doi:10.11959/j.issn.1000-436x.2015324

基于布隆过滤器的轻量级隐私信息匹配方案

万盛[1]，何媛媛[2]，李凤华[2,3]，牛犇[2]，李晖[1]，王新宇[2]

(1. 西安电子科技大学 综合业务网理论及关键技术国家重点实验室，陕西 西安 710071；
2. 中国科学院 信息工程研究所 信息安全国家重点实验室，北京 100195；3.北京电子科技学院 信息安全系，北京 100070)

摘 要：针对智能终端用户私有数据匹配中的隐私保护问题，基于布隆过滤器和二元向量内积协议，提出一种新的综合考虑用户属性及其偏好的轻量级隐私信息匹配方案，包括建立基于 Dice 相似性系数的二维向量相似度函数、设置参数、生成布隆过滤器、计算二元向量内积、计算相似度和确定匹配对象 6 个部分。该方案采用基于布隆过滤器的相似度估计和基于混淆方法的二元向量内积协议，在不依赖于可信第三方的前提下，大幅度降低计算开销，且能够有效抵御蛮力攻击和无限制输入攻击。实验结果表明，该方案与典型代表方案相比，计算效率得到明显提升。

关键词：隐私信息匹配；Dice 相似性系数；布隆过滤器；二元向量内积协议

中图分类号：TN929 **文献标识码**：A

Bloom filter-based lightweight private matching scheme

WAN Sheng[1], HE Yuan-yuan[2], LI Feng-hua[2,3], NIU Ben[2], LI Hui[1], WANG Xin-yu[2]

(1. State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China;
2. State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China;
3. Department of Information Security, Beijing Electronic Science and Technology Institute, Beijing 100070, China)

Abstract: With rapid developments of mobile devices and online social networks, users of proximity-based mobile social networks (PMSN) could easily discover and make new social interactions with others, but they enjoyed this kind of conveniences at the cost of user privacy and system overhead, etc. To address this problem, a third party free and lightweight scheme to privately match the similarity with potential friends in vicinity was proposed. Unlike most existing work, proposed scheme considered both the number of common attributes and the corresponding priorities on each of them individually. The Bloom filter-based common-attributes estimation and the lightweight confusion binary vector scalar product protocol reduce the system overhead significantly, and can resist against brute force attack and unlimited input attack. The correctness, security and performance of overhead of proposed scheme are then thoroughly analyzed and evaluated via detailed simulations.

Key words: privacy matching; Dice similarity coefficient; Bloom filter; confusion binary vector scalar product protocol

1 引言

随着大量智能移动终端的普及和社交网络等技术的出现和快速发展，以医疗、交通和求职等为主题的邻近用户移动社交网络(PMSN，proximity-based mobile social networks)已经逐渐渗入人们的日常生活。PMSN 的用户采用移动终端的蓝牙、Wi-Fi 等无线通信方式与活跃在邻近区域的其他用户即时交流，找到具有相同病症的患者、交通线路相同的出行者和相同行业的求职者等用户，与他们

收稿日期：2015-08-20；**修回日期**：2015-12-12
基金项目：国家自然科学基金-广东联合基金资助项目(U1401251)；国家高技术研究发展计划(“863”计划) 基金资助项目(2012AA013102)；教育部重点基金资助项目(209156)
Foundation Items: The National Natural Science Foundation of China –Guangdong Union Foundation(U1401251); The National High Technology Research and Development Program of China (863 Program)(2012AA013102); The Key Program of Scientific and Technology Research of Ministry of Education(209156)

相互分享治疗经验、路况和求职经历等信息，获取更多的信息资源和机会。然而，用户通过分享个人信息获取信息资源和机会时，其个人隐私例如病症病史、出行路线、个人简历等敏感信息将暴露给其他用户，从而对用户的人身、财产安全造成潜在危险。因此，对移动社交网络用户个人隐私的保护刻不容缓。

针对 PMSN 中的隐私信息匹配问题，近年来大量文献提出了解决方法。大多数方法假设每位用户拥有一个属性集合，例如多种疾病症状[1]、出行路线[2]、求职岗位等，采用匹配参与双方共同属性的个数为相似度，共同属性越多，用户越匹配，整个匹配过程不泄露双方个人敏感信息，但是，这些方案存在一定的局限性：依赖第三方对用户的输入信息进行认证；匹配函数只考虑了共同属性的个数，忽视了每个用户对各个属性偏好程度的差异[3,4]；基于交换加密、同态加密等加密方法的隐私信息匹配方案安全性较高。但是，计算开销较大，难以在移动终端上运行。

为了解决上述问题，本文利用布隆过滤器[5](Bloom filter)和二元向量内积协议[6]，提出一种轻量级、与属性偏好相关的隐私信息匹配方案，其主要贡献如下。

1) 应用布隆过滤器编码私有数据集合的所有元素，从而将私有数据集合的匹配问题转化为布隆过滤器的内积计算问题，而且无需第三方。

2) 基于 Dice 相似性系数建立一种新的二维向量相似度函数，既考虑属性匹配和偏好程度的匹配，又有助于抵御无限制输入攻击。

3) 采用混淆的方法计算二元向量内积，并结合基于布隆过滤器的相似度估计公式，构建轻量级隐私信息匹配方案，整个方案不涉及耗时的加密运算，计算开销较小。

4) 论证方案的正确性，并通过分析方案的隐私保护力度和抗攻击能力论证其安全性。

2 相关工作

近年来对于 PMSN 中私有数据匹配的隐私保护问题涌现了大量的解决方案，大致可以划分为 2 类：基于加密的方法和基于非加密的方法。

大量文献[7~14]将私有数据集合匹配中的隐私保护问题视为 PSI(private set intersection)问题，采用 OPE[7~9](oblivious polynomial evaluation)、OPRF[10~13](oblivious pseudo-random functions)或同态加密算法[14]实现 PSI。然而，这些方案均涉及大量同态加密运算，计算开销较大，或者依赖第三方对输入信息进行认证，导致用户需要承担数字签名等加密算法带来的额外计算开销。此外，基于交换加密的隐私信息匹配方案被频繁报道。2003 年，Agrawal[15]等提出了基于交换加密的双方秘密共享方案，该方案无需第三方。同一时期，Vaidya 等[16]提出了 N 方秘密共享方案。随后 Veneta 等[17]建立了移动社交网络中的 friend-of-friend 预测模型。然而，该模型无法抵御蛮力攻击。Zhang 等[18]首次提出了细粒度隐私信息匹配的概念，对每个属性赋予偏好，使用相似度函数来衡量匹配程度。在此基础上，Niu 等[19~21]设计隐私信息匹配协议时进一步考虑了属性本身的权重或等级，以及匹配参与方的社会关联(social strength)等因素。由于采用加密算法，以上方案大多安全性较强，但是，计算量较大，难以为移动用户提供高效的隐私信息匹配服务。

相对而言，采用非加密的方法实现隐私信息匹配的文献少见报道。Fu 等[22]提出了基于布隆过滤器的隐私信息匹配方案，Many 等[23]应用布隆过滤器建立了多方安全共享模型，大幅降低了计算开销。但是，这些方案明文传输布隆过滤器的交集，将无法抵御蛮力攻击，导致隐私信息泄露。Zhang 等[2]针对时空匹配场景，提出了基于布隆过滤器的轻量级隐私信息匹配方案，未使用加密算法，只涉及到散列 SHA-256 运算，计算开销较小，但是，布隆过滤器及其部分散列函数被公开，将无法抵御蛮力攻击。以上非加密的隐私信息匹配方案普遍安全性较弱，且未考虑属性偏好、权重和用户间社会关联等因素。

综上所述，现有的隐私信息匹配方案大多忽视了以下 2 点：现实生活中，用户通常对不同属性的偏好程度不同，衡量匹配程度的相似度函数除了与共同属性的个数相关，也与属性的偏好程度相关；较多隐私信息匹配方案采用同态加密、交换加密等加密算法，计算开销颇大。因此，本文旨在构建一个轻量级、与属性偏好相关的隐私信息匹配方案。

3 研究问题、思路和相关技术

3.1 研究问题与思路

假设 PMSN 中的每一位用户具有一组二维向

量$\{\langle I_1,w_1\rangle,\langle I_2,w_2\rangle,\cdots,\langle I_n,w_n\rangle\}$，其中，$I_i$表示用户设定的第 i 个属性，$w_i\in[1,l]$表示与属性I_i对应的偏好程度，w_i越大表示用户越偏好属性I_i，l 通常取较小的整数，比如 7 或 10，足以区分用户对某一属性的偏好程度即可。设匹配发起者为 Alice，响应 Alice 匹配请求的响应者记为 V_i (i=1,2,⋯)，在隐私保护的前提条件下，如果响应者与 Alice 的共同属性越多，并且相应的偏好程度越相近，那么他与 Alice 的匹配程度越高。本方案拟为 Alice 快速找到与之最匹配的用户并有效保护双方隐私信息，即求解$V_{\text{match}}=\arg\left(\max\left\{p\left(\text{Alice},V_i\right)\right\}_{i=1}^{\infty}\right)$，其中，$p$ 表示相似度函数。

求解V_{match}的基本思路如下：首先基于二元向量 Dice 相似性系数建立一种新的与属性及其偏好都相关的二维向量相似度函数，然后利用布隆过滤器和非同态加密的二元向量内积协议计算二维向量相似度 p，最后选出相似度最大的用户V_{match}为匹配对象。

3.2　布隆过滤器

布隆过滤器[5]是一种空间效率较高的随机数据结构，将集合$S=\{x_1,x_2,\cdots,x_n\}$，编码在 w 位数组中，并能判断某个元素 是否属于集合 S。构建集合 S 的 (w, m, k, H)-布隆过滤器 BF_S的步骤如下。首先选定散列函数集合$H=\{h_0,h_1,\cdots,h_{k-1}\}$，其中散列函数$h_0,h_1,\cdots,h_{k-1}$，相互独立，并且值域均为[0,$w$−1]，然后将 BF_S的所有位初始值置为 0，最后对所有 $x\in S$ 和$0\leqslant i\leqslant k-1$令$BF\left[h_i(x)\right]=1$，即可得 BF_S。但是，布隆过滤器在判断元素是否属于集合 S 时，有可能会出现误判(false positive)，即把$x\notin S$误认为 $x\in S$。当需要判断元素 y 是否属于集合 S 时，只需计算 $h_i(y)$ ($0\leqslant i\leqslant k-1$)并检查 $BF\left[h_i(y)\right]$是否全为 1，若不全为 1，$y\notin S$，否则 $y\in S$，其最大错误率[24]是

$$\varepsilon=p^k\left(1+O\left(\frac{k}{p}\sqrt{\frac{\ln w-k\ln p}{w}}\right)\right) \tag{1}$$

其中，$p=1-\left(1-\frac{1}{w}\right)^{km}$，$\varepsilon$ 为 k 的可忽略函数。允许最大错误率为ε时，w 的取值范围是$w\geqslant -m\text{lbelb}\varepsilon$，其中 e 指自然数。当$k=w\ln\frac{2}{m}$或$k=-\text{lb}\varepsilon$时，错误率最小。如无特别说明，本文默认采用以上公式计算 w 和 k。

3.3　攻击模型和隐私保护目标

设匹配发起者为 Alice，Bob 是响应 Alice 匹配请求的多个用户之一，以 Alice 和 Bob 的信息匹配过程为例描述攻击模型和隐私保护目标。

攻击模型：将诸如公钥密码基础体系(PKI, public key infrastructure)等常见的密码学技术应用到本方案通信通道之上，对各个实体间所传输的内容进行加密操作，便可以很容易地保证通信信道的安全，避免窃听等被动攻击的发生。因此，本方案忽略上述通信信道安全等问题，着重考虑来自匹配参与者的内部攻击。匹配参与者任一方可以是 HBC(honest but curious)用户甚至是恶意(malicious)攻击者。HBC[25]用户指该用户能诚实地执行既定方案，但是，用户会分析方案执行过程中获得的所有信息，试图获得对方更多的隐私信息。恶意攻击[25]指攻击者通过有策略地输入某些属性及其偏好发起攻击。

隐私保护目标：基于以上攻击模型，当匹配方案结束时，发起者 Alice 除了她与响应者 Bob 的相似度 p 外，不知道关于 Bob 属性集的任何信息，Bob 也不知道关于 Alice 属性集的任何信息。

4　基于布隆过滤器的隐私信息匹配框架

4.1　系统架构

本文为二维向量的私有数据匹配问题提供解决方案，处于移动终端无线通信(如蓝牙、Wi-Fi 等)范围内的任意 2 个用户进行一次信息交互即可完成隐私信息匹配，无需第三方介入。本匹配方案首先将测度二元向量相似度的 Dice 相似性系数扩展为二维向量集合相似度函数，然后基于此二维向量集合相似度函数建立隐私信息匹配模型。如图 1 所示，假设 Alice 和 Bob 分别是信息匹配发起者和响应者，本匹配模型主要包括设置参数、生成布隆过滤器、计算二元向量内积、计算相似度和确定匹配对象 5 个部分，其中设置参数为所有用户建立属性向量集合，并设置本匹配模型涉及到的所有参数；生成布隆过滤器部分，用户输入自己的属性向量集合，离线生成相应的生成布隆过滤器并输出；计算二元向量内积由混淆二元向量、计算二元向量内积混淆值和还原二元向量内积 3 个部分组成，输入匹配发起者 Alice 和匹配响应者 Bob 的布隆过滤器，将布隆过滤器视为二元向量，输出 2 个二元向量的内积；

计算相似度部分，发起者根据上一个部分得到的二元向量内积，采用基于布隆过滤器的相似度估计公式计算 Alice 和 Bob 的属性向量集合相似度；确定匹配对象部分，匹配发起者收集与所有响应者的相似度，选择拥有最大相似度的响应者作为匹配对象。

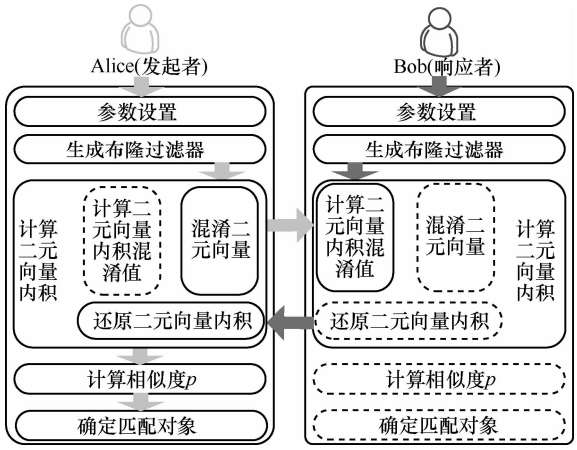

图 1 系统架构

4.2 基于 Dice 相似性系数的二维向量集合相似度函数

4.2.1 构建二维向量集合相似度函数

本节基于 Dice 相似性系数，建立多元二维向量集合 $A=\left\{\langle x_1,a_1\rangle,\langle x_2,a_2\rangle,\cdots,\langle x_{n_A},a_{n_A}\rangle\right\}$ 和 $B=\left\{\langle y_1,b_1\rangle,\langle y_2,b_2\rangle,\cdots,\langle y_{n_B},b_{n_B}\rangle\right\}$ 的相似度函数。由于 Dice 相似性系数[26]只适用于二元多维向量，所以建立 $p(A,B)$ 的关键在于把集合 A 和 B 中的所有多元二维向量转化为二元多维向量 $\overline{\boldsymbol{A}}$ 和 $\overline{\boldsymbol{B}}$，其具体过程如下。

设有二元向量 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_r)$，$\beta=(\beta_1,\beta_2,\cdots,\beta_r)$，$\alpha_i,\beta_i\in\{0,1\}$，其 Dice 相似性系数为

$$Dice(\alpha,\beta)=\frac{2\sum_{i=1}^{r}\alpha_i\beta_i}{\sum_{i=1}^{r}\alpha_i^2+\sum_{i=1}^{r}\beta_i^2} \tag{1}$$

令 $X=\left\{x_1,x_2,\cdots,x_{n_A}\right\}$，$Y=\left\{y_1,y_2,\cdots,y_{n_B}\right\}$，$C=X\bigcup Y=\left\{c_1,c_2,\cdots,c_t\right\}$ 且

$$c_q=\begin{cases}x_{i_q}=y_{j_q},q\in[1,s]\\ x_{i_q},q\in[s+1,n_A]\\ y_{j_q},q\in[n_A+1,t]\end{cases}$$

其中，$s=\left|X\bigcap Y\right|,t=\left|X\bigcup Y\right|$，那么 A 和 B 可转化为二元多维向量 $\overline{\boldsymbol{A}}=(cx_1,cx_2,\cdots,cx_t)$ 和 $\overline{\boldsymbol{B}}=(cy_1,cy_2,\cdots,cy_t)$，其中，

$$cx_q=\left(cx_{q,1},cx_{q,2},\cdots,cx_{q,l}\right)=\begin{cases}\left(\overbrace{1,1,\cdots,1}^{a_{i_q}},\overbrace{0,0,\cdots,0}^{l-a_{i_q}}\right),q\in[1,n_A]\\ \left(\overbrace{0,0,\cdots,0}^{l}\right),q\in[n_A+1,t]\end{cases}$$

$$cy_q=\left(cy_{q,1},cy_{q,2},\cdots,cy_{q,l}\right)=\begin{cases}\left(\overbrace{1,1,\cdots,1}^{b_{i_q}},\overbrace{0,0,\cdots,0}^{l-b_{i_q}}\right),q\in[1,s]\bigcup[n_A+1,t]\\ \left(\overbrace{0,0,\cdots,0}^{l}\right),q\in[s+1,n_A]\end{cases}$$

所以

$$\begin{aligned}p(A,B)&=Dice\left(\overline{\boldsymbol{A}},\overline{\boldsymbol{B}}\right)\\&=\frac{2\sum_{q=1}^{t}\sum_{i=1}^{l}cx_{q,i}cy_{qi}}{\sum_{q=1}^{t}\sum_{i=1}^{l}cx_{q,i}^{\ 2}+\sum_{q=1}^{t}\sum_{i=1}^{l}cy_{qi}^{\ 2}}\\&=\frac{2\sum_{q=1}^{s}\min\left(a_{i_q},b_{i_q}\right)}{\sum_{q=1}^{n_A}a_{i_q}+\sum_{q=1}^{s}b_{j_q}+\sum_{q=n_A+1}^{t}b_{j_q}}\\&=\frac{2\sum_{x_\mu=y_\vartheta\in X\cap Y}\min\left(a_\mu,b_\vartheta\right)}{\sum_{i=1}^{n_A}a_i+\sum_{j=1}^{n_B}b_j}\end{aligned} \tag{2}$$

显然，$p(A,B)\in[0,1]$，$p(A,B)$越大，A 与 B 越相似。$p(A,B)=0$ 表示没有共同属性，$p(A,B)=1$ 表示所有属性和偏好均相同。

4.2.2 建立二维向量集合相似度函数的计算公式

本节基于如式(2)所示多元二维向量集合的相似度函数 $p(A,B)$，推导其计算公式。计算 $p(A,B)$的关键在于计算分子即 $\sum_{x_\mu=y_\vartheta\in X\cap Y}\min\left(a_\mu,b_\vartheta\right)$，拟采用布隆过滤器计算此分子，不妨设偏好指标函数 $f_i(j)=j$（$j\in[1,a_i]$，$i\in[1,n_A]$）和 $g_i(j)=j$（$j\in[1,b_j]$，$i\in[1,n_B]$），将 A 和 B 分别转化为

$$A'=\left\{\left\{\left\langle x_i,f_i(j)\right\rangle\right\}_{j=1}^{a_i}\right\}_{i=1}^{n_A}$$

和

$$B'=\left\{\left\{\left\langle y_j,g_i(j)\right\rangle\right\}_{j=1}^{b_j}\right\}_{j=1}^{n_B}$$

那么 $\left|A'\right|=\sum_{i=1}^{n_A}a_i$，$\left|B'\right|=\sum_{j=1}^{n_B}b_j$，且

$$A'\bigcap B'=\left\{\left\langle x_\mu,f_\mu(j)\right\rangle\middle|x_\mu=y_\vartheta\in X\bigcap Y,j\in\left[1,\min\left(a_\mu,b_\vartheta\right)\right]\right\}$$

所以

$$\left|A' \cap B'\right| = \sum_{x_\mu = y_\vartheta \in X \cap Y} \min\left(a_\mu, b_\vartheta\right) \tag{3}$$

记 $m_A = \left|A'\right|$，$m_B = \left|B'\right|$，将式(3)代入式(2)得

$$p\left(A,B\right) = \frac{2\left|A' \cap B'\right|}{m_A + m_B} \tag{4}$$

4.3 基于布隆过滤器的隐私信息匹配方案

4.3.1 设置参数

每个智能终端用户首次使用本隐私信息匹配应用时，都必须设定自己的属性向量集合。假设发起者 Alice 和某个响应者 Bob 的属性向量集合依次记为 $A = \left\{\left\langle x_1, a_1\right\rangle, \left\langle x_2, a_2\right\rangle, \cdots, \left\langle x_{n_A}, a_{n_A}\right\rangle\right\}$，$B = \left\{\left\langle y_1, b_1\right\rangle, \left\langle y_2, b_2\right\rangle, \cdots, \left\langle y_{n_B}, b_{n_B}\right\rangle\right\}$，其中属性 x_i 和 y_i 的内容和个数均由 Alice 和 Bob 自行设定，并且 $n_A \leqslant n$，$a_i, b_j \in [1, l]$。事先设定常数 n 和 l，一般 n 为 100 或 200，l 为 7 或 10，足以表示用户的所有属性并区分对每个属性的偏好差异即可。为了简化匹配问题，不妨设所有用户对某一个属性的表达方式是唯一的。另外，令 $m=nl$，A' 和 B' 分别表示 A 和 B 离散后的集合，且 $\left|A'\right| = m_A$，$\left|B'\right| = m_B$，$s' = \left|A' \cap B'\right|$，$p\left(A,B\right)$ 表示 A 和 B 的相似度函数。

(w,m,k,H)-布隆过滤器中，参数 w 表示布隆过滤器的长度，m 表示编码的元素个数，k 表示散列函数的个数，H 表示散列函数集合。散列函数集合 $H = \left\{h_i\right\}_{i=0}^{k-1}$ 中的散列函数值域均为 $[0, w-1]$，且相互独立。$BF_{A'}$、$BF_{B'}$ 和 $BF_\cap$ 分别表示编码了集合 A'、B' 和 $A' \cap B'$ 中所有元素的布隆过滤器，其第 i ($0 \leqslant i \leqslant w-1$)位依次记为 $BF_{A'}[i]$、$BF_{B'}[i]$ 和 $BF_\cap[i]$。t_A 和 t_B 分别表示 $BF_{A'}$ 和 $BF_{B'}$ 中 1 的个数。

4.3.2 生成布隆过滤器

利用用户设定的属性向量集合，离线生成相应的布隆过滤器。以 Alice 和 Bob 为例说明具体过程。

算法 1　生成布隆过滤器

输入：A set $S = \left(\left|S\right| \leqslant m\right), w, m, k, H = \left\{h_i\right\}_{i=0}^{k-1}$

输出：(w,m,k,H)-Bloom filter BF_S

1) BF_S=new m-element array of bit strings;
2) for(i=0; $i \leqslant w-1$; i++) do
3) $BF_S\left[i\right] = 0$;
4) end
5) for each $s \in S$ do
6) for(i=0;$i \leqslant k-1$;i++) do
7) $j=h_i(s)$;
8) if $j \leqslant w-1$ then
9) $BF_S\left[j\right] = BF_S\left[j\right] \| 1$;
10) else
11) break
12) end
13) end
14) end

如算法 1 所示，Alice 输入 $S = A'$，生成布隆过滤器模块对于每一个 $a_{ij}' = (x_i, f_i(j)) \in A'$，把 w bit 数组 $BF_{A'}$ 的第 $h(a_{ij}')(0 \leqslant i \leqslant k-1)$ 位置为 1，其他位为 0，生成(w, m, k, H)-布隆过滤器 $BF_{A'}$，计算 $BF_{A'}$ 中 1 的个数，记为 $t_{A'}$。同样地，Bob 输入 $S = B'$，生成(w,m,k,H)-布隆过滤器 $BF_{B'}$，计算 $BF_{B'}$ 中 1 的个数，记为 $t_{B'}$。

4.3.3 计算二元向量内积

Alice 输入 $BF_{A'}$，调用计算二元向量内积模块算法 2，利用素数 θ ($\left|\theta\right| > 64$ bit)和 η $\left(\eta > \left(w+1\right)\theta^2\right)$，以及随机数 $c_i\left(i = 1, 2, \cdots, w-1\right)$ 混淆 $BF_{A'}$ 中的 1 和 0，当 $BF_{A'}[i]$ 为 1 时，令 d_i 为 $\theta + c_i + r_i\eta$，否则令 d_i 为 $c_i + r_i\eta$，以此混淆方式生成数组 $D = \left\{d_i\right\}_{i=0}^{w-1}$，并计算所有 $r_i\eta - c_i$ 之和 K 以便最后抵销混淆，输出参数 θ，η，K 和 $D = \left\{d_i\right\}_{i=0}^{w-1}$。调用算法 2 可以离线执行。$\eta$ 和 K 保密。Alice 广播匹配请求，并公开 θ 和 $D = \left\{d_i\right\}_{i=0}^{w-1}$。

算法 2　混淆二元向量

输入：Initiator's Bloom filter $BF_{A'}$ which is a binary vector

输出：θ, η, K and encrypted Bloom filter $D = \left\{d_i\right\}_{i=0}^{w-1}$

1) choose 2 large primes θ and η, where θ is of the length $|\theta| = 64$ bit and $\eta > \left(w+1\right)\theta^2$;

2) set K=0 and choose w positive random numbers $\left(c_0, c_1, \cdots, c_{w-1}\right)$ such that $\sum_{i=0}^{w=1} c_i < \theta - w$;

3) for(i=0;$i \leqslant w-1$,i++) do
4) $a_i' = BF_{A'}\left[i\right]$;
5) choose a random number r_i, compute $r_i\eta$ such that $|r_i\eta| \approx 256$ bit;
6) if $a_i' == 1$ then
7) $d_i = \theta + c_i + r_i\eta$;

8) else

9) $d_i = c_i + r_i\eta$;

10) end

11) $k_i = r_i\eta - c_i$;

12) $K=K+k_i$;

13) end

用户 Bob 为响应者之一，同意匹配，接收 θ 和 D，并执行如下算法。

算法 3 计算二元向量内积混淆值

输入： θ, $D=\{d_i\}_{i=0}^{w-1}$ and responder's Bloom filter $BF_{B'}$

输出： E which is encrypted scalar product of $BF_{A'}$ and $BF_{B'}$

1) set E=0;

2) for(i=0;$i\leqslant w-1$;i++) do

3) $b_i' = BF_{B'}[i]$;

4) if b_i'==1 then

5) $e_i = \theta d_i$;

6) else

7) $e_i = d_i$;

8) end

9) $E=E+e_i$;

10) end

Bob 输入 θ、$BF_{B'}$ 和 $D=\{d_i\}_{i=0}^{w-1}$，调用二元向量内积模块算法 3，把 $BF_{B'}[i]$ 的信息添加到数组 $D=\{d_i\}_{i=0}^{w-1}$ 中，得数组 $e_i\ (i=0,1,\cdots,w-1)$。当 $BF_{B'}[i]$ 为 1 时，$e_i=\theta d_i$，否则 $e_i=d_i$，最后计算并输出 $e_i\ (i=0,1,\cdots,w-1)$的和值 E，向 Alice 发送 E，$t_{B'}$ 和 $m_B=|B'|$。

Alice 接收 E、$t_{B'}$ 和 m_B，还原二元向量内积：

$F=(E+K)\bmod\eta$，$G=\dfrac{F-(F \bmod \theta^2)}{\theta^2}$。将布隆过滤器 $BF_{A'}$ 和 $BF_{B'}$ 视为二元向量，G 即为二元向量 $BF_{A'}$ 和 $BF_{B'}$ 的内积。其正确性将在定理 1 中的引理 2 进行论证。

4.3.4 计算相似度

Alice 计算与 Bob 的相似度

$$\hat{p}(A,B)=\frac{2\left[\ln\left(w-\dfrac{Gw-t_{A'}t_{B'}}{w-t_{A'}-t_{B'}+G}\right)-\ln w\right]}{k\left(\ln(w-1)-\ln w\right)(m_A+m_B)}$$

$\hat{p}(A,B)$ 为如式(4)所示的多元二维向量集合的相似度函数 $p(A,B)$ 的估计公式，其正确性将在定理 1 中进行论证。

4.3.5 确定匹配对象

Alice 和所有响应者 $V_i\in V$ 按上述步骤进行匹配，找到最匹配的用户

$$V_{\text{match}}=\arg\left(\max\left\{p(\text{Alice},V_i)\right\}_{i=1}^{\infty}\right)$$

5 方案正确性和安全性分析

5.1 方案正确性分析

定理 1 A 和 B 的相似度为

$$\hat{p}(A,B)=\frac{2\left[\ln\left(w-\dfrac{Gw-t_{A'}t_{B'}}{w-t_{A'}-t_{B'}+G}\right)-\ln w\right]}{k\left(\ln(w-1)-\ln w\right)(m_A+m_B)}$$

证明 要证本文方案相似度计算公式的正确性，需证引理 1 和引理 2。

引理 1 1) 已知由算法 1 生成的 w bit 布隆过滤器 $BF_{A'}$ 和 $BF_{B'}$，那么

$$\hat{s}'=\frac{\left[\ln\left(w-\dfrac{t_\wedge w-t_{A'}t_{B'}}{w-t_{A'}-t_{B'}+t_\wedge}\right)-\ln w\right]}{k\left(\ln(w-1)-\ln w\right)} \tag{5}$$

其中，$s'=|A'\cap B'|$，$t_{A'}$、$t_{B'}$ 和 $t_\wedge$ 分别表示 $BF_{A'}$、$BF_{B'}$ 和 $BF_\wedge=BF_{A'}\wedge BF_{B'}$ 中 1 的个数。

2) 对于任意 $s_0'<\hat{s}'<s_1'$，有

$$\Pr\left[s'\in\left(s_0',s_1'\right)\right]\geqslant 1-\delta\left(s_0',s_1'\right)$$

其中，$\delta\left(s_0',s_1'\right)=\left(\dfrac{\hat{t}_\wedge\left(s_0'\right)}{t_\wedge-1}\right)^{t_\wedge-1}\mathrm{e}^{t_\wedge-1-\hat{t}_\wedge\left(s_0'\right)}+\mathrm{e}^{-\frac{\left(t_\wedge+1-\hat{t}_\wedge\left(s_1'\right)\right)^2}{2\hat{t}_\wedge\left(s_1'\right)}}$，

$\hat{t}_\wedge(s')=w-t_{A'}-t_{B'}+\dfrac{t_{A'}t_{B'}}{w\left(1-\dfrac{1}{w}\right)^{ks'}}$。

证明 1)设 $BF_\cap$ 表示向算法 1 输入 $S'=A'\cap B'$ 生成的 w-bit 布隆过滤器，$t_\cap$ 为 $BF_\cap$ 中 1 的个数。

对于任意布隆过滤器 BF，令 $S_{BF}=\{i\mid BF[i]=1,\ i\in[0,w-1]\}$，$t_{A'}=\left|S_{BF_{A'}}\right|$，$t_{B'}=\left|S_{BF_{B'}}\right|$，$t_\wedge=\left|S_{BF_\wedge}\right|$，$t_\cap=\left|S_{BF_\cap}\right|$，$S_1=\left(S_{BF_{A'}}-S_{BF_\cap}\right)\cap\left(S_{BF_{B'}}-S_{BF_\cap}\right)$，$r=|S_1|$ 则 $\Pr\{i\in S_1\}=\dfrac{r}{w-t_\cap}$。

因为 H 中的散列函数相互独立，所以 $S_{BF_{A'}}-S_{BF_{\cap}}$ 和 $S_{BF_{B'}}-S_{BF_{\cap}}$ 中的元素相互独立，那么 $\Pr\{i\in S_1\}=\frac{t_{A'}-t_{\cap}}{w-t_{\cap}}\frac{t_{B'}-t_{\cap}}{w-t_{\cap}}$。

由以上两式得 $\frac{r}{w-t_{\cap}}=\frac{t_{A'}-t_{\cap}}{w-t_{\cap}}\frac{t_{B'}-t_{\cap}}{w-t_{\cap}}$，解之得 $\hat{r}=\frac{(t_{A'}-t_{\cap})(t_{B'}-t_{\cap})}{w-t_{\cap}}$，又因为散列函数的碰撞性导致 $t_{\wedge}\neq t_{\cap}$ [27]，且 $t_{\wedge}-t_{\cap}=r$，所以

$$t_{\wedge}-t_{\cap}=\frac{(t_{A'}-t_{\cap})(t_{B'}-t_{\cap})}{w-t_{\cap}} \tag{6}$$

显然，$BF_{\cap}$ 中 1 的个数 $t_{\cap}$ [2]为

$$\hat{t}_{\cap}=wp=w\left[1-\left(1-\frac{1}{w}\right)^{ks'}\right] \tag{7}$$

把式(7)代入式(6)即可得

$$\hat{s}'=\frac{\left[\ln\left(w-\frac{t_{\wedge}w-t_{A'}t_{B'}}{w-t_{A'}-t_{B'}+t_{\wedge}}\right)-\ln w\right]}{k\left(\ln(w-1)-\ln w\right)}$$

2) 任取 $s_0'<\hat{s}'<s_1'$，设随机变量

$$X_i=\begin{cases}1, BF_{\wedge}[i]=1\\ 0, BF_{\wedge}[i]=0\end{cases}, i=0,1,\cdots,w-1$$

则 X_i 相互独立同分布。不妨设 $X_i\sim b(1,p_{\wedge})$，$0<p_{\wedge}<1$，那么 $t_{\wedge}=\sum_{i=0}^{w-1}X_i$，$\hat{t}_{\wedge}=E(t_{\wedge})=\sum_{i=0}^{w-1}p_{\wedge}=wp_{\wedge}$。

根据 Chernoff 不等式[27]可知对于任意 $\delta_0>0$ 和 $\delta_1>0$，有

$$\Pr\left[t_{\wedge}>(1+\delta_0)\hat{t}_{\wedge}\right]<\left[\frac{\mathrm{e}^{\delta_0}}{(1+\delta_0)^{1+\delta_0}}\right]^{\hat{t}_{\wedge}} \tag{8}$$

$$\Pr\left[t_{\wedge}<(1-\delta_1)\hat{t}_{\wedge}\right]\hat{s}'\leqslant s_0'\mathrm{e}^{\frac{-\delta_1^2\hat{t}_{\wedge}}{2}} \tag{9}$$

而对于任意 $\hat{s}'\leqslant s_0'$，有

$$\Pr\left(t_{\wedge}\geqslant\hat{t}_{\wedge}\mid s'\right)\hat{s}'\leqslant s_0'\ \Pr\left(t_{\wedge}\geqslant\hat{t}_{\wedge}\mid s_0'\right)$$

令 $\delta_0=\frac{t_{\wedge}-1-\hat{t}_{\wedge}(s_0')}{\hat{t}_{\wedge}(s_0')}$，代入式(8)得

$$\Pr\left(t_{\wedge}\geqslant(1+\delta_0)\hat{t}_{\wedge}\mid s'\right)\leqslant\left[\frac{\hat{t}_{\wedge}(s_0')}{t_{\wedge}-1}\right]^{t_{\wedge}-1}\mathrm{e}^{t_{\wedge}-1-\hat{t}_{\wedge}(s_0')}$$

同理，对于任意 $s_0'\geqslant\hat{s}'$，令 $\delta_1=\frac{\hat{t}_{\wedge}(s_1')-t_{\wedge}-1}{\hat{t}_{\wedge}(s_1')}$，有

$$\Pr\left(t_{\wedge}\leqslant(1-\delta_1)\hat{t}_{\wedge}\mid s'\right)\leqslant\mathrm{e}^{\frac{-\left(t_{\wedge}+1-\hat{t}_{\wedge}(s_1')\right)^2}{2\hat{t}_{\wedge}(s_1')}}$$

综上可得

$$\Pr[s'\in(s_0',s_1')]=\Pr[(1-\delta_1)\hat{t}_{\wedge}\leqslant t_{\wedge}\leqslant(1-\delta_0)\hat{t}_{\wedge}]\leqslant 1-\delta$$

其中，$\delta(s_0',s_1')=\left[\frac{\hat{t}_{\wedge}(s_0')}{t_{\wedge}-1}\right]^{t_{\wedge}-1}\mathrm{e}^{t_{\wedge}-1-\hat{t}_{\wedge}(s_0')}+\mathrm{e}^{\frac{-\left(t_{\wedge}+1-\hat{t}_{\wedge}(s_1')\right)^2}{2\hat{t}_{\wedge}(s_1')}}$，且由引理 1 中式 (9) 和式 (10) 易知 $\hat{t}_{\wedge}(s')=w-t_{A'}-t_{B'}+\frac{t_{A'}t_{B'}}{w\left(1-\frac{1}{w}\right)^{ks'}}$。

引理 2　由算法 2 和算法 3 计算出 $G=t_{\wedge}$。

证明　根据算法 2 的第 7)行和算法 3 的第 5)行，当且仅当 $BF_{A'}[i]=BF_{B'}[i]=1$ 时，对应的 d_i 中会出现 θ^2 项，所以 F 中 θ^2 项的系数 G 等于 $BF_{A'}$ 和 $BF_{B'}$ 的内积，即

$$G=BF_{A'}BF_{B'}=\sum_{i=0}^{w-1}BF_{A'}[i]BF_{B'}[i]=t_{\wedge} \tag{10}$$

具体地，通过实例来说明 $G=t_{\wedge}$ 的正确性。假设 $BF_{A'}=\{1,1,0,0,1\}$，$BF_{B'}=\{1,0,1,0,1\}$，则算法 2 输出 $d_0=\theta+c_0+r_0\eta$，$d_1=\theta+c_1+r_1\eta$，$d_2=c_2+r_2\eta$，$d_3=c_3+r_3\eta$ 和 $d_4=c_4+r_4\eta$。

Bob 输入 θ、$BF_{B'}$ 和 $\{d_i\}_{i=0}^{w-1}$，调用算法 3，计算 $e_0=\theta^2+c_0\theta+r_0\theta\eta$，$e_1=\theta+c_1+r_1\eta$，$e_2=c_2\theta+r_1\theta\eta$，$e_3=c_3+r_3\eta$，$e_4=\theta^2+c_4\theta+r_4\theta\eta$，$k_i=r_i\eta-c_i$（$0\leqslant i\leqslant 4$），输出 $E=\sum_{i=0}^{4}e_i$，$K=\sum_{i=0}^{4}k_i$。

Alice 接收 E 和 K，计算

$$\begin{aligned}F&=(E+K)\bmod\eta=\sum_{i=0}^{4}(d_i+k_i)\bmod\eta\\&=\left[2\theta^2+\theta+(c_0+c_2+c_4)(\theta-1)\right]\bmod\eta\end{aligned}$$

而由算法 2 可知

$$\begin{aligned}&\left[2\theta^2+\theta+(c_0+c_2+c_4)(\theta-1)\right]\\&<2\theta^2+\theta+\sum_{i=0}^{4}\theta<2\theta^2+\theta(1+\theta-w)\\&<2\theta^2+\theta^2=3\theta^2<(w+1)\theta^2<\eta\end{aligned}$$

所以

$$F = 2\theta^2 + \theta + (c_0 + c_2 + c_4)(\theta - 1)$$

同时因为 $\theta + (c_0 + c_2 + c_4)(\theta - 1) < 2\theta^2$，可得

$$G = \frac{F - (F \bmod \theta^2)}{\theta^2} = \frac{2\theta^2}{\theta^2} = 2 = t_\wedge$$

综上所述，将式(10)代入式(5)即得

$$\hat{s}' = \frac{\left[\ln\left(w - \frac{t_\wedge w - t_{A'} t_{B'}}{w - t_{A'} - t_{B'} + t_\wedge}\right) - \ln w\right]}{k(\ln(w-1) - \ln w)}$$

代入式(4)得 $\hat{p}(A,B) = \frac{2\left[\ln\left(w - \frac{Gw - t_{A'} t_{B'}}{w - t_{A'} - t_{B'} + G}\right) - \ln w\right]}{k(\ln(w-1) - \ln w)(m_A + m_B)}$

证毕。

5.2 方案安全性分析

5.2.1 发起者的隐私安全

定理 2 如果算法 2 中的混淆方法[6]是安全的，本隐私信息匹配方案能保护 Alice 的隐私信息，即 Bob 不可能知道关于 Alice 属性集的任何信息。

证明 对发起者 Alice 而言，在整个匹配方案中，响应者 Bob 能够知道的所有关于 Alice 的信息只有参数 θ 和 $D = \{d_i\}_{i=0}^{w-1}$，其中 D 是 $BF_{A'}$ 混淆所得。因为 $d_i = \begin{cases} \theta + c_i + r_i\eta, BF_{A'}[i] = 1 \\ c_i + r_i\eta, BF_{A'}[i] = 0 \end{cases}$，对于 Bob 和窃听者而言，随机数 c_i 和 $r_i\eta$ 未知，难以区分 d_i 是由 $\theta + c_i + r_i\eta$ 表示，还是由 $c_i + r_i\eta$ 表示，所以无法推测 $BF_{A'}[i]$ 是否等于 1。此外，每一对随机数 $(c_i, r_i\eta)_{i=0}^{w-1}$ 只使用一次且相互独立，这使 d_i 和 d_i' 也相互独立，所以即使恶意攻击者 Bob 与 Alice 匹配多次，Bob 也难以从 d_i 和 d_i' 中推测出敏感信息。综上可知，响应者 Bob 和窃听者无法从 θ 和 $D = \{d_i\}_{i=0}^{w-1}$ 推测出 Alice 的任何一个属性或偏好值，所以 Alice 的个人隐私是安全的。证毕。

5.2.2 响应者的隐私安全

定理 3 如果算法 2 中的混淆方法[6]是安全的，本隐私信息匹配方案能保护 Bob 的隐私信息，即 Alice 除了相似度 p 以外，无法知道关于 Bob 属性集的任何信息。

证明 对响应者 Bob 而言，发起者 Alice 只能够知道 Bob 的 $t_{B'}$、m_B 和 E，其中 $t_{B'}$ 为布隆过滤器 $BF_{B'}$ 中 1 的个数，m_B 为 B' 的模，$E = \sum_{i=0}^{w-1} e_i$，$e_i = \begin{cases} \theta d_i, BF_{B'}[i] = 1 \\ d_i, BF_{B'}[i] = 0 \end{cases}$。显然，由于 θ 是公开的，只需知道 e_i，就能推测出 Bob 的 $BF_{B'}[i]$ 是否为 1。然而，由算法 3 可知除了 Bob 以外，其他人都不知道 e_i 的值，只知道算法 3 输出的 E。E 中含有多个 $(c_i + r_i\eta)$ 和 $(c_i + r_i\eta)\theta$，很好地隐藏了 $BF_{B'}$ 的相关信息，故发起者 Alice 和窃听者无法从 E 推测出 Bob 的 $BF_{B'}$，更不用说推断出任何一个属性或偏好值，所以 Bob 的个人隐私是安全的。证毕。

5.2.3 抵御蛮力攻击

由于布隆过滤器元素查询能以极低的错误率判断布隆过滤器 BF_A 是否编码某一属性向量 (x,a)，所以布隆过滤器不具有保密性，无法抵御蛮力攻击。文献[2]的高级方案采用布隆过滤器实现隐私信息匹配，为了隐藏 Alice 的属性集 A，Alice 在公开散列函数集合 $H_k(|H_k| = k)$ 中随机选择 $l < k$ 个散列函数，剩下的(k−1)个散列函数从 $\bar{H}_k$ 中秘密选取，这能一定程度地隐藏 Alice 的属性集合 A，但是该方案无法抵御蛮力攻击，论证过程如下。

设响应方 Bob 或窃听者对 Alice 发起蛮力攻击。由于 H_k 是公开的，攻击者取遍有可能的属性向量 (x,a)，对 Alice 的布隆过滤器 BF_A 执行基于 H_k 的元素查询，可获得属性向量集合 $\tilde{A} = \left\{(\tilde{x},\tilde{a}) \middle| \forall h \in H_k, BF_A\left[h_{i_j}(\tilde{x},\tilde{a})\right] = 1, j \in [1,d], d \geqslant l\right\}$，又因为对于任意 $(x,a) \in A$，$h \in H$，有 $BF_A[h(x,a)] = 1$，所以 $A \subset \tilde{A}$。

攻击者谎称初始属性向量集合为 $B = \tilde{A}$，作为参与方和 Alice 匹配，由于该方案的相似度为匹配双方属性集合中共同属性向量的个数即 $p = |A \cap B|$，又因 $A \subset \tilde{A}$，所示攻击者将获得最大相似度为 $\tilde{p} = |A|$，成为最优匹配用户。

此外，攻击者任取 $(\tilde{x},\tilde{a}) \in \tilde{A}$，设定初始属性集为 $B = \{(\tilde{x},\tilde{a})\}$，向 Alice 发起匹配，计算相似度函数 $p = |A \cap B|$，若 $\tilde{p} = 1$，可断定 $(\tilde{x},\tilde{a}) \in A$，否则认为 $(\tilde{x},\tilde{a}) \notin A$。攻击者按以上策略至多匹配 $|\tilde{A}|$ 次即能知道 Alice 的属性集合 A。

综上可知，该方案无法抵御蛮力攻击。然而，本文提出的隐私信息匹配方案对双方的布隆过滤器经过算法 2 和算法 3 的处理，加入了大量随机数，使匹配双方无法知悉对方的布隆过滤器，无处执行元素查询获得 $\tilde{A}$，无法进行蛮力攻击。

5.2.4 抵御无限制输入攻击

无限制输入攻击[25]指参与匹配的某一方是恶意攻击者，通过设定尽可能多的属性和尽可能大的偏好程度发起攻击，以期获得较高的相似度，成为最优匹配者。

为了与 Bob 的相似度尽可能的大，Alice 多次谎称属性集合$\{(x_i,a_i)\}_{i=1}^{n_A}$中$n_A$非常大并且每一个的偏好$a_i=l$。不妨设在此情境下 Bob 的属性集合是固定的，那么存在$n_0\in N$，当$n_A\geqslant n_0$时，有$\sum_{x_\mu=y_\theta\in X\cap Y}\min(a_\mu,b_\theta)=\sum_{j=1}^{n_B}b_j=m_B$，Alice 和 Bob 的相似度为$p=\frac{2m_B}{m_A+m_B}$，而$\frac{\mathrm{d}p}{\mathrm{d}m_A}=\frac{-2m_B}{(m_A+m_B)^2}<0$，所以当 Alice 输入的属性越多时，相似度将会减小，即本方案能抵御来自 Alice 的无限制输入攻击。同理，本方案亦能抵御来自 Bob 的无限制输入攻击。

6 实验仿真

假设在 PMSN 中，每位手机用户的属性向量集合包含 n 个属性向量，每个属性的偏好取值区间为$[1,l]$，至多离散 $m(m=nl)$ 个属性向量。散列函数个数为 k，布隆过滤器的长度为 $w=1.5km$ bit 以保证较低的错误率。为了保证方案的安全性，将算法 2 和算法 3 中的随机数取为 256 bit。在参与比较的方案里，计算开销主要涉及 SHA-1、1 024 bit 模幂、256 bit 乘法、256 bit 加减法，依次表示为 *h*、*exp*、*mul* 和 *add*/*sub*。通信开销指用户接收和发送的数据，以 bit 为单位。如表 1 所示，本文方案的在线计算开销远小于 Cristofaro[13]和 Zhang[2]的方案，离线计算开销稍大于 Zhang[2]的方案，通信开销高于 Zhang[2]方案。

SHA-1、*exp*、*mul* 和 *add*/*sub* 运算在小米智能手机 1.7 GHz, Cortex-A8 processor, 1 GB RAM, Android4.1.2 上执行的平均时间、最大时间、最小时间、中位数以及标准差如表 2 所示。图 2 比较了本文、Cristofaro[13]和 Zhang[2]各个方案的离线计算开销、在线计算开销、通信量和执行时间。本文方案在离线计算开销和通信量上稍逊于 Zhang[2]的方案，但是本文方案的在线计算开销和执行时间均为最小。图 3 比较了$m\in[100,1\,000]$和$\frac{w}{km}$=1.2、1.5 和 3 时，本文方案相似度的相对误差$r=\frac{\hat{p}-p}{p}=\frac{\hat{s}'-s'}{s'}$。总地来说，相似度的相对误差较小，精确度较高。

如图 2(a)所示，各个方案的离线计算开销关于 m 单调递增。显然，Cristofaro[13]的方案因为涉及模幂运算，离线计算开销最大。Zhang[2]的方案比本文方案的离线计算开销略小一些，然而，离线计算开销可以被预先离线计算，不占用执行时间。

图 2(b)比较了在 m 取 100 到 1 000 时各个方案的在线计算开销。在线计算开销直接影响执行时间的长短，所以期望尽可能地降低在线计算开销。本文方案在线只需计算 w 个 256 bit 乘法和 w 个 256 bit 加减法，比其他方案的在线计算开销都更小，其中 Cristofaro[13]的方案涉及大量模幂和 SHA-1 运算，

表 1　　计算开销

方案	参与方	离线计算开销	在线计算开销	通信量/bit
Cristofaro 方案	发起方	$(2m+2m^2)exp,(2m)h$	$(m+m^2)exp,(m)h$	$3m\cdot1\,024$
	响应方	$(m+m^2)exp,(2m)h$	$(2m)exp$	$4m\cdot1\,024$
Zhang 方案	发起方	—	$(km)h$	—
	响应方	$(km)h$	—	w
本文方案	发起方	$(km)h$,$(2w)mul$,$(4w)add/sub$	—	$256w$
	响应方	$(km)h$	$(2w)mul$,$(4w)add/sub$	320

表 2　　运算执行时间(单位：μs)

运算	平均值	最大值	最小值	中位数	标准差
h	6.75	14	6	7	13.81
$exp(\times10^3)$	11.21	15.57	10.59	10.82	0.76
mul	1.599	1.760	1.493	1.644	0.081
add/*sub*	0.205 5	0.345 2	0.180 6	0.193 8	0.022

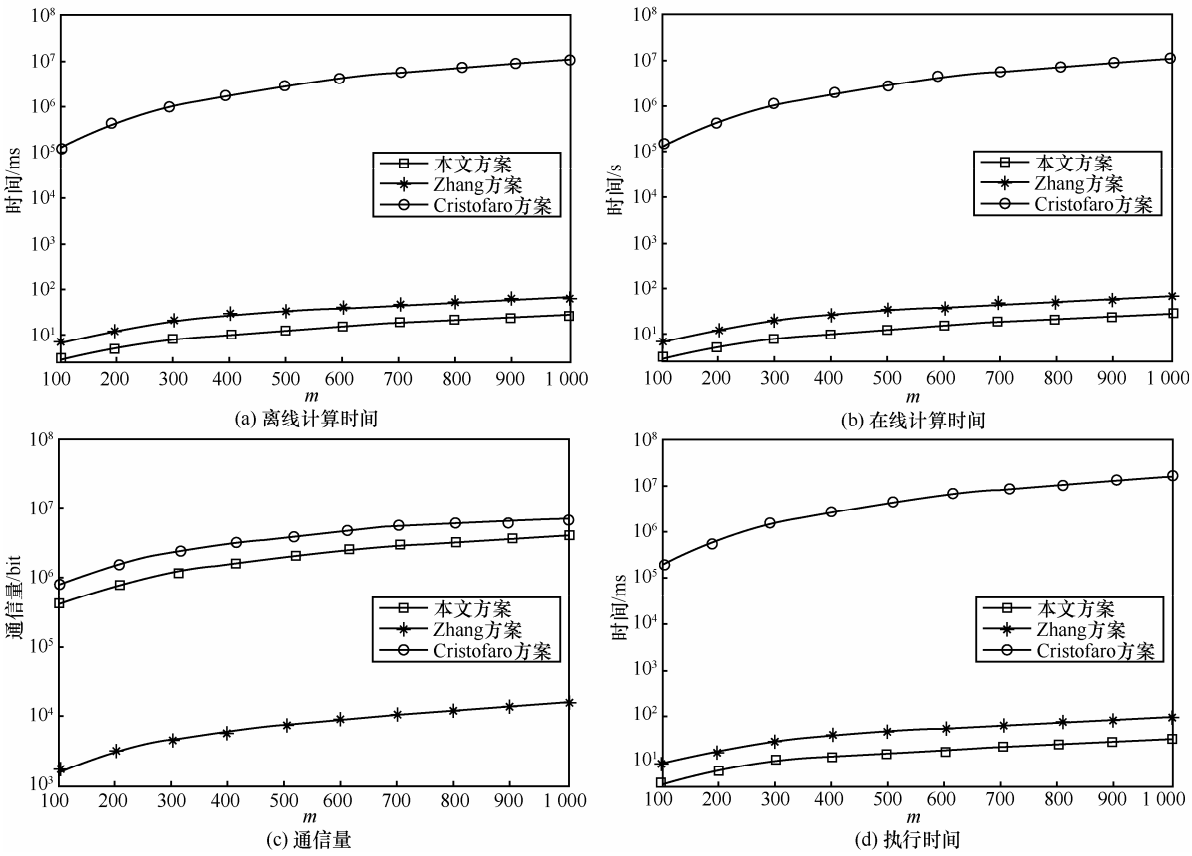

图 2　k=10，$m\in[100,1\,000]$，$\frac{w}{km}$=1.5 时，各方案的性能比较

在线计算开销最大。

图 2(c)比较各方案的通信开销，Cristofaro[13]的方案通信量最大。本文方案由于需花费较多通信量传递属性偏好信息，其通信开销大于 Zhang[2]的方案。然而，本实验采用蓝牙、Wi-Fi 等传输方式，传输速率为 900 bit/s，通信时间较短。

如图 2(d)所示，本文方案的执行时间最短，m 取 100 到 1 000 时的执行时间不超过 0.035 s。例如 m=500 时本文方案执行一次匹配仅需 0.017 s，然而，Cristofaro[13]和 Zhang[2]的方案完成一次匹配分别需要 0.050 s 和 4.144×10^3 s。

如图 3 所示，令 k=10，$s'=0.5m$，当布隆过滤器的长度 w 固定时，相似度函数的相对误差关于 m 单增。因为当 w 不变时，编码的元素个数 m 增大，出现 $BF_{A'}[i]=BF_{B'}[i]=1$ 的概率增大，导致 $t_\wedge$ 偏大，故误差增大。当 $\frac{w}{km}$ 增大时，相似度函数的相对误差减小。这是因为对于相同的 m，w 增大，出现 $BF_{A'}[i]=BF_{B'}[i]=1$ 的概率减小，使 $t_\wedge$ 偏小，误差减小。

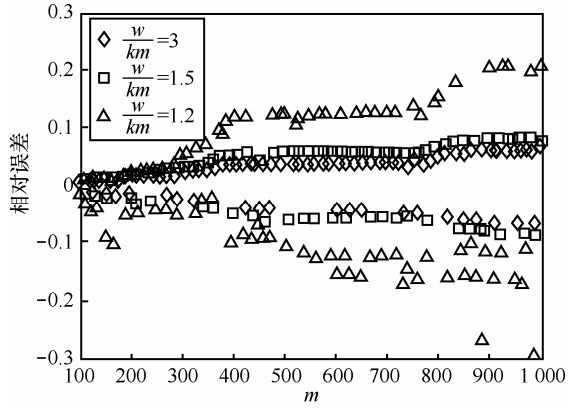

图 3　$m\in[100,1\,000]$，$\frac{w}{km}$ = 1.2、1.5 和 3 时，相似度的相对误差

7　结束语

本文针对 PMSN 智能终端用户的隐私信息匹配问题提出了一种基于布隆过滤器的轻量级隐私信息匹配方案。本方案不依赖于可信第三方，首先基于 Dice 相似性系数建立了一种新的二维向量相似度函数，然后把私有数据集合中的

所有元素编码在相应的布隆过滤器中，最后将布隆过滤器视为二元向量，采用混淆的方法计算二元向量内积，得到基于布隆过滤器的相似度估计值，从而实现用户之间的隐私信息匹配。基于布隆过滤器的相似度估计大幅降低了计算开销。正确性及安全性分析表明了方案的正确性和安全性。实验结果进一步验证了所提出方案的有效性和高效性。

参考文献：

[1] LU R, LIN X, LIANG X, et al. A secure handshake scheme with symptoms-matching for healthcare social network[J]. Mobile Networks & Applications—Special Issue on Wireless & Personal Comm, 2011, 16(6):683-694.

[2] SUN J, ZHANG R, ZHANG Y. Privacy-preserving spatiotemporal matching[A]. Proceedings of the 32nd IEEE International Conference on Computer Communications[C]. Turin, Italy, 2013. 800-808.

[3] LI M, CAO N, YU S, et al. FindU: privacy-preserving personal profile matching in mobile social networks[A]. Proceedings of the 30th IEEE International Conference on Computer Communications[C]. Shanghai, China, 2011. 2435-2443.

[4] YANG Z, ZHANG B, DAI J, et al. E-SmallTalker: a distributed mobile system for social networking in physical proximity[A]. Proceedings of the IEEE 30th International Conference on Distributed Computing Systems[C]. Genoa, Italy, 2010. 468-477.

[5] BLOOM B H. Space/time trade-offs in hash coding with allowable errors[J]. Communications of the ACM, 1970, 13(7): 422-426.

[6] LU R, LIU X, SHEN X. SPOC: a secure and privacy-preserving opportunistic computing framework for mobile-healthcare emergency[J]. IEEE Transactions on Parallel and Distributed Systems, 2013, 24(3): 614-624.

[7] FREEDMAN M J, NISSIM K, PINKAS B. Efficient private matching and set intersection[A]. Advances in Cryptology - EUROCRYPT 2004[C]. Springer Berlin Heidelberg, 2004.1-19.

[8] KISSNER L, SONG D. Privacy-preserving set operations[A]. Advances in Cryptology – CRYPTO 2005[C]. Springer Berlin Heidelberg, 2005.241-257.

[9] DONG C, CHEN L, WEN Z. When private set intersection meets big data: an efficient and scalable protocol[A]. Proceedings of the 21st ACM Conference on Computer and Communications Security[C]. Scottsdale, USA, 2013. 789-800.

[10] HAZAY C, NISSIM K. Efficient set operations in the presence of malicious adversaries[A]. Proceedings of the 13th International Conference on Practice and Theory in Public Key Cryptography[C]. Paris, France, 2010. 383-433.

[11] JARECKI S, LIU X. Fast secure computation of set intersection[A]. Proceedings of the 7th International Conference on Security and Cryptography for Networks[C].Amalfi, Italy, 2010. 418-435.

[12] CRISTOFARO E D, TSUDIK G. Practical private set intersection protocols with linear complexity[A]. Proceedings of the 14th International Conference on Financial Cryptography and Data Security[C]. Tenerife, Spain, 2010. 143-159.

[13] G. CRISTOFARO E D, KIM G, TSUDIK G. Linear-complexity private set intersection protocols secure in malicious model[A]. Proceedings of the 16th International Conference on the Theory and Application of Cryptology and Information Security[C]. Singapore, Singapore, 2010. 213-231.

[14] KERSCHBAUM F. Outsourced private set intersection using homomorphicencryption[A]. Proceedings of the Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security[C]. Seoul, Korea, 2012. 85-86.

[15] AGRAWAL R, EVFIMIEVSKI A, SRIKANT R. Information sharing across private databases[A]. Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data[C]. San Diego, USA, 2003. 86-97.

[16] VAIDYA J, CLIFTON C. Secure set intersection cardinality with application to association rule mining[J]. Journal of Computer Security, 2002, 1(13): 593-622.

[17] VON M, BADER A M, KUHN M, WATTENHOFER R. Veneta: serverless friend-of-friend detection in mobile social networking[A]. Proceedings of the 2008 IEEE International Conference on Wireless & Mobile Computing, Networking & Communication[C]. Avignon, France, 2008. 184-189.

[18] ZHANG R, ZHANG Y, SUN J, et al. Fine-grained private matching for proximity-based mobile social networking[A]. Proceedings of the 31st IEEE International Conference on Computer Communications[C]. Orlando, USA, 2012. 1969-1977.

[19] NIU B, LI X, ZHU X, et al. Are you really my friend? Exactly spatio-temporal matching scheme in privacy-aware mobile social networks[A]. Proceedings of the 10th International Conference on Security and Privacy in Communication Networks[C]. Beijing, China, 2014.

[20] NIU B, HE Y, LI F H, LI H. Achieving secure friend discovery in social strength-aware PMSNs[A]. Proceedings of the the 34th Anniversary of the Premier International Conference for Military Communications[C]. Florida, USA, 2015.

[21] NIU B, ZHU X, LIU J, et al. Weight-aware private matching scheme for Proximity-based mobile social networks[A]. Proceedings of the IEEE 2003 Global Communications Conference[C]. Atlanta, USA, 2013. 3170-3175.

[22] FU Y, WANG Y. BCE: a privacy-preserving common-friend estimation method for distributed online social networks without cryptography[A]. Proceedings of the 7th International Conference on Communications and Networking in China[C]. Kunming, China, 2012. 212-217.

[23] MANY D, BURKHART M, DIMITROPOULOS X. Fast private set

Operations with SEPIA[EB/OL]. http://www.researchgate.net/ publication/26652 4458_Fast_Private_Set_Operations_with_SEPIA.

[24] POSE P, GUO H, KRANAKISE.On the false-positive rate of bloom filters[J]. Information Processing Letters, 2008, 108(4):210-213.

[25] NIUB, ZHU X, ZHANG T. P-match: priority-aware friend discovery for proximity-based mobile social networks[A].Proceedings of the10th IEEE International Conference on Mobile Ad-Hoc and Sensor Systems[C]. Hangzhou, China, 2013. 351-355.

[26] CHEN Z, ZHU B. Some formal analysis of rocchio's similarity-based relevance feedback algorithm[J]. Algorithms and Computation, 2000, 19(69): 108-119.

[27] PAPAPERTROU W, SIBERSKI W. Cardinality estimation and dynamic length adaptation for bloom filters[J]. Distributed & Parallel Databases, 2010, 28(2-3): 119-156(38).

作者简介：

万盛（1987-），男，江苏南通人，西安电子科技大学博士生，主要研究方向为网络安全与隐私保护。

何媛媛（1985-），女，湖北松滋人，中国科学院信息工程研究所博士生，主要研究方向为信息安全、隐私保护。

李凤华[通信作者]（1966-），男，湖北浠水人，博士，中国科学院信息工程研究所副总工程师、研究员、博士生导师，主要研究方向为网络与系统安全、隐私计算、可信计算。E-mail: lfh@iie.ac.cn。

牛犇（1984-），男，陕西西安人，博士，中国科学院信息工程研究所助理研究员，主要研究方向为网络安全、隐私计算。

李晖（1968-），男，河南灵宝人，博士，西安电子科技大学教授、博士生导师，主要研究方向为密码学、无线网络安全、云计算安全、信息论与编码理论。

王新宇（1989-），男，甘肃平凉人，中国科学院信息工程研究所博士生，主要研究方向为信息安全、隐私保护。