于同态加密的密钥对(K_e,K_d)，将公钥 K_e 转发给全局计算子模块，并用 E 表示加密函数.

(2)全局计算子模块将 K_e 广播给各个参与计算的节点.

(3)中间结果隐私保护子模块生成一随机数对(v,v')，$v+v'=0$(加同态)或 $vv'=1$(乘同态)，将 $E(v')$发送给全局计算子模块.

(4)每个节点完成各自的本地计算，并将本地计算结果 d_i 用 K_e 进行加密，再将 $E(d_i)$ 发送给全局计算子模块.

(5)全局计算子模块根据同态加密的性质(加同态或乘同态)，通过 $E(v')\prod_{i=1}^{h}E(d_i)$ 得到 $E\left(\left(\sum_{i=1}^{h}d_i\right)+v'\right)$(加同态) 或 $E\left(\prod_{i=1}^{h}(d_i)v'\right)$(乘同态)，其中 h 是一次合作运算涉及到的节点数，可以是部分节点，也可以是全部节点，$1\leqslant h\leqslant n$.

(6) $E\left(\left(\sum_{i=1}^{h}d_i\right)+v'\right)$或 $E\left(\prod_{i=1}^{h}(d_i)v'\right)$ 被送到密钥生成子模块，通过私钥 K_d 对其解密，从而得到

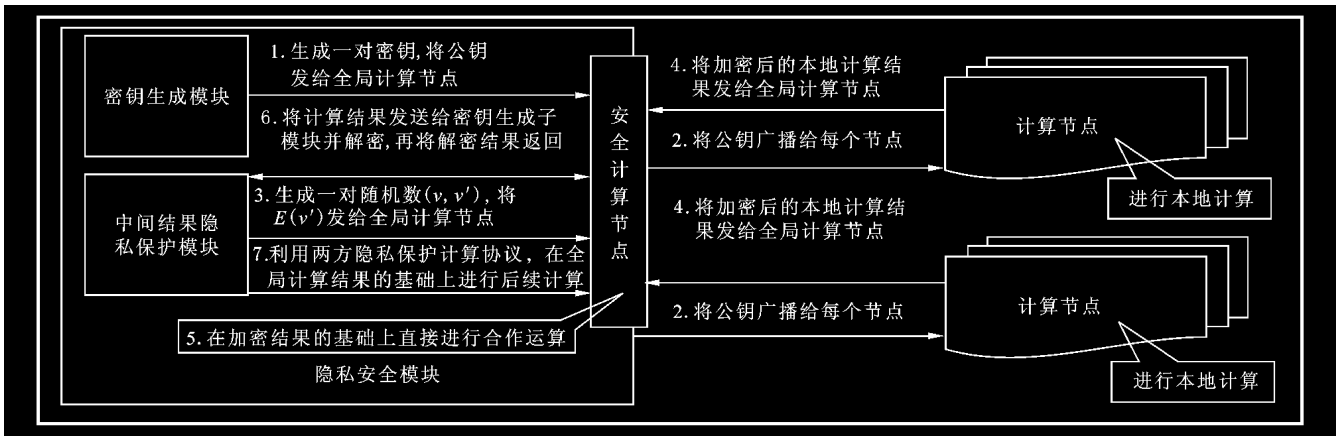

图1 隐私保护计算模型

$\sum_{i=1}^{h}(d_i)+v'$ 或 $\prod_{i=1}^{h}(d_i)v'$.

(7)通过多次重复步骤(5)、(6),得到多个 $\sum_{i=1}^{h}(d_i)+v'$ 或 $\prod_{i=1}^{h}(d_i)v'$. 全局计算子模块和中间结果隐私保护子模块共同在 $\left(\left(\sum_{i=1}^{h}d_i+v'\right)_j+v\right)$ 和 $\left(\left(\sum_{i=1}^{h}d_i+v'\right)_k+v\right)$ 或 $\left(\prod_{i=1}^{h}(d_i)v'\right)_j v$ 和 $\left(\prod_{i=1}^{h}(d_i)v'\right)_k v$ 之间使用两方隐私计算协议.

3 分布式隐私保护方差计算实例

3.1 $\{F_1,F_2,\cdots,F_n\}$个节点参加计算的具体步骤

(1)根据隐私安全模型的步骤(1)~步骤(4),F_1(密钥生成子模块)生成(K_e,K_d),各个节点得到公钥 K_e,F_3(中间结果隐私保护子模块)将 $E(v')$发送给 F_2(全局计算子模块). 各节点分别计算 x_i/n 和$-x_i/n$,并将 $E(x_i)$、$E(x_i/n)$和 $E(-x_i/n)$发送给 F_2.

(2)根据隐私安全模型的步骤(5)、步骤(6),F_2 通过计算 $\prod_{i=1}^{n}E(x_i/n)E(v')E(x_i)$ 和 $\prod_{i=1}^{n}E(-x_i/n)\cdot E(v')E(x_i)$,得到多个 $E(x_i+\bar{x}+v')$和 $E(x_i-\bar{x}+v')$,并将它们交给 F_1 解密;F_1 得到$(x_i+\bar{x}+v')$和$(x_i-\bar{x}+v')$,将它们返回给 F_2.

(3)根据隐私安全模型的步骤(7),F_2($(x_i+\bar{x}+v')$和$(x_i-\bar{x}+v')$)和 $F_3(v)$共同使用第1节中的两方隐私保护加法协议计算出$(k(x_i+\bar{x}+v')+v)((x_i-\bar{x}+v')+v)$,从而得到 $k(x_i+\bar{x})(x_i-\bar{x})$($k$ 是在两方隐私保护加法协议中仅 F_3 知道的随机数),而 F_2 无法知道 v 的值,同时 F_3 也无法知道$(x_i+\bar{x}+v')$、$(x_i-\bar{x}+v')$的值.

(4)F_2 将 n 个 $k(x_i+\bar{x})(x_i-\bar{x})$相加,得到 $\left(k\sum_{i=1}^{n}((x_i-\bar{x})(x_i+\bar{x}))/n\right)$,并将它发送给 F_3,F_3 计算出 $\left(\sum_{i=1}^{n}((x_i-\bar{x})(x_i+\bar{x}))/n\right)$.

3.2 安全性分析

(1)F_1 作为密钥生成子模块和一个独立的节点,由3.3节中的分析,它仅能得到$(x_i+\bar{x}+v')$和$(x_i-\bar{x}+v')$. v'由 F_3 所产生,F_1 不知道 v',故无法得到任何关于 x_i、$x_i-\bar{x}$、$x_i+\bar{x}$、$(x_i+\bar{x})(x_i-\bar{x})$及其他隐私信息.

(2)在整个计算中,F_2 可以得到 $E(x_i)$、$E(x_i/n)$、$E(-x_i/n)$、$E(x_i+\bar{x}+v')$、$E(x_i-\bar{x}+v')$、$E(v')$、$x_i+\bar{x}+v'$、$x_i-\bar{x}+v'$和 $k(x_i+\bar{x})(x_i-\bar{x})$,但没有 K_d,F_2 不可能知道这些加密过的数据. 对于 $x_i+\bar{x}+v'$和 $x_i-\bar{x}+v'$,由于 v'和 k 是由 F_3 产生,所以 F_2 不知道 v'和 k,也无法得到任何关于 x_i、$x_i-\bar{x}$、$k(x_i+\bar{x})(x_i-\bar{x})$及其他隐私信息.

(3)F_3 仅仅和 F_2 一起用两方加法隐私保护计算协议计算 $k((x_i+\bar{x}+v')+v)((x_i-\bar{x}+v')+v)$,由于该协议的隐私保护性,$F_3$ 不可能知道$(x_i+\bar{x}+v')$和$(x_i-\bar{x}+v')$,同时也不可能知道其他的隐私数据.

(4)其他的节点仅仅是将本地加密的计算结果传送给 F_2,而不与其他任何节点进行交互,故这些节点也无法得到任何隐私数据.

3.3 性能及动态性分析

本文算法基于PPCMLS模型,其计算复杂度为 $O(n^2)$,其通信复杂度应为 $O(n)$加上一次两方安全点积协议的通信复杂度(该通信复杂度取决于两方安全点积协议所采用的"1-out-of P 盲传输算法"中随机数 P 的大小[8]). 如在计算中,如果节点出现

机器故障或网络出现无法恢复的情况，F_2 将安排新的节点参与计算，这时仅需要将新的 $E(x_{n+1})$、$E(x_{n+1}/n)$和$E(-x_{n+1}/n)$发送给 F_2，而 F_2 仅需要重新与 F_3 一起计算一次$((x_{n+1}+\bar{x}+v')+v)\cdot((x_{n+1}-\bar{x}+v')+v)$即可. 当节点变化时，算法能以较快的速度恢复计算(仅涉及到新节点，以及 F_2 和 F_3).

4　分布式隐私保护数据聚类算法

4.1　算法步骤

(1)根据隐私安全保护计算模型的步骤(1)～步骤(4)，s_1(密钥生成子模块)生成一对加法同态加密密钥(K_e，K_d)，各节点得到 K_e，s_3(中间结果隐私保护子模块)将 $E(v')$发送给 s_2(全局计算子模块). 各节点计算从本地数据到 K 个子聚类中心的欧氏距离，并将 $E(d_{ij})$发送给 s_2.

(2)根据隐私安全计算模型的步骤(5)～步骤(6)，s_2 通过计算 $E(d_{1j})\cdots E(d_{mj})E(v'_j)(1\leqslant j\leqslant k)$得到$E((\sum_{i=1}^{m}d_{ij})+v'_j)$. s_2 将$E((\sum_{i=1}^{m}d_{ij})+v'_j)$发送给 s_3，s_3 通过随机扰乱算法对 $E((\sum_{i=1}^{m}d_{ij})+v'_j)$进行扰乱，也即是对 j 进行扰乱，得到 $P\left(E\left(\left(\sum_{i=1}^{m}d_{ij}\right)+v'_j\right)\right)$. 将 $P(E((\sum_{i=1}^{m}d_{ij})+v'_j))$发送给 s_1，s_1 对其解密，并将解密后的结果发送给 s_2.

(3)根据隐私安全计算模型的步骤(7)，s_2 和 s_3 一起通过两方加法隐私保护协议计算 $\left(P\left(\sum_{i=1}^{m}(d_{ij})+v_j'\right)+P(v_j)\right)/\left(P\left(\sum_{i=1}^{m}(d_{il})+v_l'\right)+P(v_l)\right)$，如果该值大于 1，则$\left(P\left(\sum_{i=1}^{m}(d_{ij})+v_j'\right)+P(v_j)\right)$大于$\left(P\left(\sum_{i=1}^{m}(d_{il})+v_l'\right)+P(v_l)\right)$，反之亦然. 通过$k-1$次比较，得到$P(\min(i))$. s_2 将$P(\min(i))$发送给 s_3，s_3 进行解扰，将 $\min(i)$ 发送给 s_2.

(4)s_2 将 $\min(i)$广播给所有的计算节点，并重复 N 次计算模型的步骤(1)～步骤(7)，从而让每个记录归并到最近的聚类中.

(5)每个节点分别在本地计算新的子聚类中心，并计算前后两次子聚类中心的差值 $D(u'_{ij}-u_{ij})$，再用 K_e 对 $D(u'_{ij}-u_{ij})$加密，并将 $E(D(u'_{ij}-u_{ij}))$发送给 s_2. 然后，算法第 2 次运用隐私安全保护计算模型，用类似的步骤(6)方法，将 $P\left(\sum_{i=1}^{m}D(u'_{ij}-u_{ij})+v'_j\right)$ 与预先设置好的门限值进行比较(该门限值仅 s_2 知道). 重复以上的步骤，直到所有的 $P\left(\sum_{i=1}^{m}D(u'_{ij}-u_{ij})+v'_j\right)$ 皆小于门限值.

4.2　安全性分析

(1)s_1 仅能得到 $P\left(\sum_{i=1}^{m}(d_{ij}+v'_j)\right)$ 和 $P\left(\sum_{i=1}^{m}D(u'_{ij}-u_{ij})+v'_j\right)$，由于受 v' 的干扰，s_1 无法从它们中得到任何隐私数据.

(2)s_2 仅得到 $E_k(d_{ij})$、$E(D(u'_{ij}-u_{ij}))$、$E(v'_j)$、$P((\sum_{i=1}^{m}d_{ij})+v'_j)$和$P\left(\sum_{i=1}^{m}D(u'_{ij}-u_{ij})+v'_j\right)$，如果没有 K_d 和 v'，s_2 也得不到任何敏感信息.

(3)s_3 仅得到 $E_k((\sum_{i=1}^{m}d_{ij})+v'_j)$、$E_k\left(\left(\sum_{i=1}^{m}D(u'_{ij}-u_{ij})\right)+v'_j\right)$，$s_3$ 和 s_2 一起通过两方隐私保护计算协议合作计算并消去 v'的干扰，若没有 K_d，s_3 无法知道任何隐私信息，且 s_3 仅知道最后的 $P(\min(i))$，故它也不可能知道 k 个聚类中心距离的中间比较结果.

(4)除了部署 s_1、s_2、s_3 的节点以外的其他节点，仅将本地加密的计算结果$E(d_{ij})$和 $E(D(u'_{ij}-u_{ij}))$传送给 s_2，而不与别的任何节点发生数据交换，故这些节点也无法得到任何隐私数据. 对于记录到 k 个聚类中心距离的中间比较结果，由于 k 个 $\sum_{i=1}^{m}(d_{ij})$和 $\sum_{i=1}^{m}D(u'_{ij}-u_{ij})$ 的序列被随机扰乱为 $P\left(\sum_{i=1}^{m}(d_{ij})\right)$ 和 $P\left(\sum_{i=1}^{m}D(u'_{ij}-u_{ij})\right)$，所以如果没有 s_3 解扰，s_2 及任何节点均无法知道具体的比较结果.

4.3　性能和动态性分析

若对于一条数据记录，寻找“最近聚类部分”的计算复杂度为 $O(km)$，那么 N 条记录的计算复杂度为 $O(Nkm)$，通信复杂度应为 $O(nm)$加上 N 次两方安全点积协议的通信复杂度. 对于“二次聚类的差值与门限值比较”部分，其计算时间复杂度为 $O(km)$，通信复杂度为 $O(m)$.

对于聚类算法，由于每条记录都需要找到最近距离的聚类中心，所以在整个隐私计算中，每次循环

时寻找“最近的聚类”部分需执行 n 次. 如果一些节点退出计算，s_2 将安排新的节点参与计算，这时仅需要将新的 $E(d_{(n+1)j})$ 发送给 s_2，而 s_2 仅需重新和 s_3 执行一次计算模型的步骤(1)～步骤(7)即可. 所以，当有节点离开时，算法通过寻找“最近的聚类”部分就能以较快的速度恢复计算. 既然聚类中心已经改变，需要重新计算聚类中心，所以应比较新的聚类中心的差值和门限值. 由于这 2 部分在一次循环中仅需执行一次，所以它们的重复不会对整个算法的动态性产生较大影响. 从上述分析可见，当节点变化时，算法能以较快的速度、较小的代价恢复计算. 与文献[9]、文献[10]的隐私保护分布式数据聚类算法比较，本文算法不需要一组分布到每个节点上的、总和为 0 的随机数，当节点变化时，恢复计算也无需涉及所有节点. 又因本文算法将加密分配到了每个本地节点，避免了文献[9]、文献[10]中的交互加密过程，所以运行效率提高了.

5 结束语及展望

本文针对当前分布式应用中存在的隐私保护问题，提出了一种适用于大规模分布式环境下的隐私保护计算模型，以及基于该模型的分布式隐私保护方差和分布式隐私保护数据聚类算法. 安全及动态性分析结果表明，本文模型及算法在对各节点数据及敏感中间结果进行隐私保护的同时，能成功地避免繁琐的一对多(多对多)的交互式加密和涉及全局的恢复计算.

对隐私保护计算的研究，本文仅停留在底层数据及中间计算结果上，这是远远不够的. 各节点通过对最后结果的猜测和比较，仍然可以得到私密数据的蛛丝马迹. 如何通过形式化的方法建立一个系统的隐私保护安全策略，将是下一步需要研究的内容. 在实际应用中，从管理学的角度合理地利用各节点之间的关系，建立一个安全的隐私保护规范，也是极为重要的.

参考文献：

[1] Vaidya J, Clifton C. Privacy-preserving data mining: why, how, and when [J]. Security and Privacy Magazine, 2004, 2(6): 19-27.

[2] Estivill-Castro V, Brankovic L. Data swapping: balancing privacy against precision in mining for logic rules [C]// Proceedings of the First Conference in Data Warehousing and Knowledge Discovery. Berlin: Springer-Verlag, 1999: 389-398.

[3] Kantarcioglu M, Clifon C. Privacy-preserving distributed mining of association rules on horizontally partitioned data [J]. IEEE Transaction on Knowledge and Data Engineering, 2004, 16(9): 1026-1037.

[4] Bresson E, Catalano D, Pointcheval D. A simple public key cryptosystem with a double trapdoor decryption mechanism and its applications [C] // Proceedings of 9th International Conference on the Theory and Application of Cryptology and Information Security. Berlin: Springer-Verlag, 2003: 37-54.

[5] Yao A C. Protocols for secure computation [C]// Proceedings of 23rd Annual IEEE Symposium on Foundations of Computer Science. Piscataway, USA: IEEE, 1982: 160-164.

[6] 孙中伟，冯登国，武传坤. 基于加同态公钥密码体制的匿名数字指纹方案[J]. 软件学报，2005，16(10)：1816-1821.
Sun Zhongwei, Feng Dengguo, Wu Chuankun. An anonymous fingerprinting scheme based on additively homomorphic public key cryptosystem [J]. Journal of Software, 2005, 16(10): 1816-1821.

[7] Phillip J C. Algorithm 383: permutations of a set with repetitions [J]. Communications of the ACM, 1970, 13(6): 368-369.

[8] Du W L, Atallah M J. Privacy-preserving cooperative statistical analysis [C] // Proceedings of 17th Annual Computer Security Applications Conference. Piscataway, USA: IEEE, 2001: 102-110.

[9] Estivill-Castro V. Private representative-based clustering for vertically partitioned data [C]// Proceedings of the 5th Mexican International Conference on Computer Science. Piscataway, USA: IEEE, 2004: 160-167.

[10] Vaidya J, Clifton C. Research track: privacy-preserving k-means clustering over vertically partitioned data [C]// Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 206-215.

(编辑 苗凌)