DOI:10.16515/j.cnki.32-1722/n.2018.04.003

一种高效的隐私集合交集协议

邱 硕[1,2],柳亚男[1,2],阎 浩[1,2],张 正[1,2]

(1.金陵科技学院软件工程学院,江苏 南京 211169;2.金陵科技学院网络安全学院,江苏 南京 211169)

摘 要:隐私集合交集计算已经被广泛地应用到各个行业中。随着数据量的不断增长,传统的隐私集合交集协议不再能有效地满足实际需求。主要从隐私保护的角度出发,设计满足实际需求的大数据隐私集合交集协议。结合布隆过滤器,构造了一个安全高效的隐私集合交集协议,该协议在半诚实敌手模型下安全通过实验测试。当集合大小为百万级时,协议在并行模式下的执行时间仅有15 s,同时可达到128位的安全级别。

关键词:大数据;隐私保护;集合交集;布隆过滤器

中图分类号:TP309 **文献标识码**:A **文章编号**:1672-755X(2018)04-0010-05

An Efficient Private Set Intersection Protocol

QIU Shuo, LIU Ya-nan, YAN Hao, ZHANG Zheng

(Jinling Institute of Technology, Nanjing 211169, China)

Abstract: Private Set Intersection (PSI) has extensive practical applications. With the increasing of data volume, traditional PSI protocols can no longer meet the actual needs. From the point of view of privacy protection, an efficient private set intersection protocol over large-scale datasets is designed. We construct a secure and efficient PSI protocol over large-scale datasets in the server-aided setting based on Bloom Filter algorithm. Our protocol is secure against a semi-honest server. Furthermore, experimental results show that our protocol only needs around 15 (128-bit security in parallel mode) over one million-element datasets.

Key words: large-scale dataset; privacy protection; set intersection; Bloom Filter

Freedman等[1]由2004年首先提出了隐私集合交集(Private Set Intersection, PSI)的概念。隐私集合交集广泛应用在各种新兴行业中,如基因匹配[2]、移动社交网络[3]以及电子医疗系统[4]等。由于传统的PSI协议已经无法满足大规模数据集合的计算[5-12]。Dong等[13]结合布隆过滤器和不经意传输技术构造了一个双方交互型的PSI协议,此协议可适应与百万级数据集大小的交集运算。随后,Pinkas等[14]利用扩展的不经意传输技术对Dong的方案进行了改进,提升了计算效率。但以上方案均是基于双方交互型的设计,由于需要多次的交互,无法适应于计算能力比较弱的用户端。Kamara等[15]基于伪随机置换提出了一个基于第三方辅助计算型的PSI协议,在并行计算模式下通过硬件加速后,可用于十亿数据级集合大小的交集计算。但此协议仍需要用户和第三方服务器之间多次交互。

本文设计了两个基于第三方辅助计算型的隐私集合交集协议,可实现高效安全的隐私集合交集计算。协议利用布隆过滤器存储集合元素,最后输出一个和精确值非常相近的交集结果。协议在半诚实敌手模

收稿日期:2018-11-18

基金项目:金陵科技学院高层次人才启动基金(jit-b-201726,jit-b-201639);江苏省高等学校自然科学研究面上项目(17KJD520003);网络安全专项项目(2017YFB0802800)

作者简介:邱硕(1989—),女,安徽阜阳人,讲师,博士,主要从事大数据安全、云数据安全以及应用密码学研究。

型下设计的，Server 可以估测输入集合的大小和计算出来的集合交集的大小。实验结果表明，在并行模式下，对于百万级的数据集合大小，协议的平均运行时间为 15 s，并且可达到 128-bit 的安全性。

1　相关工作

Freedman 等[1]首次提出隐私数据集合交集的概念(Private Set Intersection，PSI)，并给出了相应的解决方案。PSI 方案大体上可以分为两种类型：第一种是双方交互型，通过双方之间的交互计算出隐私集合交集；而第二种是借助第三方辅助服务器的作用，数据拥有者把数据外包存储在服务器端或者将部分计算外包给服务器，从而实现隐私集合交集的计算。

1)双方交互计算型。双方计算型的隐私集合交集(Two-party PSI)是指用户 Alice 和 Bob 之间直接通过交互计算出交集结果。Freedman 等[1]基于不经意多项式根植的思想，并分别给出了在半诚实敌手和恶意敌手模型下的安全设计。随后学者在此基础上进行了效率和安全性的提升[16－18]。相关学者还提出了基于不经意伪随机函数构造的隐私集合交集协议[19－22]，以及基于盲签名[5]构造的隐私集合交集协议。随着数据量的快速增长，Dong 等[13]实现了支持大数据集的集合交集计算。之后，相关学者为了实现更为高效的双方交互型的集合交集计算协议，采用扩展的 OT 协议[14]等技术大大提高计算效率。

2)第三方辅助计算型。基于第三方辅助计算型的隐私集合交集的研究，已经取得了许多成果。Kerschbaum 等[23]利用 Goldwasser-Micali 同态加密[24]给出了外包计算的隐私集合交集协议，将一部分密文交集计算授权给第三方服务器，大大地节约了用户端的计算开支。后续学者也有利用可检索加密[25－27]和密文相等性测试[28－29]的思想实现密文集合的匹配计算，但复杂的密文计算使得上述方案无法满足大规模数据的集合交集计算。

Kamara 等[15]使用确定性的 AES 加密方案加密数据，分别给出抵抗恶意第三方和隐藏交集大小的隐私集合交集协议。文献[15]的方案大大提高了协议运行效率使其能够满足 TB 数据级集合大小的计算需求。但不足的是，该方案在用户端 Alice 和 Bob 与服务器端 Server 之间进行交互时，其交互过程中的通信开支与集合大小呈线性增长关系。

本文基于布隆过滤器构造了一个安全高效的隐私集合交集协议，同时在不泄露隐私信息的条件下将隐私集合操作外包给第三方服务器来完成。实验结果表明我们的协议能够满足大数据集合的实际计算需求。

2　模型定义

本文 PSI 协议中包括三个实体：两个客户端参与方 Alice 和 Bob，一个第三方辅助服务器 Server。主要执行如下，两个参与方 Alice 和 Bob 分别有一个隐私集合 S_A 和 S_B，然后分别将两个集合进行加密上传至服务器 Server，Server 通过计算将集合 S_A 和 S_B 的交集结果返回给参与方 Alice 和 Bob。

一般情况下我们考虑半诚实敌手模型：简单的说，一个半诚实敌手是指能够正确地按照协议步骤来执行，但是会记录协议过程中计算的所有中间结果，并根据这些中间结果尽可能地推测出额外的隐私信息。同时，我们假设任意两个参与方之间没有共谋，也就是说服务器 Server 是不允许与任一参与方 Alice(Bob)进行共谋，且参与方之间也不存在任何共谋。

3　预备知识

3.1　布隆过滤器(Bloom Filter，BF)

布隆过滤器可以用于快速检索一个元素是否在一个集合中。一个空的布隆过滤器是一个 m 位的二进制向量，由 k 个相互独立的哈希函数组成，表示为 $H=\{h_0,h_1,\cdots,h_{k-1}\}$。设集合 $S=\{s_1,s_2,\cdots,s_n\}$，$BF_S=\{BF_S[0],BF_S[1],\cdots,BF_S[m-1]\}$表示集合 S 对应的布隆过滤器，$BF_S[j](0\leqslant j<m)$表示 BF_S 的第 j 位。在一个布隆过滤器 BF_S 中添加和查询元素的整个过程描述见图 1。

• 初始化：首先初始化

$$BF_S[0]=0,\cdots,BF_S[m-1]=0,$$

然后选 k 个相互独立的哈希函数

$$h_0,h_1,\cdots,h_{k-1}:\{0,1\}^*\rightarrow\{0,1,\cdots,m-1\}$$

• 构造：将元素 $s\in S$ 添加到 BF_S 中，计算 $h_i(s)(0\leqslant i\leqslant k-1)$，然后设置 $BF_S[h_0(s)]=1,\cdots,BF_S[h_{k-1}(s)]=1$，若 $BF_S[j]$ 已被设置为 1，将不再改变其值。重复上述步骤，直至将 S 中的所有元素添加到 BF_S 中。

• 查询：查询元素 s' 是否在 S 中，如果

$$BF_S[h_0(s')]=1\wedge\cdots\wedge BF_S[h_{k-1}(s')]=1,$$

则 s' 可能在 S 中。否则 s' 不在 S 中。

图 1 布隆过滤器原理

3.2 确定性加密(Deterministic Encryption，DE)

确定性加密是指加密算法是确定性的，算法包含三部分：密钥产生 KeyGen，加密 Enc，解密 Dec。其中，KeyGen 是一个概率性算法，Enc 和 Dec 是确定性算法。具体描述如下：

• $sk\leftarrow\mathrm{KeyGen}(1^\lambda)$：给定安全参数 λ，输出对称密钥 $sk\in\boldsymbol{K}$，$\boldsymbol{K}$ 是密钥空间。

• $c\leftarrow\mathrm{Enc}(sk,m)$：该算法属于一个确定性算法，给定 sk 和消息 $m\in\boldsymbol{M}$，输出密文 c。

• $m\leftarrow\mathrm{Dec}(sk,c)$：该算法是一个确定性算法。

4 协议设计

设计了一个安全高效的隐私集合交集协议。协议在半诚实敌手模型下设计的，Server 可以估测输入集合的大小和计算出来的集合交集的大小，而有时候集合的大小也属于隐私信息。

为了保证交集结果计算的正确性，我们根据布隆过滤器中值的不同对元素进行以下不同方式的加密，

$$C[i]=\begin{cases}\mathrm{Enc}(sk,i||BF[i]),BF[i]=1,\\ \mathrm{Enc}(sk,r_i||BF[i]),BF[i]=0\end{cases}\tag{1}$$

其中，$r_i(i\in\{0,1,\cdots,m-1\})$ 是一个随机数，|| 表示链接两个数值，$C=\{C[1],C[2],\cdots,C[m-1]\}$ 是布隆过滤器中的元素被加密后的密文。

该协议在半诚实辅助服务器基础上执行 Alice 和 Bob 的隐私集合交集。主要思想是：每个客户端在本地构造一个自己集合对应的布隆过滤器，然后按照等式(1)加密布隆过滤器的每一个比特，最后将密文外包给半诚实服务器进行集合交集计算。当收到两个客户端的密文后，服务器执行相等性测试并且返回交集布隆过滤器给双方。每个客户端 Alice(Bob)通过查询自己的集合布隆过滤来恢复交集。

在整个过程中，为了保护客户端的输入隐私，Alice 和 Bob 预先共享一个确定性加密的密钥。此外，双方需要约定一个秘密的伪随机置换函数，以此防止服务器获得布隆过滤器中相等元素对应的真实位置。协议的详见图 2。

协议的形式化安全定义如下定理。

定理 1 根据确定性加密 DE 和置换函数 π 的伪随机性，图 2 中所述的隐私集合交集不会泄露任何额外的隐私给半诚实服务器。

证明：首先，构造理想世界的模拟器 S 去模拟真实世界中服务器的行为与 Alice 和 Bob 之间交互。如果 A 在真实世界中的联合分布视图和 S 在理想世界中的视图不可区分，那么协议是安全的。

模拟器 S 在理想世界中的信息只有加密的布隆过滤器 $C_A\leftarrow\mathrm{Enc}(sk,BF_A),C_B\leftarrow\mathrm{Enc}(sk,BF_B)$。然后，$S$ 计算 $C_A\cap C_B$ 且返回计算给敌手 A。根据 DE 加密的伪随机性，A 不能区分真实世界和理想世界中的密文。敌手 A 和模拟器 S 视图的唯一区别是返回的交集结果长度。但是长度是在 $[0,m]$ 中随机分布的(m 是布隆过滤器的长度)，以此保证了敌手 A 无法区分真实世界和理想世界。此外，根据置换函数 π，敌手 A 不能猜测到布隆过滤器中真实比特的位置。因此，无法根据获得的 $C_A\cap C_B$ 通过暴力攻击获得交集中的任何明文元素。

系统初始化：设确定性加密 DE=(KeyGen，Enc，Dec)，Alice 和 Bob 的输入集合分别为 S_A 和 S_B，π 是一个随机置换函数。协议 I 按照如下方式进行：

• Alice(Bob)分别产生一个 m-bit 的布隆过滤器

$$BF_A \leftarrow \mathrm{Creat}BF(S_A),$$
$$BF_B \leftarrow \mathrm{Creat}BF(S_B)$$

• Alice(Bob)分别计算

$$C_A \leftarrow Enc(sk, BF_A),$$
$$C_B \leftarrow Enc(sk, BF_B)$$

根据公式(1)的规则加密。然后 Alice(Bob)计算 $T_A \leftarrow \pi(C_A)$ 和 $T_B \leftarrow \pi(C_B)$，对密文进行随机置换，并将 T_A 和 T_B 发给 Server。

• Server 测试是否

$$T_A[i] = T_B[i], i \in \{0,1,\cdots,m-1\}$$

返回结果 $I=\{i \mid T_A[i]=T_B[i]\}$ 给 Alice 和 Bob。

• Alice 和 Bob 计算相对应交集的布隆过滤器，即

$$BF_{S_A \cap S_B} = \{BF_{S_A \cap S_B}[i] \mid i \in \pi^{-1}(I)\},$$

通过查询 $\mathrm{Query}(S_j, BF_{S_A \cap S_B}), j \in \{A,B\}$ 得到交集结果。

图 2　协议的详细过程

5　性能评估

实验中利用 Crypto＋＋库中的密码操作，实验中利用 Crypto＋＋库中的密码操作，利用 128-bit 安全的 AES-ECB 加密模式来实现确定性加密，并利用 MD5 算法来构造布隆过滤器。实验中分别给出了串行和并行模式下的运行结果。表 1 给出了集合大小 n 从 $10^4 \sim 10^8$ 的平均运行时间。结果显示，并行模式可大大提升协议的执行效率，例如，当集合大小为 10^6 时，协议的平均执行时间分别为 79 s，而在并行模式下，平均执行时间可被减少到 15 s，显著提升了执行效率。

表 1　协议平均运行时间

运行模式	不同集合大小时的运行时间/s				
	10^4	10^5	10^6	10^7	10^8
串行	0.77	8.10	79	804	7 985
并行	0.18	1.58	15	138	1 330

6　结　语

本文基于布隆过滤器技术设计了一个安全高效的隐私集合交集协议。协议可在保护数据隐私的条件下适应百万数据集合大小之间的交集计算。实验结果表明，并行计算模式下，本文中的协议可高效地完成大规模集合大小的交集。

参考文献：

[1] Freedman M J，Nissim K，Pinkas B. Efficient private matching and set intersection[J]. Proceedings of Advances in EUROCRYPT，2004(4)：1－19

[2] Baldi P，Baronio R，Cristofaro E D，et al. Countering gattaca：efficient and secure testing of fully-sequenced human genomes[C]//Proceedings of the 18th ACM Conference of Computer and Communications Security. New York：ACM，2011：691－702

[3] Li M，Yu S，Cao N，et al. Privacy-preserving distributed profile matching in proximity-based mobile social networks[J]. IEEE Transactions on Wireless Communications，2013，12(5)：2024－2033

[4] Tang Q. Public key encryption supporting plaintext equality test and user-specified authorization[J]. Security and Communication Networks，2012，5(12)：1351－1362

[5] Cristofaro E D，Tsudik G. Practical private set intersection protocols with linear complexity[C]//Financial Cryptography

and Data Security. Tenerife:Canary Islands,2010:143－159
[6] Huang Y,Evans D,Katz J. Private set intersection:Are garbled circuits better than custom protocols[C]. San Diego,19th Network and Distributed Security Symposium,2012
[7] Asharov G,Jain A,LópezAlt,et al. Multiparty computation with low communication,computation and interaction via threshold FHE[M]. Berlin Advances in Cryptology-EUROCRYPT,2012:483－501
[8] Dong C,Chen L,Camenisch J,et al. Fair private set intersection with a semi-trusted arbiter[M]. Heidelberg:Data and Applications Security and Privacy XXVII,2013:128－144
[9] Qiu S,Liu J,Shi Y. Identity-based symmetric private set intersection[C]//IEEE. Social Computing (SocialCom),2013 International Conference on. Alexandria:Virginia,2013:653－658
[10] Shikfa A,Onen M,Molva R. Broker-based private matching[C]//Privacy Enhancing Technologies. Waterloo:ON,2011:264－284
[11] Shao Z Y,Yang B. Private set intersection via public key encryption with keywords search[J]. Security and Communication Networks,2015,8(3):396－402
[12] Canetti R,Paneth O,Papadopoulos D,et al. Verifiable set operations over outsourced databases[J]. ECM,2014(3):85－92
[13] Dong C,Chen L,Wen Z. When private set intersection meets big data:an efficient and scalable protocol[C]. Berlin:ACM Sigsac Conference on Computer & Communications Security,2013:789－800
[14] Pinkas B,Schneider T,Zohner M. Faster private set intersection based on OT extension[C]. NewYork:USENIX Association,Usenix Conference on Security Symposium,2014:797－812
[15] Kamara S,Mohassel P,Raykova M,et al. Scaling private set intersection to billion-element sets[J]. ECM,2014(5):195－215
[16] Kissner L,Song D. Privacy-preserving set operations[M]. San Diego:Advances in Cryptology-CRYPTO,2005:241－257
[17] Dachmansoled D,Malkin T,Raykova M,et al. Efficient robust private set intersection[J]. Lecture Notes in Computer Science,2009,2(4):289－303
[18] Hazay C,Nissim K. Efficient set operations in the presence of malicious adversaries[J]. Journal of Cryptology,2012,25(3):383－433
[19] Hazay C,Lindell Y. Efficient protocols for set intersection and pattern matching with security against malicious and covert adversaries[J]. Journal of Cryptology,2010,23(3):422－456
[20] Stanisław J,Liu X. Efficient oblivious pseudorandom function with applications to adaptive OT and secure computation of set intersection[C]. Waterloo:Theory of Cryptography Conference on Theory of Cryptography,2009:577－594
[21] Stanisław J,Liu X. Fast secure computation of set intersection[C]. Verlag:International Conference on Security & Cryptography for Networks,2010:418－435
[22] Cristofaro E D,Jarecki S,Kim J,et al. Privacy-preserving policy-based information transfer[C]. Springer-Verlag:International Symposium on Privacy Enhancing Technologies,2009:164－184
[23] Kerschbaum F. Outsourced private set intersection using homomorphic encryption[C]//ACM. Proceedings of Proceedings of the 7th ACM Symposium on Information,Computer and Communications Security. NewYork,2012:85－86
[24] Goldwasser S,Micali S. Probabilistic encryption[J]. Journal of Computer & System Sciences,1984,28(2):270－299
[25] Cash D,Jarecki S,Jutla C,et al. Highly-scalable searchable symmetric encryption with support for boolean queries[C]//Springer. Annual Cryptology Conference. Berlin:Heidelberg,2013:353－373
[26] Curtmola R,Garay J,Kamara S,et al. Searchable symmetric encryption:improved definitions and efficient constructions[C]//Proceedings of the 13th ACM conference on Computer and communications security. NewYork:ACM,2006:79－88
[27] Kamara S,Papamanthou C,Roeder T. Dynamic searchable symmetric encryption[C]//Proceedings of the 2012 ACM conference on Computer and communications security. New York:ACM,2012:965－976
[28] Lipmaa H. Verifiable homomorphic oblivious transfer and private equality test[J]. Lecture Notes in Computer Science,2003,28:416－433
[29] Yang G,Tan C H,Huang Q,et al. Probabilistic public key encryption with equality test[C]. Verlag:International Conference on Topics in Cryptology,2010:119－131

(责任编辑:湛 江)