# 基于0-1编码的参与式感知隐私保护的数据价值匹配方案

刘梦君[1,3] 刘树波[2] 丁永刚[1,2]
(湖北大学教育学院 武汉 430062)[1] (武汉大学计算机学院 武汉 430072)[2]
(湖北大学计算机与信息工程学院 武汉 430062)[3]

**摘 要** 在参与式感知中,满足数据请求者对数据类型和数据价值匹配的要求,同时保护请求者和提供者的个人隐私,是普及参与式感知需要解决的问题。鉴于此,提出了一种基于0-1编码的隐私保护的数据价值匹配方案,它将用户数据价值转换成0-1编码,然后使用时空高效的布隆过滤器执行价值匹配,在保护了用户数据价值隐私的同时,完成了数据价值的高效匹配。理论分析和仿真实验论证了所提方案的正确性、安全性和高效性。

**关键词** 参与式感知,隐私保护,数据价值,0-1编码
**中图法分类号** TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.03.021

## 0-1 Code Based Privacy-preserving Data Value Matching in Participatory Sensing

LIU Meng-jun[1,3] LIU Shu-bo[2] DING Yong-gang[1,2]
(School of Education, Hubei University, Wuhan 430062, China)[1]
(School of Computer, Wuhan University, Wuhan 430072, China)[2]
(School of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China)[3]

**Abstract** In participatory sensing, protecting both the privacy of requestor and provider while satisfying the special requirement of data types and data value of data requesters at the same time, is a crucial problem before the widespread of participatory sensing application. This paper put forward a 0-1 encode based privacy-preserving data value matching scheme. It first converts two users' data value into two 0-1 code sets, and then matches the two sets with a spatial-timing efficient data structure—bloom filter, thus preserving the privacy of data value while completing efficient data value matching. Theoretical analysis and simulation experiment prove the correctness, safety and effectiveness of the proposed scheme.

**Keywords** Participatory sensing, Privacy-preserving, Data value, 0-1 code

## 1 引言

移动智能终端的迅速普及催生了一类新型的分布式共享应用——参与式感知[1-3]。用户出于个人兴趣或经济因素,有意识地将私人手机内置传感器(GPS、加速度、时间、图像、温度等)采集到的信息通过移动社交媒体进行共享,从而构成了一个感知能力强大的巨型"传感器网络",这就催生了一类特殊的无线传感器应用——参与式感知。

出于共享效率的考虑,参与式感知中的数据请求者既对数据提供者共享的数据类型数有要求,也对数据的量有要求,前者涉及的是数据类型匹配,后者则涉及的是数据价值匹配。例如在环境数据共享应用中,用户不仅对数据类型(如PM2.5)感兴趣,某些时候也对采集者采集数据的频率有一定的需求,如需要10次/小时的PM2.5的值、20次/小时的空气污染物的值等。显然,采集频率高的数据蕴含的信息更为丰富,价值也更高。与数据类型数匹配相似,数据价值匹配需要数据请求者和数据提供者在共享数据前进行匹配操作。

然而,根据笔者在前期数据类型匹配研究工作中的论述[4-6],若直接开展数据匹配操作,则往往会泄露用户的个人隐私。由于用户共享的数据价值信息和数据类型信息都反映了他的个人感知偏好,若直接将这些信息公开,则会泄露用户个人隐私,给用户带来各类潜在的安全风险。但遗憾的是,虽然参与式感知中的隐私保护数据类型匹配问题已经得到了一定研究[7-14],但数据价值匹配问题还未能得到较好解决。

虽然数据价值匹配问题可以看作是多次隐私保护集合交问题[7],但在参与式感知环境中简单应用现有解决方案会面临多种挑战。直观上,可以对需求者和采集者的每种数据类型的价值进行隐私保护集合交求解,如果对于所有(或达到某

到稿日期:2016-12-12 返修日期:2017-03-20 本文受国家自然科学基金面上项目:面向移动位置服务的空间位置大数据差分隐私保护研究(41671443),武汉市科技局应用基础研究计划:支持移动位置服务的时空数据隐私保护技术研究(2016010101010024)资助。
刘梦君(1988—),男,博士,讲师,主要研究方向为移动/无线网络、移动社交/分布式系统上的安全与隐私,E-mail:lmj_whu@163.com;刘树波(1970—),男,教授,博士生导师,主要研究方向为物联网安全与隐私保护、数据隐私挖掘与发布;丁永刚(1965—),女,博士,副教授,主要研究方向为数据挖掘,E-mail:21269974@qq.com(通信作者)。

个数量)的数据类型,采集者能提供的价值都大于需求者需要的价值,则采集者满足需求者的数据价值需求。但现有的隐私保护集合交要么计算量大[8,11,14],要么只适用于特定的应用环境[9-10]或者需要第三方服务器的参与[12-13],不能直接用于资源受限的参与式感知下隐私保护的数据价值匹配问题中。

基于这些挑战和不足,文中研究如何在没有第三方服务器参与的参与式感知环境下,为数据请求者找到满足需求的采集者,同时保障参与双方的数据价值隐私。文中将通过一个精心设计的基于 0-1 编码及散列消息鉴别码(HMAC)的数据价值匹配方案来高效地解决上述问题,并通过理论分析和仿真实验验证所提方案的安全性和高效性。

本文第 2 节给出了系统模型和问题定义;第 3 节给出了详细的方案设计;第 4 节给出了方案的性能分析和仿真实验结果;最后总结全文。

## 2 系统模型与问题定义

### 2.1 系统模型

系统中主要有两类实体,分别是数据采集者和数据需求者,用户既可以是数据采集者也可以是数据需求者。数据需求者因为某种需求,需要请求相关用户采集的数据;数据采集者则是拥有智能终端的用户,其根据自身条件,可以采集不同类型的数据,由于采集时间、地点、频率等的不同,不同数据采集者的数据价值也不同。

### 2.2 安全模型

系统中的用户是诚实而好奇的,即用户会按照方案描述的步骤执行,但是会尽可能地从系统输入中获取更多的不应被其知晓的信息。例如,不满足匹配条件的用户可能想获知需求者需要的数据类型和每种数据的需求量。本文不考虑恶意用户的欺骗行为,即本文中的用户在匹配过程中的数据类型与价值均是其真实拥有的。本文只考虑数据价值匹配阶段的用户数据价值隐私匹配问题,不涉及用户身份验证等问题,系统初始化时所需要的操作也不在本文的考虑范围内。

### 2.3 问题描述

对于系统中的两个用户 Alice(数据需求者)和 Bob(数据采集者),他们已经利用现有的隐私保护数据类型匹配方案证明了双方是符合类型匹配条件的用户。假设 Alice 需要的数据类 $NA_i=\{a_i^1,a_i^2,\cdots,a_i^{t_i}\}$,其中 $t_i=n$,即 $|NA_i|=n$。Alice 对每种数据的需求量 $CA_i=\{x_1,x_2,\cdots,x_n\}$,其中 $x_k(1\leqslant k\leqslant n)$为 $a_i^k$ 的需求量;Bob 与 Alice 匹配的 $n$ 种数据类型可提供的数据价值组成集合 $CA_j=\{y_1,y_2,\cdots,y_n\}$,Alice 需要知道 Bob 的每个数据值是否满足自己的需求,即需要判断 $\forall i\in[1,n]$时 $x_i\geqslant y_i$ 是否都成立或至少 $\delta$($\delta$ 为 Alice 选定的阈值)个成立。本文将该过程定义为数据价值的匹配,如果 $\forall i\in[1,n]$时 $x_i\geqslant y_i$ 都成立或至少 $\delta$ 个成立,则匹配成功,否则匹配失败。

该数据价值匹配过程的目标为:

1)整个匹配过程中只涉及 Alice 和 Bob 的交互,不需要第三方服务器的参与;

2)Alice 能够知道 $\forall i\in[1,n]$时 $x_i$ 与 $y_i$ 的大小关系。

此外,无论结果是否匹配成功,Alice 都不知道 Bob 任何类型数据的数据价值,Bob 也不知道 Alice 任何类型数据的数据价值,即匹配双方的数据价值具有隐私性。

## 3 基于 0-1 编码的隐私保护数据价值匹配方案

### 3.1 准备知识和预处理

#### 3.1.1 HMAC

HMAC 是一种利用密码散列函数构造消息认证码的方法,其利用 hash 散列函数,以一个密钥和一个消息作为输入,生成一个消息摘要作为输出。HMAC 的实现原理:用公开函数和密钥产生一个固定长度的值作为认证标识,用这个标识鉴别消息的完整性;使用一个密钥生成一个固定大小的小数据块,即 MAC,并将其加入到消息中,然后进行传输;接收方利用与发送方共享的密钥进行鉴别认证等[15]。可将 HMAC 描述如下:

$$HMAC(K,M)=H[(K^+\oplus opad)\,|\,H(K^+\oplus ipad|M)] \tag{1}$$

各符号及其定义如表 1 所列。

表 1 HMAC 中的各符号及其定义

Table 1 Symbols in HMAC and their definitions

| 符号 | 定义 |
|---|---|
| $H$ | 嵌入的散列函数(如 SHA-1,SHA-256 等) |
| $M$ | HMAC 函数的输入消息 |
| $K$ | 生成认证使用的密钥 |
| $B$ | 每一个分组长度 |
| $K^+$ | $K$ 经过左边填充几位 0,使得密钥长度达到 $b$ 位 |
| $opad$ | 0x5C 重复 $B$ 次得到的结果 |
| $ipad$ | 0x36 重复 $B$ 次得到的结果 |

以计算“$message$”的 HMAC 为例来说明 HMAC,过程如下:

1)在秘钥 $K$ 后面添加 0,创建长度为 $B$ 的字符串 $K^+$;

2)将 $K^+$ 与 $ipad$ 做异或运算,即 $K^+\oplus ipad$;

3)将 $message$ 添加到 $K^+\oplus ipad$ 后,即 $K^+\oplus ipad|message$;

4)计算 $m'=H(K^+\oplus ipad|message)$;

5)将 $K^+$ 与 $opad$ 做异或运算,即 $K^+\oplus opad$;

6)将 $m'$添加到 $K^+\oplus opad$ 后,得到 $K^+\oplus opad|m'$;

7)计算 $H(K^+\oplus opad|m')=H[(K^+\oplus opad)\,|\,H(K^+\oplus ipad|message)\,]$,输出结果。

HMAC 具有单向性和较强的抗碰撞性,其安全性依赖于嵌入的散列函数。HMAC 只需要 SHA-1 等作为嵌入散列函数就能实现较强的安全性。利用 HMAC 的单向性质,攻击者即使知晓 HMAC 的密钥 $K$,也无法逆推出原始数据。

#### 3.1.2 0-1 编码

Lin 等人[16]于 2005 年首先提出 Z-O 编码,即 0-1 编码。0-1 编码主要是为了在不暴露两个数值的前提下比较两个数值的大小。其主要过程是:分别通过两个数值的二进制表示形式得到其对应的 0-1 编码集合,然后判断两个 0-1 编码集合是否存在公共元素。

1)0-1 编码的定义

对于数值 $x$,假设二进制表示为$x=x_nx_{n-1}\cdots x_1$,$x_i\in\{0,1\}$,$i\in[1,n]$,对 $x$ 的二进制进行 0-1 编码得到的集合为:

$$S_x^0=\{x_nx_{n-1}\cdots x_{i+1}\mid x_i=0,1\leqslant i\leqslant n\} \tag{2}$$

$$S_x^1=\{x_nx_{n-1}\cdots x_{i+1}x_i\mid x_i=1,1\leqslant i\leqslant n\} \tag{3}$$

一个二进制表示的 $n$ 位数值对应的 0-1 编码具有以下性质：

① $S_x^0\cap S_x^1=\varnothing$

② $|S_0^x|\ \cap|S_x^1|=n$

对于两个数值 $x,y$，其二进制对应的 0-1 编码集合分别为 $S_x^0,S_x^1,S_y^0,S_y^1$，则：

$$S_x^1\cap S_y^0\neq\varnothing\Leftrightarrow x>y \tag{4}$$

$$S_x^1\cap S_y^0=\varnothing\Leftrightarrow x\leqslant y \tag{5}$$

例如，$x=10,y=13$，则 $x,y$ 分别表示为二进制形式 $x=1010_2,y=1101_2$（如果 $x,y$ 中某数的位数不够，那么在左边补 0，使位数一致），$S_x^1=\{1,101\},S_x^0=\{1011,11\},S_y^0=\{111\},S_y^1=\{1101,11,1\}$。由于 $S_x^1\cap S_y^0=\varnothing$，故 $x\leqslant y$，而 $S_y^1\cap S_x^0=\{11\}\neq\varnothing$，故 $y>x$。

2）0-1 编码的数值化

0-1 编码技术的最后步骤是判断两个 0-1 编码集合是否存在公共元素，而 0-1 编码集合中的元素都是由{0,1}组成的二进制串，如果直接对二进制串逐个进行比较，会导致较大的计算量，因此进一步对 0-1 编码进行数值化，即将 0-1 编码集合中的二进制串用其对应的具体数值表示，从而减少计算集合交集所需的开销。

但是直接将二进制串转换为数值会存在一定的错误，如果最高位为 0，则会被忽略，从而造成误判。例如 $x=6=0110_2,y=9=1001_2$，则 $S_x^1=\{011,01\},S_y^0=\{11,101\},S_x^1\cap S_y^0=\varnothing\Rightarrow x\leqslant y$，这与实际情况相符。而数值化后的编码集合为 $f(S_x^1)=\{1,3\},f(S_y^0)=\{3,5\},f(S_x^1)\cap f(S_y^0)\neq\varnothing\Rightarrow x>y$，$f$ 表示数值化函数，$x>y$ 与实际情况不相符。因此，在对 0-1 编码进行数值化时，不能直接将元素转换为二进制对应的数值。

本文在将 0-1 编码集合中的二进制转换为具体数值时，对集合中的二进制表示进行统一变换。因为误判是由最高位为 0 引起的，所以本文在所有的二进制元素前统一加“1”，即对于 $\forall e=x_nx_{n-1}\cdots\ x_i\in S_x^1\cup S_y^0,N(e)=f(1x_nx_{n-1}\cdots x_i)$。对于 $\forall e_1,e_2\in S_x^1\cup S_y^0$，当且仅当 $e_1=e_2$ 时，$N(e_1)=N(e_2)$；同样，如果 $e_1\neq e_2,N(e_1)\neq N(e_2)$。则式(4)与式(5)可变化为：

$$N(S_x^1)\cap N(S_y^0)\neq\varnothing\Leftrightarrow x>y \tag{6}$$

$$N(S_x^1)\cap N(S_y^0)=\varnothing\Leftrightarrow x\leqslant y \tag{7}$$

3）HMAC 数值化 0-1 编码

如果攻击者获取了某个元素对应的 0-1 编码集合，则其不难恢复出元素的值，例如包含 4 个二进制位的数值 0 编码集合 $S_x^0=\{1,011\}$，通过 0-1 编码的计算方法，可以逆推得出原始数据的二进制表示为 0101。为了避免 0-1 编码的可逆推性，采用 3.1.1 节中 HMAC 的单向性对数据进行保护。

由 HMAC 抗碰撞性可知，对于 0-1 编码集合中任意两个元素 $a_1$ 和 $a_2$，当且仅当 $a_1=a_2$ 时，才有 $HMAC_k(N(a_1))=HMAC_k(N(a_2))$；$a_1\neq a_2$ 时，$HMAC_k(N(a_1))\neq HMAC_k(N(a_2))$，因此本方案通过 HMAC 计算实现了 0-1 编码集合的隐私性保护。首先将 0-1 编码得到的两个集合进行数值化；再经过 HMAC 操作，得到两个不可逆的 HMAC 集合；然后在 HMAC 集合上判断大小关系，以防止恶意用户通过 0-1 编码集合逆推得到原始数据。式(6)与式(7)可变化为：

$$HMAC_k(N(S_x^1))\cap HMAC_k(N(S_y^0))\neq\varnothing\Leftrightarrow x>y \tag{8}$$

$$HMAC_k(N(S_x^1))\cap HMAC_k(N(S_y^0))=\varnothing\Leftrightarrow x\leqslant y \tag{9}$$

### 3.1.3　布隆过滤器

布隆过滤器（bloom filter）[17] 是由一个二进制向量和一组随机映射函数组成的数据结构，具有很好的空间和时间效率，通常被用来检测一个元素是否属于某个集合。其原理如下。

假设布隆过滤器有 $\omega$ 位，初始化时将所有的位置 0，如图 1(a)所示。将集合 $S=\{x_1,x_2,\cdots,x_n\}$ 用布隆过滤器表示，即利用 $H=\{h_1,h_2,\cdots,h_k\}(h_i(x)\in[0,\omega-1])$ 将 $S$ 中的每一个元素映射到布隆过滤器中。对于元素 $x_i$，分别将布隆过滤器中 $h_j$ 计算值对应的位置为 1。图 1(b)表示分别将元素 $x_1$ 和 $x_2$ 映射到布隆过滤器中，其中 $k=4$。

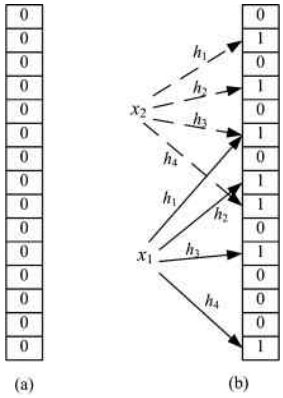


图 1　布隆过滤器的映射原理

Fig. 1　Schematic of bloom filter

如果要判断一个元素 $x$ 是否属于布隆过滤器所表示的数据集合，就利用 $H=\{h_1,h_2\cdots,h_k\}$ 分别计算 $k$ 次哈希值，并判断 $k$ 个哈希值在布隆过滤器中的对应位是否全部为 1，如果至少有一位不为 1，则元素 $x$ 不属于该布隆过滤器表示的集合，否则就认为该元素属于集合。

## 3.2　主要思想

为了秘密地对 Alice 和 Bob 的数据价值进行比较，先采用 0-1 编码分别对每种数据类型的价值进行 0-1 编码，然后根据 Alice 和 Bob 各自数据价值的 0-1 编码集合的交集判断他们各自数据价值的大小；但如果直接应用 0-1 编码，那么攻击者可以逆推出数据价值，因此本文继续使用 HMAC 对 0-1 编码的结果进行加密，并利用 HMAC 的单向性使攻击者无法逆推出 0-1 编码对应的原始数据价值；最后，通过判断 HMAC 加密后的 0-1 编码集合是否存在公共元素来确定对于每一种数据类型，Alice 的需求值和 Bob 能提供价值的大小关系。

此外，由于 0-1 编码会产生较多的元素，因此在进行匹配时产生了较大的通信开销和计算开销。本文将 0-1 编码存储在时空高效的布隆过滤器中，以进一步提高匹配过程的效率并减少通信开销。

### 3.3 详细设计

本节提出的隐私保护数据价值匹配方案主要分为两个阶段:1)双方分别对每种数据类型的数据计算数据价值对应的数据编码,将对数据价值大小的比较转变为对两个集合是否存在交集进行判断;2)利用布隆过滤器,对于每个数据类型分别判断两个集合是否存在交集,从而得到数据价值的匹配结果。

假设 Alice 和 Bob 在利用基于 PSI-CA 的隐私保护数据匹配方案完成数据类型匹配后,得到会话秘钥 $sk$ 和会话 ID $uid$ 等信息;然后假定 Alice 的每种请求数据的价值需求量为 $CA_i=\{x_1,x_2,\cdots,x_n\}$,Bob 与 Alice 匹配的 $n$ 种数据组成的数据价值集合为 $CA_j=\{y_1,y_2,\cdots,y_n\}$。Alice 和 Bob 之间的数据价值匹配过程如图 2 所示,交互过程使用的符号说明如表 2 所列。本节根据图 2 对方案进行详细描述。

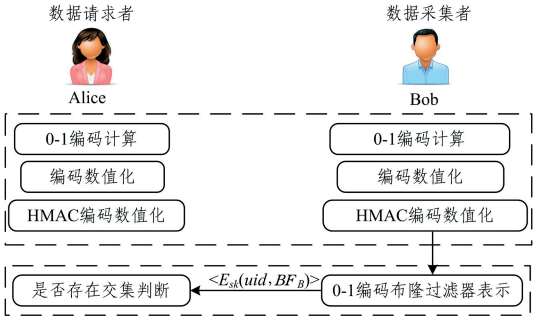


图 2 隐私保护的数据价值匹配方案框架

Fig. 2 Framework of privacy-preserving data value matching scheme

表 2 符号说明

Table 2 Notations

| 符号 | 说明 |
|---|---|
| $CA_i$ | Alice 的数据价值需求集合 |
| $CA_j$ | Bob 提供的数据价值集合 |
| $x_i$ | Alice 对 $NA_i$ 中第 $i$ 种数据的需求量 |
| $y_i$ | Alice 对 $NA_i$ 中第 $i$ 种数据的提供量 |
| $uid$ | Alice 与 Bob 的会话 $ID$ |
| $sk$ | Alice 与 Bob 的会话秘钥 |
| $BF$ | 布隆过滤器简称 |
| $k'$ | 构造 $BF$ 的 hash 函数个数 |
| $\omega'$ | $BF$ 的位数 |
| $\mathcal{F}$ | 公开的 hash 函数池 |
| $\lambda$ | 0-1 编码时统一的二进制位数 |
| $S^1_{x_i}$ | 根据 $x_i$ 得到的 1 编码集合 |
| $S^1_x$ | $CA_i$ 中所有 1 编码集合组成的集合 |
| $S^0_{y_i}$ | 根据 $y_i$ 得到的 0 编码集合 |
| $S^0_y$ | $CA_j$ 中所有 0 编码集合组成的集合 |
| $N(e)$ | 对编码集合中的元素进行数值化 |
| $HMAC_k(N(e))$ | 利用秘钥 $k$ 对数值化后的值进行 HMAC 加密 |
| $S^i_B$ | Bob 第 $i$ 种数据的价值 HMAC 数值化之后的 0 编码集合 |
| $S^i_A$ | Alice 第 $i$ 种数据的价值 HMAC 数值化之后的 0 编码集合 |
| $H_B$ | $S^i_B(1\leqslant i\leqslant n)$组成的集合 |
| $H_A$ | $S^i_A(1\leqslant i\leqslant n)$组成的集合 |
| $BF_i$ | $S^i_B(1\leqslant i\leqslant n)$生成的布隆过滤器 |
| $BF_B$ | $BF_i(1\leqslant i\leqslant n)$组成的集合 |

阶段 1 HMAC 数值化 0-1 编码

在阶段 1,交互双方分别将数据需求量和数据价值集合进行 0-1 编码,然后利用 HMAC 将 0-1 编码集合进行数值化。HMAC 数值化 0-1 编码的过程分为 3 个步骤,下面分别对 Alice 和 Bob 的执行步骤进行说明。

Alice(请求者):

步骤 1 对于 $\forall x_i\in CA_i$,Alice 根据式(3)得到 $S^1_{x_i}$,组成集合 $S^1_x=\{S^1_{x_i}\}^n_{i=1}$;

步骤 2 对于 $\forall S^1_{x_i}\in S^1_x$,$\forall e\in S^1_{x_i}$,分别计算 $N(e)$,得到集合 $N(S^1_x)$;

步骤 3 对于 $\forall N(e)\in N(S^1_x)$,分别计算 $HMAC_k(N(e))$,得到 $HMAC_k(N(S^1_x))$。

Bob(采集者):

步骤 1 对于 $\forall y_i\in CA_j$,Bob 根据式(3)得到 $S^0_{y_i}$,组成集合 $S^0_y=\{S^0_{y_i}\}^n_{i=1}$;

步骤 2 对于 $\forall S^0_{y_i}\in S^0_{y_i}$,$\forall e\in S^0_{y_i}$,分别计算 $N(e)$,得到集合 $N(S^0_y)$;

步骤 3 对于 $\forall N(e)\in N(S^0_y)$,分别计算 $HMAC_k(N(e))$,得到 $HMAC_k(N(S^0_y))$。

其中,Alice 对 1 编码集合进行 HMAC 数值化(步骤 3)和 Bob 对 0 编码集合进行 HMAC 数值化时使用的秘钥 $k=sk$。Alice 和 Bob 分别执行以上步骤之后,Alice 得到集合 $HMAC_{sk}(N(S^1_x))$,Bob 得到集合 $HMAC_{sk}(N(S^0_y))$。

阶段 2 交集判断

如果 Bob 直接将 $HMAC_{sk}(N(S^0_y))$发送给 Alice,则Alice需要分别对每种数据的需求与 Bob 对应的价值进行比较,如果全部满足条件(或者达到 Alice 预先设置的阈值),则 Bob 满足 Alice 的数据价值需求,匹配成功;反之,匹配不成功。如果对两个集合中的数据分别比较判断,则需要的开销为 $O(n\lambda^2)$,$n$ 为 Alice 需要的数据类型个数,$\lambda$ 表示对价值进行 0-1 编码时统一的 bit 长度。由 3.1.3 节布隆过滤器的性质可知,布隆过滤器可以在 O(1)的时间内判断一个元素是否属于某个集合。如果利用布隆过滤器判断是否存在公共元素,则需要的开销为 $O(n\lambda+kn\lambda)$,$n\lambda$ 表示对所有的元素判断是否属于集合的开销,$kn\lambda$ 表示为所有元素生成所有布隆过滤器的开销。另外,直接传输 0-1 编码集合的通信开销也较大。因此,本方案采用布隆过滤器来判断在每种数据类型下 Alice 的 1 编码集合和 Bob 的 0 编码集合是否存在交集。

阶段 2 主要分为以下两个步骤:

步骤 1 设 $HMAC_{sk}(N(S^0_y))=H_B=\{S^i_B\}^n_{i=1}$,$S_i$ 表示 Bob 根据第 $i$ 种数据的价值计算出的对应的 0 编码集合。Bob 利用 $uid$ 选择 $k$ 个 hash 函数,分别利用算法 1 对 $S_i$ 生成布隆过滤器 $BF_i$,得到 $BF_B=\{BF_i\}^n_{i=1}$。Bob 将 $BF_B$ 和 $uid$ 加密后发送给 Alice。

Bob→Alice:$\langle uid,BF_B\rangle$

步骤 2 Alice 收到 $uid$ 和 $BF_B$ 之后,对阶段 1 计算得到的 HMAC 数值化后的 1 编码集合 $HMAC_{sk}(N(S^1_x))=H_A=\{S^i_A\}^n_{i=1}$,采用算法 1 对每个数值对应的集合 $S^i_A$ 进行判断,如果 $S^i_A\cap S^i_B=\varnothing$,即 $S^i_A$ 中的元素都不属于 $BF_i$ 表示的集合,则 $x_i\leqslant y_i$,Bob 对第 $i$ 种数据提供的价值满足 Alice 对第 $i$ 种数据的需求值;否则,$x_i\geqslant y_i$,Bob 对第 $i$ 种数据提供的价值不满足 Alice 对第 $i$ 种数据的需求。如果对于 $\forall i\in[1,n]$,$x_i\leqslant y_i$ 均成立(或者成立的个数大于 Alice 设定的阈值 $\delta$),则 Bob 满足 Alice 的数据价值需求,匹配成功;否则匹配失败。

**算法 1**　利用 $BF$ 比较数据的大小

输入：表示元素 x 的 0-1 编码集合 S，表示元素 y 的 BF，$k'$，$\omega'$，$H=\{h_i\}_{i=1}^{k'}$

输出：如果 x ＞ y，输出 0；否则输出 1

```
1. for each t∈S //遍历集合中的元素
2.   for(i=0;i<k′−1;)
3.     j=h_i(t);
4.     if(BF[j]==1 )
5.         i++;
6.     else break; //t 不属于集合，判断下一个元素
7.   end for
8.   if(i==k′−1) //元素 t 属于集合，x>y
9.     return 0;
10.  end if
11. end for
12. return 1; //S 中的所有元素都不属于 BF，x≤y
```

## 4　性能评估

### 4.1　方案分析

本节首先对上一节提出的方案进行正确性分析，然后对方案的隐私性和复杂度进行分析与量化，并且与现有方案进行比较。

#### 4.1.1　正确性分析

本文方案的误差主要来自于采用布隆过滤器判断 Alice 的 1 编码集合 $HMAC_{sk}(N(S_x^1))$ 和 Bob 的 0 编码集合 $HMAC_{sk}(N(S_y^0))$ 是否存在交集，从而判断 Bob 是否满足 Alice的数据价值需求。下面主要从两个方面分析方案的正确性。

(1)false negative 分析

布隆过滤器由于存在一定的假阴性误判率(false negative rate)，因此可能会将没有公共元素的两个集合判定为有公共元素。假设对于数据类型 $a$，Alice 的需求量为 $x_a$，$x_a$ 对应的 1 编码集合为 $HMAC_{sk}(N(S_{x_a}^1))=\{3,5,7\}$，Bob 根据自身能提供的数据 $a$ 的价值 $y_a$，计算出 $y_a$ 对应的 0 编码集合为 $HMAC_{sk}(N(S_{y_a}^0))=\{4,6,8\}$，$S_{x_a}^1 \cap S_{y_a}^0=\varnothing$，由式(9)知，$x\leqslant y$。但是由于布隆过滤器的误判，有可能将 $S_{x_a}^1$ 中的某个元素(如 3)判定为属于 $S_{y_a}^0$，从而导致结果为 $x>y$，即对于数据类型 $a$，Bob 能够提供的数据价值不满足 Alice 需要的价值。因此，本文方案存在一定的假阴性误判率，即会将满足条件的用户判定为匹配失败，误判率与布隆过滤器的误判率相同，而布隆过滤器自身的假阴性概率为：

$$p=(p')^k=(1-e^{-nk/\omega})^k \tag{10}$$

(2)false positive 分析

为了减少计算量与通信开销，本方案采用布隆过滤器判断 Alice 的 1 编码集合 $HMAC_{sk}(N(S_x^1))$ 和 Bob 的 0 编码集合 $HMAC_{sk}(N(S_y^0))$ 是否存在交集，从而判断 Bob 是否满足 Alice 的数据价值需求。但是布隆过滤器存在一定的假阳性误判率(false positive rate)，可能会将没有公共元素的两个集合判定为有公共元素。假设对于数据类型 $a$，Alice 对数据 $a$ 的需求量为 $x_a$，计算得到的 1 编码集合为 $S_{x_a}^1=\{3,5,7\}$，Bob 能提供的价值 $y_a$ 得到的 0 编码集合为 $S_{y_a}^0=\{4,6,8\}$。显然，$S_{x_a}^1 \cap S_{y_a}^0=\varnothing$，由式(9)知，$x\leqslant y$。但是由于布隆过滤器的误判，有可能将 $S_{x_a}^1$ 中的某个元素(如 3)判定为属于 $S_{y_a}^0$，从而导致结果为 $S_{x_a}^1 \cap S_{y_a}^0=\{3\}\neq\varnothing$，即 $x>y$，也即对于数据类型 $a$，Bob 能够提供的数据价值不满足 Alice 需要的价值。因此，本文方案存在一定的假阳性误判率，即会将满足条件的用户判定为匹配失败，误判率与布隆过滤器的误判率相同。为了提高匹配的成功率，允许 Alice 设置阈值，即只要在 Bob 的 $n$ 种数据中有 $\delta$ ($\delta\leqslant n$)种的数据价值满足 Alice 的需求，则 Alice 认为 Bob 满足条件。另外，Alice 也可以重复利用布隆过滤器进行二次匹配来减小误判，即将阶段 2 重复执行一次，对于某个数据类型 $a$，只有两次都判定 $x>y$ 才认为不满足条件。另一方面，如果 $S_{x_a}^1 \cap S_{y_a}^0\neq\varnothing$，即如果 $HMAC_{sk}(N(S_{x_a}^1)) \cap HMAC_{sk}(N(S_{y_a}^0))\neq\varnothing$，假设 $S_{x_a}^1=\{3,5,7\}$，$S_{y_a}^0=\{4,6,7\}$，由于布隆过滤器一定会将属于集合的元素判定为属于该集合，因此结果一定是 $S_{x_a}^1 \cap S_{y_a}^0\neq\varnothing$，即 $HMAC_{sk}(N(S_{x_a}^1)) \cap HMAC_{sk}(N(S_{y_a}^0))=\{7\}\neq\varnothing$，也即 Bob 不满足 Alice 的匹配要求。因此，本方案判定结果是匹配成功的用户一定是真的匹配成功，即不存在 false positive。

因此，若 Alice 判定 Bob 为满足数据价值的用户，则 Bob 一定满足数据价值需求。

#### 4.1.2　隐私性分析

Alice 与 Bob 之间的数据交互只有在阶段 2 中 Bob 将 $BF_B$ 发送给 Alice 时才判断是否存在交集，而 Bob 不知道Alice 的任何数据价值信息，所以在整个方案中 Alice 的数据价值信息是安全的，因此本节主要分析 Bob 的数据价值隐私性。

本文使用香农熵[18]来量化 Alice 对 Bob 数据的不确定程度，即 Bob 数据的隐私程度。由于布隆过滤器的可恢复性，如果已知整个属性集合 $A$，对于 $\forall A'\subset A$，已知 $A'$ 利用 $H=\{h_1,h_2,\cdots,h_k\}(h_i(x)\in[0,\omega-1])$ 构造的布隆过滤器 $BF$，则通过穷举集合 $A$，对于集合 $A$ 中的每个元素都判断其是否属于 $BF$，则有可能恢复属性集合 $A'$。

如果 Alice 知道 HMAC 对 0-1 编码数值化之后的集合 $N$，则 Alice 可通过穷举 $N$ 恢复 $BF_i$ 所表示的集合，其中 $BF_i\subset BF_B$。另外，Bob 构造 $BF_i$ 的元素均属于 $N$，假设 HMAC 生成的密文长度为 $\eta$，则整个集合 $N$ 的元素个数为 $2^\eta$。因此，$BF_i$ 的每个元素 $y$ 的香农熵为 $\log_2 2^\eta$，而 $BF_i$ 的元素个数为数据价值对应的二进制长度 $\lambda$，则 Bob 每一个数据价值信息($BF_i$)的隐私程度值为 $\lambda\cdot\log_2 2^\eta=\lambda\eta$。

进一步分析，即使 Alice 恢复了 Bob 生成 $BF_i$ 的元素集合，也由于该集合中的元素都属于 HMAC 将 0-1 编码数值化之后的集合，而使得 Alice 不能直接得到原来的 0-1 编码集合进而恢复数据价值。而 HMAC 计算的安全性依赖于嵌入散列函数的抗强碰撞性，对于 $\gamma$ 位长度的散列码，其抗强碰撞性的计算代价[19]为 $2^{\gamma/2}$，因此每个数据价值的隐私度为 $\log_2 2^{\gamma/2}$；对于 Bob 的 $n$ 个数据价值，其隐私度则为 $n\cdot\log_2 2^{\gamma/2}=n\gamma/2$。

因此，Bob 的隐私性香农熵为：

$$E=\lambda\eta\cdot n\gamma/2=n\lambda\eta\gamma/2 \tag{11}$$

#### 4.1.3　复杂度分析

本节主要通过计算开销和通信开销对方案的复杂度进行分析。

假设方案中的数据价值集合中有 $n$ 个数据，每种数据元素的二进制长度为 $\lambda$，构造布隆过滤器采用 $k'$ 个 hash 函数，布

隆过滤器的长度为 $\omega'$，HMAC 开销为 $\Pi$。

(1)计算开销

在阶段 1，Alice 和 Bob 首先分别对 $n$ 个数据进行 0-1 编码，由式(2)和式(3)可知，每一个数据的计算开销为 $\lambda$，则对 $n$ 个数据分别进行 0-1 编码需要的计算开销为 $n\lambda$；然后对 0-1 编码进行 HMAC 数值化，需要的开销为 $n\Pi$。因此在阶段 1 Alice 和 Bob 的计算开销都为 $O(n(\lambda+\Pi))$。

在阶段 2，Bob 对 $n$ 个数据对应的 HMAC 数值化之后的 0-1 编码集合分别构造布隆过滤器。由 0-1 编码集合的性质可知，每个元素的 0 编码或者 1 编码最多有 $\lambda$ 个元素，因此每个数据构造布隆过滤器的开销为 $k'\lambda$，$n$ 个数据需要的计算开销为 $O(nk'\lambda)$。Alice 分别判断 $n$ 个数据是否满足价值需求，因此最大开销为 $O(nk'\lambda)$。

因此，本方案中 Alice 和 Bob 的计算开销均为 $O(n(\lambda+\Pi)+nk'\lambda)$。

(2)通信开销

本方案的通信开销主要来自于 Bob 给 Alice 发送 $BF_B$，每个数据对应的布隆过滤器的长度为 $\omega'$，因此 $n$ 个布隆过滤器的长度为 $n\omega'$，即本方案的通信开销为 $O(n\omega')$。

4.1.4 方案对比

为了说明本文方案的性能，将直接使用文献[11]、文献[12]方案的理论开销与本文方案的开销进行了对比，比较结果见表 3。其中，exp 为模幂运算，add 为加减法运算；文献[12]方案给出的是最坏情况下的计算量。

表 3 各方案开销的比较

Table 3 Comparison of overhead in each scheme

| 方案 | 计算量 | 通信量 |
|---|---|---|
| 文献[11]方案 | $2((2^{\lambda}-1)\text{exp})$ | $2^{\lambda}$ |
| 文献[12]方案 | $((2^{\lambda}-1)\text{add}+2^{\lambda}\text{exp})$ | $2\lambda(2^{\lambda}-1)$ |
| 本文方案 | $2(\lambda+\Pi)+k'\lambda$ hash | $\omega'$ |

由表 3 知，本文方案只需要对数值的二进制表示计算 0-1 编码，然后构成布隆过滤器来判断是否存在交集，同时本方案只需要发送布隆过滤器。文献[11]对于任意二进制表示为 $\lambda$ 的数字，均需要计算 $(2^{\lambda}-1)$ 次模幂运算与解密运算，而且在多次交互过程中最多需要发送 $2^{\lambda}$ 个元素组成的集合，因此该方案的计算量与通信量均高于本文方案；文献[12]在特定情况(两个大数相等时)下的复杂度较大。综上所述，本文提出的基于 0-1 编码的数据价值匹配方案具有较好的实用性。

下面将通过具体的实验对复杂度进行比较验证。

### 4.2 仿真实验

由于数据价值匹配是在数据类型匹配成功的基础上考虑的，因此在执行数据价值匹配方案之前，Alice 已经认为 Bob 为满足条件的用户。假设 Alice 需要的数据类型集合元素个数为 $n$，因此 Alice 和 Bob 需要对 $n$ 个元素进行价值比较。假设每个数据价值的二进制表示为 $\lambda=32$bit，由于本文的 HMAC 采用 SHA-1 生成密文，因此密文长度为 $\eta=20$Byte$=160$bit，HMAC 处理的块大小为 $\gamma=64$Byte$=512$bit。

本节主要从方案的计算开销和通信复杂度两个方面对数据价值匹配方案进行实验分析。实验基于 Java JDK＋Android SDK 对方案进行实现，实现后的方案主要运行于手机客户端。其中，PC 端的硬件环境为联想小新 V4000：Intel(R) Core(TM) i7-5500U CPU @ 2.4GHz 2.4GHz；8GB DDR3 内存，1TB 硬盘，Windows 10 系统。智能手机为小米 4：Qualcomm® 骁龙™ 801 四核 2.5GHz 处理器；3GB RAM；16GB ROM。

为了说明本方案的性能，将直接使用文献[11]、文献[12]方案的开销与本文方案的开销进行了对比。

4.2.1 计算开销

图 3 和图 4 给出了在相同实验参数的情况下，文献[11]、文献[12]及本文方案比较一对数值在 PC 端与手机端的计算开销。由图 3 可知，本文方案的计算开销小于文献[11]和文献[12]方案的开销。又由于本文方案只涉及 HMAC 和 hash 操作，不涉及复杂的模幂运算和大数乘法，因此随着元素位数的增加，本文方案的计算开销增长得不明显，计算开销增长率远低于涉及模幂运算的文献[11]和文献[12]方案。

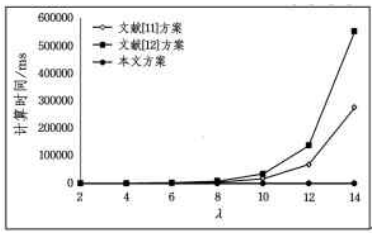


图 3 PC 端计算开销

Fig. 3 Computation overhead of PC

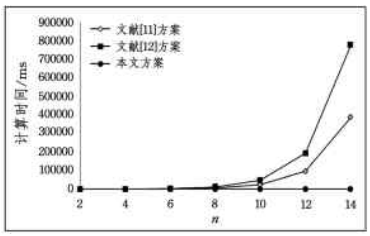


图 4 手机端的计算开销

Fig. 4 Computation overhead of smartphone

4.2.2 通信开销

图 5 给出了在相同实验参数的情况下，文献[11]、文献[12]及本方案比较一对数值的通信开销。由图 5 可知，本文方案的通信开销小于文献[11]和文献[12]方案的开销。由于本文的通信只涉及到布隆过滤器的传输，而文献[11]与文献[12]在数值比较过程中均需要发送 $2^{\lambda}$ 个整数，同时文献[12]还涉及到不经意传输，因此随着 $\lambda$ 的增大，文献[11]、文献[12]方案的通信开销增长明显；为了保持布隆过滤器的误差最小，本文布隆过滤器的位数也会随着 $\lambda$ 的增大而增长，通信开销也会相应增长，但是增长速度比文献[11]和文献[12]方案慢。

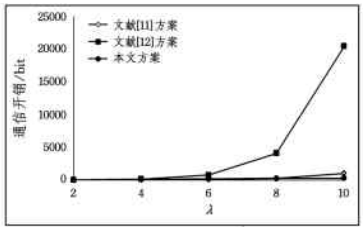


图 5 通信开销

Fig. 5 Communication overhead

分析图 3－图 5 可知，本方案的计算量与通信量都较小，同时本文的匹配结果是精确匹配，因此本文方案比文献[11]、文献[12]方案更适应用于资源有限的智能终端。

**结束语**　本文针对直接应用隐私保护集合交方案在解决参与式感知数据价值匹配问题上的不足，利用 0-1 编码方法，将数据价值大小的比较问题转换成两个集合是否存在交集的判断问题；然后进一步采用 HMAC 数值化 0-1 编码集合，并采用布隆过滤器对两个集合是否存在交集进行判断，在较好地保护用户数据价值的基础上减小了匹配过程中的计算和通信开销。理论分析和仿真实验证明了所提方案的正确性和高效性。将本文方案应用于日益蓬勃的各类移动环境下的参与式感知应用是我们后期的研究方向。

## 参 考 文 献


[1] BURKE J A，ESTRIN D，HANSEN M，et al. Participatory sensing[J]. Center for Embedded Network Sensing，2006，13(4)：117-134.

[2] AHMADI H，ABDELZAHER T，HAN J，et al. The sparse regression cube：A reliable modeling technique for open cyber-physical systems[C]// Proc. 2nd International Conference on Cyber-Physical Systems (ICCPS'11). 2011：87-96.

[3] LI H Y，ZHU H，XIAO H，et al. Location Based Participatory Sensing Service[J]. Acta Scientiarum Naturalium Universitatis Pekinensis，2014，43(2)：341-347. (in Chinese)
李怀瑜，朱瀚，肖汉，等. 基于位置的参与式感知服务[J]. 北京大学学报(自然科学版)，2014，50(2)：341-347.

[4] LIU S B，WANG Y，LIU M J. Privacy-preserving Data Sharing and Access Control in Participatory Sensing[J]. Computer Science，2015，42(6)：139-144. (in Chinese)
刘树波，王颖，刘梦君. 隐私保护的参与式感知数据分享与访问方案[J]. 计算机科学，2015，42(6)：139-144.

[5] LI Y K，LIU S B，YANG Z H，et al. Efficient and privacy－preserving profile matching protocols in opportunistic networks [J]. Journal on Communications，2015，36(12)：163-171. (in Chinese)
李永凯，刘树波，杨召唤，等. 机会网络中用户属性隐私安全的高效协作者资料匹配协议[J]. 通信学报，2015，36(12)：163-171.

[6] LIU S B，WANG Y，LIU M J，et al. Privacy-preserving various data sharing protocol in participatory sensing [J]. Journal of Computer Applications，2015，35(7)：1865-1869. (in Chinese)
刘树波，王颖，刘梦君，等. 参与式感知中隐私保护的差异化数据分享协议[J]. 计算机应用，2015，35(7)：1865-1869.

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

[8] DE CRISTOFARO E，GASTI P，TSUDIK G. Fast and private computation of cardinality of set intersection and union[M]// Cryptology and Network Security. Springer Berlin Heidelberg，2012：218-231.

[9] FREEDMAN M J，NISSIM K，PINKAS B. Efficient private matching and set intersection[M]// Advances in Cryptology-EUROCRYPT 2004. Springer Berlin Heidelberg，2004：1-19.

[10] YE Q，WANG H，PIEPRZYK J. Distributed private matching and set operations[M]// Information Security Practice and Experience. Springer Berlin Heidelberg，2008：347-360.

[11] ZHANG R，ZHANG R，SUN J，et al. Fine-grained private matching for proximity-based mobile social networking[C]// INFOCOM. IEEE，2012：1969-1977.

[12] LI H，CHENG X，LI K，et al. Efficient Customized Privacy Preserving Friend Discovery in Mobile Social Networks[C]// 2015 IEEE 35th International Conference on Distributed Computing Systems (ICDCS). IEEE，2015：225-234.

[13] NIU B，LI X，ZHU X，et al. Are You Really My Friend? Exactly Spatiotemporal Matching Scheme in Privacy-Aware Mobile Social Networks[C]// International Conference on Security and Privacy in Communication Networks. Springer International Publishing，2014：33-40.

[14] SUN J，ZHANG R，ZHANG Y. Privacy-preserving spatiotemporal matching[C]// INFOCOM. IEEE，2013：800-808.

[15] BELLARE M，ROGAWAY P. Collision-resistant hashing：Towards making UOWHFs practical [M]// Advances in Cryptology－CRYPTO'97. Springer Berlin Heidelberg，1997：470-484.

[16] LIN H Y，TZENG W G. An efficient solution to the millionaires' problem based on homomorphic encryption[M]// Applied Cryptography and Network Security. Springer Berlin Heidelberg，2005：456-466.

[17] BRODER A，MITZENMACHER M. Network applications of bloom filters：A survey[J]. Internet Mathematics，2005，1(4)：485-509.

[18] SHANNON C E. A mathematical theory of communication[J]. ACM SIGMOBILE Mobile Computing and Communications Review，2001，5(1)：3-55.

[19] WILLIAM S. Cryptography and Network Security：Principles and Practice(Fifth Edition)[M]. Beijing：Publishing House of Electronics Industry，2012. (in Chinese)
斯托林斯. 密码编码学与网络安全：原理与实践(第 5 版)[M]. 北京：电子工业出版社，2012.


(上接第 101 页)


[8] LUO L，ROY S. Efficient spectrum sensing for cognitive radio networks via joint optimization of sensing threshold and duration [J]. IEEE Transactions Wireless Communications，2012，60(10)：2851-2860.

[9] EI-SHERIF A A，MOHAMED A. Decentralized Throughput Maximization in Cognitive Radio Wireless Mesh Networks[J]. IEEE Transactions on Mobile Computing，2014，13(9)：1967-1980.

[10] GORSKI J，PFEUFFER F，KLAMROTH K. Biconvex sets and optimization with biconvex functions：asurvey and extensions [J]. Mathematical Methods of Operations Research，2007，66(3)：373-407.

[11] WENDELL R E，HUNTER A J. Minimization of non-separable objective function subject to disjoint constraints[J]. Operations Research，1976，24(3)：643-657.

[12] BOYD S，VANDENBERGHE L，FAYBUSOVICH. Convex OPTIMIZATION[J]. IEEE Transactions on Automatic Control，2016，51(11)：1859.