

参与式感知中隐私保护的差异化数据分享协议

刘树波^{1,2*}, 王颖^{1,2}, 刘梦君^{1,2}, 朱光军³

(1. 武汉大学 空天信息安全与可信计算教育部重点实验室, 武汉 430072; 2. 武汉大学 计算机学院, 武汉 430072;

3. 湖北省水利厅 信息中心, 武汉 430071)

(*通信作者电子邮箱 liu.shubo@whu.edu.cn)

摘要:参与式感知中用户不仅对数据匹配度有要求,对数据差异化也同样有要求,为了既能满足用户对数据匹配度和差异化数据的需求,也能保护用户的偏好隐私,提出了一种隐私保护的差异化数据分享协议。该协议首先将交互双方的数据表示为两个整数集合,并且利用计数布隆过滤器(CBF)计算两个集合的集合交,以集合交的结果作为数据类型匹配度;其次利用CBF能删除元素的功能,计算两个集合的差异化数据值;最后将数据类型匹配度和差异化数据值与预先设定的阈值比较,判断是否符合交互条件,同时,对CBF的构造方法进行了改进,用以保护用户的偏好隐私。理论分析和实验结果表明,与基于布隆过滤器(BF)的非加密匹配协议相比,该协议克服了匹配结果偏大的缺陷,同时计算开销减少了50%以上。该协议在保护用户偏好隐私和满足用户对差异化数据需求的同时,具有较高的匹配精度和效率。

关键词:参与式感知;差异化数据;数据匹配度;计数布隆过滤器;隐私保护

中图分类号: TP393.08; TP309.2 **文献标志码:** A

Privacy-preserving various data sharing protocol in participatory sensing

LIU Shubo^{1,2*}, WANG Ying^{1,2}, LIU Mengjun^{1,2}, ZHU Guangjun³

(1. Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education (Wuhan University), Wuhan Hubei 430072, China;

2. College of Computer, Wuhan University, Wuhan Hubei 430072, China;

3. Information Center, Hubei Provincial Water Resource Bureau, Wuhan Hubei 430071, China)

Abstract: In the process of participatory sensing, not only data matching level but also data variation is required by users. In order to meet the aforementioned two requirements, meanwhile, to protect users' preference privacy, a privacy-preserving various data sharing protocol was proposed. Firstly, both interactive data were processed to two sets of integer and Counting Bloom Filter (CBF) was utilized to calculate the intersection of the two sets of integer, the result of which was used as data matching level. Secondly, the function to delete elements of CBF was utilized to calculate the value of various data. Lastly, the data matching level and the difference between various data were compared with pre-set threshold, so as to decide whether they complied with interactive condition. In the meantime, the structuring method of CBF was improved to protect users' preference privacy. Theoretical analysis and experiment results show the following facts: compared with protocols based on non-cryptographic Bloom Filter (BF), the problem of relatively large results is overcome and computational overhead is saved by more than 50%. Users' preference privacy is protected and the need of various data is met in the proposed protocol. In addition, the proposed protocol enjoys higher matching precision and efficiency.

Key words: participatory sensing; various data; data match; Counting Bloom Filter (CBF); privacy-preserving

0 引言

参与式感知^[1]是近几年出现的一种感知技术,利用具有传感器的移动设备对各种信息进行交互式或自助式的采集、分类、传输和分析。参与式感知强调感知过程中人的参与,人们利用移动设备的各种传感器对各种数据进行采集,一个人感知的信息或者群体感知的信息可以被其他人或者群体使用,从而实现数据的广泛采集和共享使用^[2]。

参与式感知的感知主体是一个个具有思想的人,而个人的安全与隐私是每个用户在加入参与式感知时必定会考虑的

问题。目前关于参与式感知中用户的安全与隐私的研究主要集中在用户与服务器的交互过程中^[3-4]。随着WiFi (Wireless Fidelity)等近场通信(Near Field Communication, NFC)技术的发展,用户与用户之间的数据共享应用将越来越广泛。而用户之间进行数据共享过程中的安全与隐私问题还鲜有关注。

参与式感知中,用户之间会进行频繁的数据共享^[5],出于个人的安全与隐私考虑,用户在向周围用户发送数据请求时,不希望自己需要的数据类型被交互双方之外的人知晓,并且用户总是期望通过单次交互就能获得全部所需的数据类

收稿日期:2015-01-23;**修回日期:**2015-03-23。 **基金项目:**国家973计划项目(2011CB302306);中央高校基本科研业务费专项(211-274230);国家自然科学基金资助项目(41371402);水利部“948”项目(201044);湖北省水利厅农村饮用水水资源远程监控项目(211-230912)。

作者简介:刘树波(1970-),男,湖北武汉人,教授,博士生导师,CCF会员,主要研究方向:信息安全、物联网、嵌入式系统;王颖(1991-),女,湖北武汉人,硕士研究生,主要研究方向:信息安全、嵌入式系统;刘梦君(1988-),男,湖北武汉人,博士研究生,主要研究方向:移动计算与无线网络、移动社交与分布式系统中的安全及隐私;朱光军(1968-),男,湖北武汉人,教授级高级工程师,主要研究方向:水利信息化。

型,同时希望获得的差异化数据价值能够满足需求。在实际应用中用户对差异化数据的价值需求不尽相同,如供水质量监控的用户和普通用户对水质数据就有差异化要求,负责供水质量监控的用户需要水质标准中几乎全部的108种数据^[6],需要长时间的大量测量数据;而普通用户只关心当前水质是否达标或少量常规数据,需要的是实时数据少量数据。这两类用户由于关心的内容不同,所以各种数据对他们的价值也不同。不同类型的用户可以根据提供方数据对他们的价值不同,进行取舍。可以看出,仅仅只是保证获取全部数据类型并不能保证用户对差异化数据的不同需求。

本文考虑用户对差异化数据的不同需求,通过计数布隆过滤器(Counting Bloom Filter, CBF)实现数据价值的计算,这个数据价值的值就是用来衡量用户对数据的差异化需求。本文的数据分享协议既保护用户对数据的偏好隐私,又实现了不同用户对数据的差异化需求。

1 相关研究

与本文相关的研究主要是隐私保护的集合交,根据使用的数学理论,隐私保护的集合交计算方法主要分为:基于交换加密的匹配协议、基于线性多项式的匹配协议和基于伪随机函数的匹配协议。

1.1 基于交换加密的匹配协议

Agrawal等^[7]提出了一种可交换加密协议用于解决PSI(Private Set Intersection)和PCSI(Private Cardinality of Set Intersection)问题,实现了两个数据集中的交集运算。该协议的安全性依赖于DDH(Decisional Diffie-Hellman)假设,但对受到恶意攻击的情况没有考虑。

在文献[7]的基础上,Xie等^[8]提出了一种移动社交网络中的匹配协议,能够抵御一定恶意攻击。该协议计算量较大,占用资源较多。

1.2 基于线性多项式的匹配协议

Freedman等^[9]提出了一种基于多项式估值和加法同态加密的协议——FNP(Freedman-Nissim-Pinkas)协议。该协议通过将数据集中的数据作为多项式的根构造出一个多项式,并对多项式系数同态加密。该协议复杂度低,适用于半诚实模型,但对恶意攻击抵御能力较弱。

为将文献[9]协议应用到分布式环境中,Ye等^[10]把数据方集合用一个多项式表示,然后分发多项式系数到多个服务器,实现密钥分享这种分布式协议不适合在参与式感知环境中应用。

另外,在双线性映射函数基础上,Lu等^[11]提出了一个双线性映射匹配算法,并且运用到了疾病监控的具体案例中,使具有相同病症的人可以分享信息。该算法只适用于匹配一个属性的场景,难以扩充到多属性的应用中。

1.3 基于伪随机函数的匹配协议

Yang等^[12]设计了一种分布式手机社交网络系统:E-SmallTalker。运用布隆过滤器(Bloom Filter, BF)作为属性存储结构,通过伪随机函数进行多轮迭代映射计算交集,可以有效减少存储空间和避免对方知道共同属性以外的其他信息。然而,由于布隆过滤器不能按需增删元素,若要改变元素集合,只能重置布隆过滤器,因此会增加额外的工作量。

Sun等^[13]提出了两种计算集合交的方法:一种是基于PSI的加密方法;一种是基于布隆过滤器的非加密方法。加密方法通过适应性量化技术,将用户的每个元素对应到一个单元索引,计算集合交的双方通过PSI计算公共元素。这种方法的计算量和通信量都较大,不适用于资源有限的移动终端。非加密方式通过对布隆过滤器进行改进,双方分别采用不同的方式计算各元素对应的布隆过滤器,再计算公共元素,这种方法的计算量和通信量都较小,但计算结果存在一定误差。

以上三类关于集合交的办法,都只考虑了是否存在的问题,而没有考虑存在多少的问题,如集合 $C = \{a, b\}$, $B = \{a, b, c, d, a, b, a\}$,目前研究得到的是公共属性集合 $\{a, b\}$ 或者判断 C 是 B 的子集,但是都不能知道 B 中含有 C 中元素 a 和 b 的具体数目。本文提出一种新型数据分享协议,使请求者在数据分享过程中不仅能判断采集者是否拥有自己需要的数据类型,还能判断采集者对每一类型数据的拥有量,即数据的价值,同时在数据分享过程中保护双方对数据的偏好隐私。

2 模型与假设

2.1 系统模型

如图1所示,本文的系统模型主要由数据采集者和数据请求者构成,用户既可以是数据采集者也可以是数据请求者。图1中的数据采集者利用自己智能终端的传感器采集数据。数据请求者由于业务或者其他需求,需要获取一定价值的数

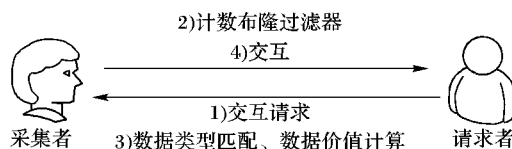


图1 系统模型

2.2 安全模型

假设用户是理性、诚实而好奇的。理性是指获得满足自己所需数据时才会付出相应代价;诚实而好奇是指每个用户都希望隐藏自己的偏好隐私,但是却希望知道其他用户的偏好隐私,同时,用户总是希望通过单次交互就能获取全部所需数据。除此之外,用户的操作都会遵循系统的要求。

2.3 问题描述

假设系统包括 m 个用户,分别表示为 U_1, U_2, \dots, U_m ,系统一共有 n 种数据,数据类型对应为一个固定长度的集合 $A = \{A_1, A_2, \dots, A_n\}$ 。每个用户既可以是数据使用者也可以是数据提供者,因此每个用户都拥有两个集合 NA_i 和 PA_i 。 NA_i 表示用户需要的数据类型对应的集合, PA_i 表示用户拥有的数据类型对应的集合。假设用户之间是通过WiFi/Bluetooth等近场通信技术通信,并且任意两个用户之间都建立了一个可信的通信通道。

进一步定义两个用户 U_i (Alice)和 U_j (Bob),Alice向Bob发出数据交易请求,假设Alice需要的数据类型对应的集合为 $NA_i = \{A_{i1}, A_{i2}, \dots, A_{im}\}$,其中Alice需要的数据类型的相应位为1,其他位为0。Bob拥有的数据类型对应的集合为 $PA_j =$

$\{N_{i1}, N_{i2}, \dots, N_{in}\}$, Bob 每次获取一种类型的数据, 就将相应数据类型的值加 1, A_{ij} 表示 Bob 拥有第 j 类数据的数目。由于 Alice 希望能直接从一个数据采集者处获得全部所需的数据类型并且希望获取数据的价值能满足自己的需求, 因此 Alice 需要将 Bob 拥有的数据对应的数据类型与自己需要的数据类型进行匹配, 另外, Alice 还会对 Bob 拥有的数据的价值进行计算判断是否达到要求(如, Alice 需要第 3 种数据 100 个, 第 7 种数据 30 个等)。如果 Bob 拥有的数据类型满足 Alice 需要的数据类型且 Bob 拥有的数据价值达到了 Alice 的要求, 则 Alice 与 Bob 进行交互, 获得想要得到的数据, 如果 Bob 没有 Alice 想要的全部的数据类型或者数据的价值不满足需求, 则 Alice 终止与 Bob 的交互, 并继续与其他用户进行上述相同过程直到找到一个满足交互条件的用户, 然后 Alice 与满足交互条件的用户进行交互。

3 隐私保护的数据分享协议

3.1 设计思想

由于整个过程是在能量有限的手机端进行操作, 因此需要一个计算量小、既能计算数据类型是否匹配又能计算数据价值的方法, 同时请求者在交互过程中不希望不满足条件的采集者知道自己对数据的需求, 而采集者也不希望将自己的数据细节暴露给任何请求者, 因此在交互过程中还需满足双方对数据的偏好隐私。本文采用计数布隆过滤器^[17]实现用户对数据类型和数据价值的计算, 计数布隆过滤器将标准的布隆过滤器^[18]的每一位扩展为一个计数器, 这个计数器恰恰可以用来衡量不同数据的价值, 同时对计数布隆过滤器的构造过程进行改进, 保护交互双方对数据的偏好隐私。

3.2 相关知识

计数布隆过滤器(CBF)由标准布隆过滤器扩展而来, 它将标准布隆过滤器的每一位扩展为一个小的计数器(Counter), 如图 2 所示。在插入元素时给对应的 k (k 为哈希函数的个数) 个 Counter 的值分别加 1, 删除元素时将对应的 k 个 Counter 的值分别减 1。计数布隆过滤器通过多占用几倍存储空间为代价, 在标准布隆过滤器的基础上增加了删除元素的功能^[19], 通过这个删除功能计算元素集合里每个元素的个数, 而数据个数作为数据价值的衡量标准。

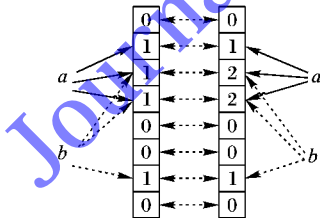


图2 计数布隆过滤器工作原理

根据 CBF 原理, 可将 CBF 记为 $CBF = \langle H, V \rangle$, 其中: V 为 n 维的整数向量, 记为 $V = \langle v_1, v_2, \dots, v_n \rangle$, 并且 $\forall v_i \in V, v_i \in [0, 2^m]$, m 为 V 在每个维度所占的位数; H 为 k 个哈希函数的集合, 记为 $H = \{h_i | 0 \leq i < k, 0 \leq h_i < d\}$ 。

在图 2 的基础上, 假设现在拥有数据 a 的数量为 10, 数据 b 的数量为 5, 则得到的计数布隆过滤器向量为:

0	10	15	15	0	0	5	0
---	----	----	----	---	---	---	---

如果将数据 a 的数量删除无效的 2 个, 得到的计数布隆

过滤器向量为:

0	8	13	13	0	0	5	0
---	---	----	----	---	---	---	---

在计算数据价值时, 对 a 分别利用 k (此处为 3) 个 Hash 函数进行计算, 找到计数布隆过滤器向量中对应数值最小的位, 以该位的数值作为 a 的数据价值, 在此处 a 的值为 8。然后将对应的 k 位分别减去 8, 同样的方法计算得到 b 的数据价值为 5。

3.3 协议描述

假设 \mathcal{F} 为一个很大的公开的 Hash 函数池。选择 Hash 函数 $H: \{0, 1\}^* \rightarrow \mathbf{Z}^*$ 。

1) Bob 随机选择整数 z , 利用 $H(z)$ 在 Hash 池 \mathcal{F} 里面选择 k 个 Hash 函数, 从得到的 k 个 Hash 函数里面选择 l 个 ($l < k$), 同时选取 $k - l$ 个 $H(z)$ 对应的 Hash 函数之外的 Hash 函数, 得到 k 个 Hash 函数集合 $Hash_Bob[1..k]$, 通过选取的 k 个 Hash 函数, 计算自己拥有属性集合 $PA_j = \{N_{j1}, N_{j2}, \dots, N_{jn}\}$ 对应的 ω 位计数布隆过滤器 (CBF_B), N_{jx} 表示拥有的第 x 种数据的个数。构造算法如下:

```

for (y = 1 to n)
  if ( $A_{jy} \neq 0$ )
    for (x = 1 to k)
       $Hash\_Bob[x](y) = position$ 
       $CBF_B[position] += A_{jy}$ 
    end for
  end for
end for

```

2) Bob 将自己的 CBF_B 以及 $H(z)$ 发送给 Alice。Alice 利用 $H(z)$ 在 Hash 池 \mathcal{F} 里面选择 k 个 Hash 函数, 得到 k 个 Hash 函数集合 $Hash_Alice[1..k]$ 。

3) Alice 需要的数据对应的集合 $NA_i = \{A_{i1}, A_{i2}, \dots, A_{in}\}$, 其中 A_{in} 表示 Alice 需要的第 x 种数据个数。Alice 对数据要求的权值集合为 $value[n] = \{v_{i1}, v_{i2}, \dots, v_{in}\}$, Alice 计算匹配度和价值的过程如下:

将 $min[n]$ 中的元素都赋值为 10000

```

for (y = 1 to n)
  if ( $A_{iy} \neq 0$ )
    for (x = 1 to k)
       $Hash\_Alice[x](y) = position$ 
      if ( $CBF_B[position] < min[y]$ )
         $min[y] = CBF_B[position]$ 
      if ( $CBF_B[position] == 0$ )
        num ++
      end for
    end for
  end for

```

将对应的 k 位分别减去 $min[y]$

end for

将 num 与 τ_1 进行比较, 若 $num > \tau_1$, 算法结束, 不符合交互要求

for ($i = 1$ to n)

$value += min[i] * value[i]$

end for

将 $value$ 和 τ_2 进行比较, 若 $value < \tau_2$, 不符合交互要求, 算法结束; 否则, 符合交互要求, 交互继续。

4) 如果符合交互要求, 则 Alice 认为 Bob 是满足交互, 双方通过安全信道进行交互。

4 协议分析

4.1 偏好隐私保护

在匹配成功之前, 数据采集者对用户需要的数据类型是

完全未知的。数据采集者选择 $k-l$ 个不在 \mathbb{F} 里面的 Hash 函数,并将这 $k-l$ 个 Hash 函数对用户保密,在数据采集者将 $H(z)$ 和 CBF_B 发送给用户之后,用户不能直接计算出数据采集者拥有的数据类型^[18]。数据采集者的偏好隐私是保密的。

4.2 计数布隆过滤器位数的选择

参与式感知中,数据的采集者数目较多^[2],为了防止数据过多,对数据的新鲜度有一定的要求,假设用户采集的数据有效时间是 3 d,用户每天采集的次数不超过 20 次,同一种类型的数据最大的数量为 60,而 $2^6 = 64 > 60$,因此,存储一种数据数量的位数选择为 6 位。由于不同的输入,Hash 可能会得到相同的输出,因此还需考虑由于其他数据类型造成的 Counter 增加。

第 i 个 Counter 被其他数据类型增加 j 次的概率为:

$$P(c(i) = j) = \binom{nk}{j} (1/m)^j (1 - 1/m)^{nk-j} \quad (1)$$

第 i 个 Counter 的值大于 j 的概率可以限定为:

$$P(c(i) \geq j) \leq \binom{nk}{j} (1/m)^j \leq m \binom{enk}{jm}^j \quad (2)$$

假设每个元素都是等概率地 Hash 到 ω 位的布隆过滤器的任何一位,与其他元素被 Hash 到哪个位无关,则对某一特定位,在一个元素由某个特定 Hash 函数插入时没有被置位为 1 的概率为:

$$1 - 1/\omega \quad (3)$$

则 k 个 Hash 函数全没有将其置位为 1 的概率为:

$$(1 - 1/\omega)^k \quad (4)$$

对于插入 n 个元素的集合,布隆过滤器中任意一位为 0 的概率为:

$$p = (1 - 1/\omega)^{nk} = e^{-nk/\omega} \quad (5)$$

将不属于集合中的元素误判为属于集合中的元素时,布隆向量所对应的 k 个位置必须全部为 1,则误判率(False Positive Rate, FPR) 为:

$$FPR = (1 - p)^k = (1 - e^{-nk/\omega})^k \quad (6)$$

当 m, ω 一定时,对式(6)求导,可知当 $e^{-nk/\omega} = 1/2$,即 $k \leq \ln 2 \times \omega/n \approx 0.7 \times \omega/n$ 时,布隆过滤器的误判率最低。假设 $k \leq \ln 2 \times \omega/n$,则由式(2)得:

$$P(\max c(i) \geq j) \leq m ((e \ln 2)/j)^j \quad (7)$$

如果每个 Counter 分配 4 位,则 Counter = $2^4 = 16$ 溢出的概率为:

$$P(\max c(i) \geq 16) \leq 1.37 \times 10^{-15} \times m$$

这个值足够小,对于本系统足够,因此 Counter 的位数为 $4 + 6 = 10$ 位。

4.3 数据价值计算的准确性

本方案只有匹配成功才需要考虑数据的价值是否满足条件,因此对数据价值的计算是建立在已经判定匹配成功的基础上。

在 CBF_B 中加入一个元素时, k 个哈希位置的 Counter 都要加 1。也就是说,如果不考虑碰撞,出现次数为 n 的元素对应的 k 个 Counter 的值都为 n 。即使考虑到碰撞的因素,只要 k 个位置不全出现碰撞, k 个 Counter 中的最小值仍是 n 。令元素 x 对应的 k 个 Counter 的最小值为 m_x , x 的出现频率为 f_x ,从上面的分析可知, $f_x \neq m_x$ 的概率和标准布隆过滤器的误判率相同,因为二者出现的充要条件都是 k 个哈希位置同时出现碰

撞^[20]。

在本方案中,每一次对某个元素 A_{ij} 进行 k 次 Hash,判断每种数据对应的价值时,都是选择 CBF_B 对应 k 位的最小值,而 Alice 计算得到的价值是和预先选定的价值阈值相比较,因此通过选择合适的系统参数和阈值,能使数据价值计算的准确性达到系统要求。

4.4 系统匹配误差阈值的选择

由 4.2 节可知,布隆过滤器的误判概率在 $k = \ln 2 \times \omega/n$ 时最小为 $FPR = (1 - 1/2)^k = 2^{-k} = 2^{-\ln 2 \times \omega/n} = 0.6185^{\omega/n}$ 。

匹配过程中的误差除了布隆过滤器的误判概率之外,主要是因为交互双方选择的 Hash 函数不完全相同,在此主要讨论在保证误判率最低的情况下 Hash 函数不同个数 $t = (k-l)$ 对匹配精度的影响。

为了便于存储,设置 $\omega = 2^{11} = 2048$, $n = 100$,则当 $\ln 2 \times \omega/n \approx 0.7 \times \omega/n \approx 14$,取 $k = 14$,此时布隆过滤器的误判率最低。

Hash 函数不同个数 $(k-l)$ 对匹配精度和数据价值的影响分别如图 3 和图 4 所示。

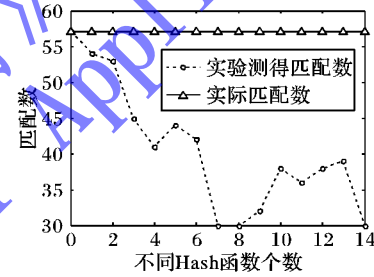


图3 Hash 函数不同个数 $(k-l)$ 对匹配精度的影响

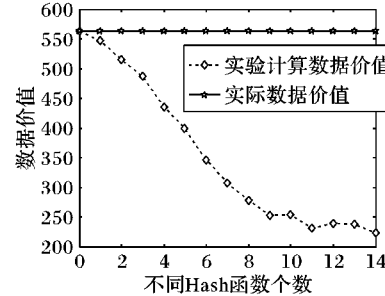


图4 Hash 函数不同个数 $(k-l)$ 对数据价值精度的影响

实验过程中的 $PB_j = \{A_{j1}, A_{j2}, \dots, A_{jn}\}$ 是利用随机函数生成的 $[0, 60]$ 区间的随机数。 $NA_i = \{A_{i1}, A_{i2}, \dots, A_{in}\}$ 是利用随机函数生成的随机 0,1。数据价值的权重是利用随机函数生成的 $[0, 1)$ 区间的随机数。

由图 3 和图 4 可知,当 Hash 函数不同个数 $t = (k-l)$ 在增大时,实验的匹配个数和实际价值都在减少,但两者不呈现线性关系,由于不同的输入,经过相同的 Hash 函数计算可能会有相同的输出,而不同的输入、不同的 Hash 函数计算也可能会有相同的输出,因此当 $t = k$ 时,匹配个数也不会减小到 0。

根据图 3 和图 4 可知,当 k 一定时,通过选取合理的 l 和阈值能够使匹配精度和数据价值精度都保持在满足要求的范围内。如, $k = 14$,取 $l = 12$,此时匹配个数为 53(实际匹配个数为 56),数据价值为 520(实际数据价值为 563),而此时文献[13]利用同样的方法和 Hash 函数得到的匹配个数是 64,匹配个数比实际的多,增大了错判率。在文献[13]中作者也说

明了实验得到的匹配结果比实际的匹配数大。在实际的应用中,如果协议计算的结果为匹配的数目比需要的多则结论一定是匹配成功,当这种“多”是由于误差引起时,实际的误判率会变大,到交易阶段用户才能知道实际是否匹配,如果不匹配会消耗更多的资源。针对本文的结果,此时若选取允许 CBF_B 为 0 的个数,即误差阈值 $\tau_1 = 5$, 数据最小价值 $\tau_2 = 500$, 则能保证数据分享成功且保护用户对数据的偏好隐私。实际应用中,在可接受的误差范围内,合理地选择 Hash 函数及阈值 τ_1 和 τ_2 , 能使匹配精度和数据价值误差满足应用要求,使得用户通过单次分享就能获取满足需求的数据类型和数据量,同时保护用户对数据的偏好隐私。

4.5 性能分析

整个协议中的数据类型匹配和数据价值计算都是基于计数布隆过滤器的,没有采用复杂的加解密操作,计算量小。用户只需要计算 nk 个 Hash 函数操作和简单的整数加减法,加减法对协议的计算复杂度影响可忽略,因此整个协议的计算复杂度为 $O(kn)$ 。与第 2 章中相关研究相比,本文的计算复杂度较小。

文献[10]涉及到服务器的交互,与本文协议没有可比性,文献[11]只能用于特定环境,因此也不予以比较。对比如表 1 所示。

表 1 各种方案的计算复杂度对比

方案	计算复杂度	方案	计算复杂度
文献[7] 方案	$O(2C(i+j))$	文献[12] 方案	$O(kn)$
文献[8] 方案	$O(2C(i+j) + 2D)$	文献[13] 方案	$O(2(kn + \omega))$
文献[9] 方案	$O(E(i+j))$	本文方案	$O(kn)$

表 1 中, i, j 分别为交互双方拥有的属性数目, C 是利用 $f_e(x) = x^e \bmod p$ 加解密的开销, D 为文献[8] 涉及的 Diffie-Hellman 计算开销, E 为文献[9] 涉及到的加解密的计算开销。文献[7-11] 都只实现了计算类型匹配度;没有实现对数据价值的计算,并且计算开销都比本文的高。文献[12] 采用的是布隆过滤器因此与本文开销一样;但是该方案只能计算数据类型匹配度,不能计算数据价值,同时本协议保护了交易双方的数据隐私。文献[13] 也是对布隆过滤器进行改造,保护了交易双方的数据隐私;但是该方案除了构造布隆过滤器的开销,还需要对两个布隆过滤器进行遍历,当布隆过滤器位数增多时,该开销会显著增加。

5 结语

在参与式感知的用户之间的交互过程中,为了实现既能保护用户隐私,又能使用户仅通过单次交互便能获得全部所需数据类型且获得的数据满足价值需求,本文将用户对数据价值的差异化需求考虑进来,采用计数布隆过滤器,使用户不仅能计算数据类型的匹配度,也能判断差异化数据价值是否符合需求,同时也保护用户对数据的偏好隐私。分析表明协议既能够隐私保护的计算数据类型的匹配度,也能对数据价值进行计算,为用户提供差异化服务。接下来将对布隆过滤器的构造方法进行进一步的研究与改进,将数据类型匹配和价值计算误差进一步减小。

参考文献:

- [1] BURKE J A, ESTRIN D, HANSEN M, *et al.* Participatory sensing [C]// WSW'06: Proceedings of the 1st Workshop on World-Sensor-Web. New York: ACM, 2006: 117-134.
- [2] MUN M, REDDY S, SHILTON K, *et al.* PEIR, the personal environmental impact report, as a platform for participatory sensing systems research [C]// Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services. New York: ACM, 2009: 55-68.
- [3] ZHANG J, MA J, WANG W, *et al.* A novel privacy protection scheme for participatory sensing with incentives [C]// CCIS 2012: Proceedings of the 2012 IEEE 2nd International Conference on Cloud Computing and Intelligent Systems. Piscataway: IEEE, 2012: 1017-1021.
- [4] AHMADI H, PHAM N, GANTI R, *et al.* Privacy-aware regression modeling of participatory sensing data [C]// Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems. New York: ACM, 2010: 99-112.
- [5] LEE J S, HOH B. Sell your experiences: a market mechanism based incentive for participatory sensing [C]// PerCom 2010: Proceedings of the 2010 IEEE International Conference on Pervasive Computing and Communications. Piscataway: IEEE, 2010: 60-68.
- [6] The Ministry of Health in China. GB5749-2006 Drinking water standards [S]. Beijing: Standards Press of China, 2006: 1-8. (中华人民共和国卫生部. GB5749—2006 生活饮用水卫生标准[S]. 北京: 中国标准出版社, 2006: 1-8.)
- [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. New York: ACM, 2003: 86-97.
- [8] XIE Q, HENGARTNER U. Privacy-preserving matchmaking for mobile social networking secure against malicious users [C]// PST 2011: Proceedings of the 2011 Ninth Annual International Conference on Privacy, Security and Trust. Piscataway: IEEE, 2011: 252-259.
- [9] FREEDMAN M J, NISSIM K, PINKAS B. Efficient private matching and set intersection [C]// Advances in Cryptology—EUROCRYPT 2004. Berlin: Springer, 2004: 1-19.
- [10] YE Q, WANG H, PIEPRZYK J. Distributed private matching and set operations [C]// ISPEC 2008: Proceedings of the 4th International Conference on Information Security Practice and Experience, LNCS 4991. Berlin: Springer, 2008: 347-360.
- [11] LU R, LIN X, LIANG X, *et al.* Secure handshake with symptoms-matching: the essential to the success of mhealthcare social network [C]// Proceedings of the Fifth International Conference on Body Area Networks. New York: ACM, 2010: 8-15.
- [12] YANG Z, ZHANG B, DAI J, *et al.* E-SmallTalker: a distributed mobile system for social networking in physical proximity [C]// ICDCS 2010: Proceedings of the 2010 IEEE 30th International Conference on Distributed Computing Systems. Piscataway: IEEE, 2010: 468-477.
- [13] SUN J, ZHANG R, ZHANG Y. Privacy-preserving spatiotemporal matching [C]// INFOCOM 2013: Proceedings of the 2013 IEEE Conference on Computer Communications. Piscataway: IEEE, 2013: 800-808.

(下转第 1896 页)

窗口大小设置为5 min。系统检测准确率为99.7%,而误报率仅为0.3%。

4 结语

本文通过分析P2P网络和P2P僵尸网络的基本特性,提出了3个用于检测P2P主机的统计特征(节点链接不稳定性、节点发现模式、远端IP分布广度)和两个用于检测P2P僵尸主机的统计特征(通信报文占比和远端IP回访数)。在此基础上,本文设计并实现了一个P2P僵尸主机实时检测系统。相比已有的系统,本文系统具有以下特点:1)不需要检测报文负载,能检测加密流量;2)不依赖于僵尸主机的恶意活动,能检测处于隐匿阶段的P2P僵尸主机;3)所用特征少,时间窗口小,检测精度高,实时性强。实验结果证明,本文系统能在5 min内检测出监控网络内的所有P2P僵尸主机,检测准确率达到99.7%,误报率仅为0.3%。实验结果有效地证明了本文系统在准确率和实时性上的优越性。

参考文献:

- [1] LIU L, CHEN S, YAN G, *et al.* Bot Tracer: execution-based bot-like malware detection [M]// ISC'08: Proceedings of the 11th International Conference on Information Security, LNCS 5222. Berlin: Springer, 2008: 97–113.
- [2] SZYMCHYK M. Detecting botnets in computer networks using multi-Agent technology [C]// DepCos-RELCOMEX'09: Proceedings of the Fourth International Conference on Dependability of Computer Systems. Piscataway: IEEE, 2009: 192–201.
- [3] STINSON E, MITCHELL J C. Characterizing bots' remote control behavior [C]// DIMVA'07: Proceedings of the 4th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin: Springer, 2007: 89–108.
- [4] XU K, YAO D, MA Q, *et al.* Detecting infection onset with behavior-based policies [C]// NSS 2011: Proceedings of the 5th International Conference on Network and System Security. Piscataway: IEEE, 2011: 57–64.
- [5] GU G, PORRAS P, YEGNESWARAN V, *et al.* BotHunter: detecting malware infection through IDS-driven dialog correlation [C]// SS'07: Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium. Berkeley: Usenix Security, 2007: 1–16.
- [6] SINGH K, GUNTUKU S C, THAKUR A, *et al.* Big data analytics framework for peer-to-peer botnet detection using random forests [J]. Information Sciences, 2014, 278: 488–497.
- [7] JIANG H, SHAO X. Detecting P2P botnets by discovering flow dependency in C&C traffic [J]. Peer-to-Peer Networking and Applications, 2014, 7(4): 320–331.
- [8] SILVA S S C, SILVA R M P, PINTO R C G, *et al.* Botnets: a survey [J]. Computer Networks, 2013, 57(2): 378–403.
- [9] YU X, DONG X, YU G, *et al.* Online botnet detection based on incremental discrete Fourier transform [J]. Journal of Networks, 2010, 5(5): 568–576.
- [10] ZHANG J, PERDISCI R, LEE W, *et al.* Building a scalable system for stealthy P2P-botnet detection [J]. IEEE Transactions on Information Forensics and Security, 2014, 9(1): 27–38.
- [11] EN T F, REITER M K. Are your hosts trading or plotting? Telling P2P file-sharing and bots apart [C]// ICDCS 2010: Proceedings of the 2010 IEEE 30th International Conference on Distributed Computing Systems. Piscataway: IEEE, 2010: 241–252.
- [12] RAHBARINIA B, PERDISCI R, LANZI A, *et al.* PeerRush: mining for unwanted P2P traffic [J]. Journal of Information Security and Applications, 2014, 19(3): 194–208.
- [13] ZHAO D, TRAORE I, GHORBANI A, *et al.* Peer to peer botnet detection based on flow intervals [C]// SEC 2012: Proceedings of the 27th IFIP TC 11 Information Security and Privacy Conference on Information Security and Privacy Research. Berlin: Springer, 2012: 87–102.
- [14] FINSTERBUSCH M, RICHTER C, ROCHA E, *et al.* A survey of payload-based traffic classification approaches [J]. IEEE Communications Surveys & Tutorials, 2014, 16(2): 1135–1156.
- [15] LIU C, YANG Y, TANG C. A classification method of unstructured P2P multicast video streaming based on SVM [C]// MINES'09: Proceedings of the 2009 International Conference on Multimedia Information Networking and Security. Piscataway: IEEE, 2009: 68–72.
- [16] HE J, YANG Y, QIAO Y, *et al.* Accurate classification of P2P traffic by clustering flows [J]. China Communications, 2013, 10(11): 42–51.
- [17] HALL M, FRANK E, HOLMES G, *et al.* The WEKA data mining software: an update [J]. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10–18.
- [18] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection [C]// IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1995: 1137–1145.
- [14] ZHU X. Security and privacy preservation mechanisms in vehicular Ad Hoc network [D]. Hefei: Hefei University of Technology, 2013. (朱晓玲. VANET 安全和隐私保护机制研究[D].合肥:合肥工业大学, 2013.)
- [15] SWEENEY L. *k*-anonymity: a model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557–570.
- [16] CAMENISCH J, GROTH J. Group signatures: better efficiency and new theoretical aspects [C]// Proceedings of the 4th International Conference on Security in Communication Networks, LNCS 3352. Berlin: Springer, 2005: 120–133.
- [17] FAN L, CAO P, ALMEIDA J, *et al.* Summary cache: a scalable wide-area Web cache sharing protocol [J]. IEEE/ACM Transactions on Networking, 2000, 8(3): 281–293.
- [18] BLOOM B H. Space/time trade-offs in Hash coding with allowable errors [J]. Communications of the ACM, 1970, 13(7): 422–426.
- [19] BONOMI F, MITZENMACHER M, PANIGRAHY R, *et al.* An improved construction for counting bloom filters [M] // ESA 2006: Proceedings of the 14th Annual European Symposium on Algorithms, LNCS 4168. Berlin: Springer, 2006: 684–695.
- [20] GUO D, LIU Y, LI X, *et al.* False negative problem of counting bloom filter [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(5): 651–664.

(上接第1869页)