Priority-Aware Private Matching Schemes for Proximity-Based Mobile Social Networks

Ben Niu, Tanran Zhang, Xiaoyan Zhu, Hui Li and Zongqing Lu

Abstract—The rapid developments of mobile devices and online social networks have resulted in increasing attention to Mobile Social Networking (MSN). The explosive growth of mobile-connected and location-aware devices makes it possible and meaningful to do the Proximity-based Mobile Social Networks (PMSNs). Users can discover and make new social interactions easily with physical-proximate mobile users through WiFi/Bluetooth interfaces embedded in their smartphones. However, users enjoy these conveniences at the cost of their growing privacy concerns. To address this problem, we propose a suit of priority-aware private matching schemes to privately match the similarity with potential friends in the vicinity. Unlike most existing work, our proposed priority-aware matching scheme (*P-match*) achieves the privacy goal by combining the commutative encryption function and the Tanimoto similarity coefficient which considers both the number of common attributes between users as well as the corresponding priorities on each common attribute. Further, based on the newly constructed similarity function which takes the ratio of attributes matched over all the input set into consideration, we design an enhanced version to deal with some potential attacks such as unlimitedly inputting the attribute set on either the *initiator* side or the *responder* side, etc. Finally, our proposed *E-match* avoids the heavy cryptographic operations and improves the system performance significantly by employing a novel use of the Bloom filter. The security and communication/computation overhead of our schemes are thoroughly analyzed and evaluated via detailed simulations and implementation.

Index Terms—Priority, Privacy, Private Matching, PMSNs

1 Introduction

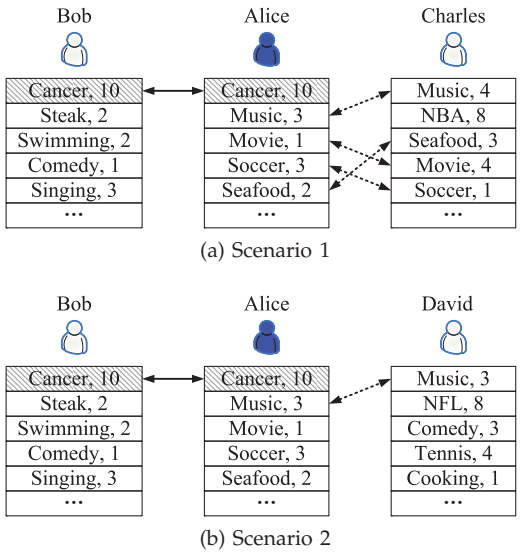
Social networking is one of the fastest-growing activities among mobile users domestically and worldwide. According to eMarketer [1], they estimate the number of US smartphone users will reach 192.4 millions by 2016, and 2.28 billions worldwide. By smartphones equipped with WiFi/Bluetooth modules, users can communicate with others easily and exchange information, content and media on shared communities such as Facebook or Foursquare. Among these services, an important classification is Proximity-based Mobile Social Networks (PMSNs), which deeply relies on mobile users' physical proximity. This kind of applications can provide us more opportunities to discover and make new social interactions within some public places, such as airports, bars or other social spots. PMSNs thus gain increasing attention in social networking.

Normally, to enjoy these activities, people always need to reveal some information such as their attributes or personal information to potential friends nearby as the first step. A straightforward way is that, an *initiator* broadcasts her attributes to nearby users directly, and the *responder*s decide whether to contact her based on common attributes. Obviously, the user's privacy is revealed during such process. Since some of these attributes may be sensitive or private to the user, it is harmful to leak them to everyone nearby, especially the potential malicious users.

To address this problem, many research solutions [2], [3], [4], [5], [6], [7], [8], [9], [**?**] have been proposed. Among these solutions, most of them employ third party servers, and thus, they become the bottlenecks from both security/privacy and system performance points of view. Although this kind of third party servers can be set to offline, the mobile users need to access them to register identities and obtain the matching results, which bring extra 3G/4G communication cost. Some researchers consider this situation as a Private Set Intersection (PSI) problem [10], [11], [12], [13] and try to achieve private matching while avoiding the third party servers by employing Secure Multi-party Computation (SMC) [8] and Paillier Cryptosystem [6], [9]. Unfortunately, the PSI-based solutions can avoid the trusted server effectively but always fail to improve the system performance due to the heavy cryptographic operations.

Moreover, existing schemes do not always produce accurate matching results. For example, [8] and [14] measure the similarity between users by simply counting the number of common attributes, and the matching decisions is made by checking whether the proximity measurement of two profiles is larger, equal, or smaller than a pre-defined threshold value in [15]. However, in reality, user interests may be associated with different priorities. We illustrate our concerns with an example in Fig. 1, the two-tuple represents user's interests with the corresponding priorities. The scenario is that *Alice*'s father is suffering from cancer and she needs to go to hospital everyday after work. In this situation, the most

- *B. Niu, X. Zhu and H. Li are with the School of Telecommunications Engineering, Xidian University, 710071, China. E-mail: xd.niuben@gmail.com and {xyzhu, lihui}@mail.xidian.edu.cn.*
- *T. Zhang is with GSIS, Tohoku University, Sendai, Japan. E-mail: xubu3@163.com.*
- *Z. Lu is with the Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802. E-mail: zongqing@cse.psu.edu.*



(a) Scenario 1

(b) Scenario 2

Fig. 1: Motivation

interested person she wants to know is someone who is facing the same situation. Therefore, in the scenario shown in Fig. 1a, the potential friend of *Alice* is *Bob* instead of *Charles*, even though there are four common attributes between *Alice* and *Charles* and only one with *Bob*. However, this kind of schemes suffers the attack of choosing the attributes as many as possible on the adversary side, leading to the exposure of users' personal information. Although, the authors in [8] proposed a solution by limiting the number of input attributes (e.g., 200) to avoid this attack, it is hard to define a proper number of input attributes for different users. To achieve a fine-grained private matching, Zhang *et al.* [9] consider the priority on each common attribute and define several privacy levels in their work, but they only pay attention on the difference of priorities on each common attribute between users, and ignore the priority value itself. That is to say, it works well for most cases except one shown in the scenario of Fig. 1b. *Alice* has one common attribute with either *Bob* or *David*, it makes sense that she prefers to make friends with *Bob* since she pays more attention on the attribute "cancer". However, the approach in [9] cannot differentiate *David* and *Bob* from the *Alice*'s point of view, since *Alice* and *Bob* have the common attribute of "cancer" at priority 10, and *Alice* and *David* have the common attribute of "music" at priority 3.

From the aforementioned analysis, it is clear that existing work either rely on third party servers, or employ heavy cryptographic tools, or do not fully consider the users' privacy in terms of the priority assigned on each attribute, or produce inaccurate matching results. In this paper, we propose a set of priority-aware private matching schemes to accelerate the widely used PMSNs. The main contributions of this paper are shown as follows.

- We propose a set of schemes to achieve private matching for different privacy goals. *P-match* achieves the priority-aware private matching with considering both the number of common attributes and the corresponding priorities. In the enhanced version *P-match*$^+$, we construct a priority-aware Ochiai similarity coefficient to consider the ratio of attributes matched over all the input set in our similarity function, which can effectively prevent several attacks, such as unlimitedly inputting users' attribute sets. Finally, to make the private matching process more efficiently, we propose *E-match*, which improves the system performance by avoiding the heavy commutative encryption function.
- We provide theoretic and experimental evidences that our proposed matching schemes are secure and can achieve our privacy goals. In addition, they are quite efficient compared to many existing work in terms of the computation cost, communication cost and energy consumption.
- We implement the proposed schemes into smartphones. The experimental results indicate the effectiveness and efficiency of our work.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the preliminaries. Section 4 describes the designs of our proposed schemes. Section 5 and 6 show the security analysis and the evaluation results. Finally, we conclude the paper in Section 7.

2 Related Work

There is a series of applications to provide private matching between users in PMSNs. Most of these solutions employ third party servers, which are always trusted and acting as matching centers to serve users. Specifically, each user sends her attributes to the server, the server replies users with the matching result to indicate the potential "friends". The servers need to know the users' personal information to perform the matching process, it is thus much dangerous when the servers are compromised. Social serendipity [2] provided mobile users more opportunities to make social interactions with potential friends nearby. However, it deeply relied on a trusted server, which keeps all users' profiles and computes the similarity between users when needed. The authors in [3] improved this problem by replacing the trusted server with a service provider, such as Facebook. However, users' profiles are exchanged in plaintext, which lead to serious privacy leakage. SmokeScreen [16] introduced opaque identifier into the information exchanging phase to protect user's real identity. It employed a broker, which also acts as a trusted server to provide matching results to users. As a result, the broker known who is interested in solving whose opaque identifier, which means the broker can infer the relationship between users with high possibility. In their follow-on work [4], they avoided this problem by not disclosing the personal information for matching, and using the location and time information as a replacement. Unfortunately, this kind of servers are still bottlenecks. The servers still need to know these information

to perform the matching process. To avoid the third party servers, many cryptographic tools-based solutions have been proposed over recent years. Some researchers conclude this situation into the Private Set Intersection (PSI) or Authorized PSI (APSI) problem [10], [11], [12], [13], which can effectively avoid the third party servers. Based on PSI, Li *et al.* [8] proposed a set of privacy-preserving profile matching schemes, where an initiating user can find the best match with minimal information leakage to others based on the security properties of Secure Multi-party Computation (SMC). However, the expensive computation cost brought by the heavy cryptographic operations decrease the system performance significantly. In addition, their schemes may fail in reality due to the ignorance on the priorities assigned on each attributes. Zhang *et al.* presented a set of fine-grained private matching schemes to achieve the requirements of mobile users in reality. Under the protection of Paillier Cryptosystem, their schemes achieved fine-grained private matching with considering both the number of common attributes and the assigned priorities. However, they only paid attention on the differences of the priorities.

Agrawal *et al.* [17] proposed a privacy-preserving protocol by using the commutative encryption function, which is more lightweight, to realize secret information sharing between users. The keyed hash functions are also employed to protect the sensitive attributes for mobile users. Then, Vaidya *et al.* [18] extended the secret information sharing phase into N-party setting, and Veneta *et al.* [19] implemented this idea to detect friend-of-friend in mobile social networks. However, since the inherent weaknesses of the scheme in [17], such as unlimitedly inputting behavior and lying behavior, simply employing the commutative encryption function-based solution cannot provide thorough security and privacy properties. In this paper, we provide more properties by combining commutative encryption function with other techniques, such as similarity functions.

3 Preliminaries

In this section, we first state the problem and give the adversary models. Then, we describe the design goals and the cryptography tools in this paper.

3.1 Problem Statement

In PMSNs, each user holds a profile with two dimensional vectors, $U = \{\langle u_1, v_1\rangle, \langle u_2, v_2\rangle, \cdots, \langle u_m, v_m\rangle\}$, where u_i represents the user's attribute and v_i means the priority value assigned to this particular attribute by the user, such as $v_i = 1, \ldots, 9$. Normally, bigger value indicates higher priority. Given a user and her profile U, the problem is how to find the best matched friend for the user with privacy-preserving based on the criteria that the best matched friend should have more common attributes with the user, especially the same priorities on the attributes.

3.2 Adversary Models

Since cryptography tools such as Public Key Infrastructure (PKI) can be easily implemented to protect current communication systems, the attacks from outside adversaries, such as eavesdropping the wireless communication channels or modifying, replying and injecting the captured messages can be easily prevented. Therefore, in this paper, we assume the *initiator* is honest-but-curious [10] or even an attacker directly. That means the *initiator* will honestly follow the protocols, but tries to learn more information than allowed in the honest-but-curious model; or the *initiator* can directly be an attacker, who may illegally input his attributes and the related priorities, to learn more information of nearby users. We also assume that the *responder* is legal or honest-but-curious with two reasons: the identity of a *responder* can be easily authorized by another authority entity, i.e., office in the hospital in the example mentioned in Sec. 1; another reason is that we can reverse our protocols to achieve a dual matching. We further assume that neither the *initiator* nor the *responder* can modify and obtain the result of parameters in the running protocols.

3.3 Design Goals

Our main goal is to thwart the aforementioned threats from either *initiator* or *responder* side. According to the amount of information disclosed during the protocols execution, we define two privacy levels from the *initiator Alice*'s point of view, which can also be equivalently defined from *Bob*'s viewpoint.

Definition 1 (Privacy Level I)**.** *When the protocol ends,* Bob *learns the set of common attributes with* Alice*, as well as priorities on these common attributes, and* Alice *learns the similarity value.*

In this case, the *initiator* believes that all the nearby users are legal or honest-but-curious, and they can be authorized by another entity. For instance, *Alice* may be a new patient in the hospital ward. Her aim is to find a best match.

Definition 2 (Privacy Level II)**.** *When the protocol ends,* Bob *learns the number of common attributes and the similarity value only when it exceeds the pre-defined threshold,* Alice *learns the number of common attributes as well as the similarity value.*

This case indicates two threats from both sides. The illegally input on the *initiator* side and curiously adjust the threshold on the *responder* side.

3.4 Cryptography Tools

Existing private matching solutions always rely on Private Set Intersection (PSI) [10], [12], [13] or Private Cardinality of Set Intersection (PCSI) [20], which is deeply based on heavy cryptographic operations such as Secure Multi-party Computation (SMC) [21]. To avoid the heavy cryptographic operations, our protocols utilize the commutative encryption function [17], which is

more computational friendly and satisfies the condition: $E_{k_1}(E_{k_2}(x)) = E_{k_2}(E_{k_1}(x))$. Thus, a user who has the key k_1 or k_2 learns $x_1 = x_2$ iff. $E_{k_1}(E_{k_2}(x_1)) = E_{k_2}(E_{k_1}(x_2))$, but cannot learn any other x_is of other user if x_i is not a common attribute. Since the priority for every attribute is considered here, it is required that the encryption function needs to be easily deciphered to compute the similarity of two users. So we adopt the power function $f_k(x) = x^k \mod p$ as our encryption function for a safe prime p, i.e. p and $(p-1)/2$ are both prime numbers. For all integers k_1, k_2 and $x \in \mathbb{Z}_p^*$, $\exists$ an integer n, s.t.

$$\begin{aligned} f_{k_1}(f_{k_2}(x)) &= f_{k_1}(x^{k_2} \mod p) \\ &= f_{k_1}(x^{k_2} - np) \\ &= (x^{k_2} - np)^{k_1} \mod p \\ &= x^{k_2 k_1} \mod p, \end{aligned} \tag{1}$$

the last equality follows from the binomial theorem. Similarly it holds $f_{k_2}(f_{k_1}(x)) = x^{k_1 k_2} \mod p$. Therefore $f_{k_1}(f_{k_2}(x)) = f_{k_2}(f_{k_1}(x))$. To obtain the decryption function, we need a corresponding number k' to every k. Choose k' such that $k'k = 1 \mod \phi(p)$, where $\phi(p)$ is the Euler phi-function of p and $\phi(p) = p - 1$ since p is a prime. We use the Extended Euclidean Algorithm to yield k' and let $g_k(y) := y^k \mod p$ for $y \in \mathbb{Z}_p^*$. Then

$$\begin{aligned} g_{k'}(f_k(x)) &= g_{k'}(x^k \mod p) \\ &= x^{k'k} \mod p \\ &= x^{n\phi(p)+1} \mod p \\ &= x. \end{aligned} \tag{2}$$

The last equality holds because of the Euler Theorem. To guarantee that x and y are both in $\mathbb{Z}_p^*$, we need a cryptographic hash function $h_p(\cdot)$ which has the quadratic residues modulo p as its range.

4 The Proposed Schemes

In this section, we present a suit of priority-aware private matching schemes. The basic version, *P-match*, satisfies Privacy Level I. As an improvement to achieve Privacy Level II, we propose the enhanced version *P-match*$^+$. Finally, the efficient version, *E-match*, improves the performance significantly by avoiding the heavy cryptographic tools such as commutative encryption function.

In our scenario, there may be several users in a particular area at a particular time period. Specifically, each user holds a set of messages $\{\langle x_i, a_i\rangle, K_A, k_A, x_i \in X\}$, where X is a set of attributes, a_i is the corresponding priority of x_i, K_A and k_A are two secret keys. Moreover, all the procedures are $\mod p$ in our schemes. We use the notations shown in Table 1.

4.1 Basic Version

4.1.1 Introducing the basic similarity function

Let *Alice*'s attribute set and corresponding priority vector be X and A respectively, and *Bob*'s attribute set and corresponding priority vector be Y and B respectively. The set of common attributes between *Alice* and *Bob* is denoted as $S = X \cap Y$, where $q = |S|$ and $S = \{s_1, s_2, \ldots, s_q\}$. Then we arrange the corresponding priorities of *Alice* and *Bob* on the common attributes into vector $V_A = (a_1, a_2, \ldots, a_q)$ and $V_B = (b_1, b_2, \ldots, b_q)$, respectively. The most widely applied similarity function is cosine similarity:

TABLE 1: Notations

$\|A\|$	Number of elements in the set A
R	Public attribute pool for all users, $n = \|R\|$, $R = \{r_i\}_{i=1}^n$
X	Attribute set of *Alice*, $n_1 = \|X\|$, $X = \{x_i\}_{i=1}^{n_1}$, $x_i \in R$
Y	Attribute set of *Bob*, $n_2 = \|Y\|$, $Y = \{y_i\}_{i=1}^{n_2}$, $y_i \in R$
S	$S = X \cap Y$, $q = \|S\|$
$a_i(b_i)$	Priority of $x_i(y_i)$, a_i, $b_i = \{1, 2, \ldots, 10\}$
V_A	$V_A = \{a_i\}_{i=1}^q$, where each a_i is the priority of $x_i \in S$
V_B	$V_B = \{b_i\}_{i=1}^q$, where each b_i is the priority of $y_i \in S$
ξ^*	the expected value of the random variable ξ

$$\cos(\theta) = \frac{V_A \cdot V_B}{\|V_A\| \cdot \|V_B\|}, \tag{3}$$

where θ is the angle between V_A and V_B. It is often used to measure the angular similarity between two vectors. However, cosine similarity is orthogonal to the priorities on a common attribute. That is cosine similarity can be high when the priorities on the common attribute are quite different or exactly the same. This implies its defect. We do not accept the high similarity if the priorities differs far from each other. Thus, it is not the best choice in our scenario.

Different from cosine similarity, Jaccard similarity coefficient better fits our scenario, which considers the quotient of the size of the intersection over the size of the union of X and Y. To simplify the computation we employ its variant form, Tanimoto similarity coefficient [22]:

$$T(A, B) = \frac{V_A \cdot V_B}{\|V_A\|^2 + \|V_B\|^2 - V_A \cdot V_B}, \tag{4}$$

where V_A and V_B are the same as in (3). The inner product appears in the numerator and the denominator of (4) displays the difference between V_A and V_B, and the norm term adjusts the size of the unit. After these refinement, Tanimoto similarity coefficient can embody the difference between two priority set on common attributes.

4.1.2 P-match

Based on the commutative encryption function [17] and Tanimoto similarity coefficient, we design our basic privacy-aware private matching scheme, *P-match*, which considers both the number of common attributes and the corresponding priorities on them. As an initialization, users encrypt their attributes and priorities under their secret keys. Fig. 2 shows the details and the procedure is as follows:

(i) When two users are within the communication range of each other, the *initiator Alice* begins the matching process by broadcasting a message $\langle h_p(x_i)^{K_A}, a_i^{k_A}\rangle$;
(ii) as one of the *responder*s, *Bob* replies *Alice* with $h_p(y_i)^{K_B}$;

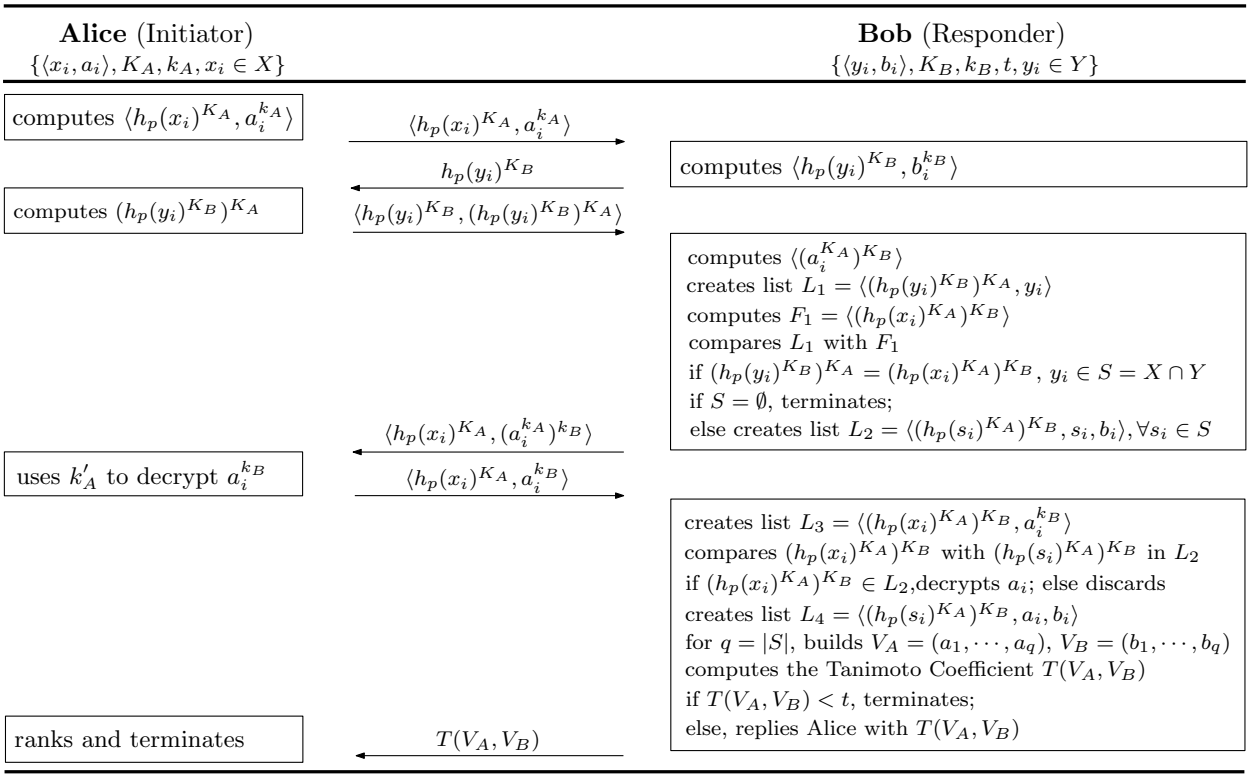


Fig. 2: Basic Private Matching Scheme

(iii) based the received message, *Alice* computes $(h_p(y_i)^{K_B})^{K_A}$ and sends the two-tuple $\langle h_p(y_i)^{K_B}, (h_p(y_i)^{K_B})^{K_A}\rangle$ together to *Bob*;

(iv) *Bob* finds matches and creates a list $L_1 = \langle (h_p(y_i)^{K_B})^{K_A}, y_i\rangle$, and computes a set $F_1 = \langle (h_p(x_i)^{K_A})^{K_B}\rangle$. Then, he compares the first element in L_1 with F_1 to compute the common attributes with *Alice*, the common ones form as set S, i.e., $S = X \cap Y$. *Bob* creates another list $L_2 = \langle (h_p(s_i)^{K_A})^{K_B}, s_i, b_i\rangle, \forall s_i \in S$ unless $S = \emptyset$, and sends back the $\langle h_p(x_i)^{K_A}, (a_i^{k_A})^{k_B}\rangle$;

(v) *Alice* computes her k'_A by Extended Euclidean algorithm and decrypts the second part of the received message, then sends out the results $\langle h_p(x_i)^{K_A}, a_i^{k_B}\rangle$;

(vi) upon the received messages, *Bob* creates another list $L_3 = \langle (h_p(x_i)^{K_A})^{K_B}, a_i^{k_B}\rangle$, then compares the first element in L_3 with $(h_p(s_i)^{K_A})^{K_B}$ in L_2. He obtains the corresponding a_i by comparing if the received $a_i^{k_B} = b_i^{k_B}$ when he can find $(h_p(x_i)^{K_A})^{K_B} \in L_2$, otherwise, discards them. When he gets these information, another list is created as $L_4 = \langle (h_p(s_i)^{K_A})^{K_B}, a_i, b_i\rangle$. Then for $q = |S|$, *Bob* uses priorities on each common attribute of *Alice* and himself to build two vectors $V_A = (a_1, a_2, \cdots, a_q)$ and $V_B = (b_1, b_2, \cdots, b_q)$, respectively. To measure the similarity, he computes the Tanimoto Coefficient $T(V_A, V_B) = \frac{V_A \cdot V_B}{||V_A||^2 + ||V_B||^2 - V_A \cdot V_B}$, and compares with the threshold t which is predefined by *Bob*. If $T(V_A, V_B) < t$, the process terminates, otherwise, *Bob* replies *Alice* with $T(V_A, V_B)$.

As the *initiator* may receive several replies from others, a ranking of Tanimoto Coefficient is provided for the *initiator* to choose the best match.

By utilizing *P-match*, we achieve Privacy Level I. *Bob* can learn the common attributes with *Alice*, as well as the corresponding priorities, while *Alice* learns nothing except the similarity value. However, there may be a problem when the scenario changes, i.e., *Alice* wants to know some knowledge since *Bob* may cheat on her. Thus, we propose an enhanced version *P-match*$^+$.

4.2 Enhanced Version

4.2.1 Constructing the enhanced similarity function

When we only consider the common attributes and their priorities, Tanimoto similarity coefficient is energetic, but not effective enough if all the attributes and priorities are taken into account, because it is impossible to invent an efficient way to put *Alice*'s and *Bob*'s all attributes in the same order if $X \neq Y$. Hence we need another function, Ochiai similarity coefficient [23]

$$O(A, B) = \frac{|A \cap B|}{\sqrt{|A| \cdot |B|}}. \tag{5}$$

This coefficient was firstly applied in biology and is useful in comparing the faunistic feature between two different localities. It is also a ratio, where the numerator is the size of the intersection of two sets A, B, and the denominator is the geometric average between the size of A and B. However, since all the priorities, on common and uncommon attributes, are considered here, and we cannot do the intersection between two ordered priority

sets, we need some more effective similarity function. For that purpose, we notice that our priority is a kind of weight function. For a finite set $Z = \{z_1, z_2, \ldots, z_\lambda\}$ where every z_i has the weight $w(z_i)$, $i = 1, 2, \ldots, \lambda$, let the weight function be $w : Z \to \mathbb{I}$, then the un-weighted sum on Z is defined by $\sum_{i=1}^{\lambda} z_i$ while the weighted sum on Z is defined by $\sum_{i=1}^{\lambda} z_i w(z_i)$. If we let $Z = X$, $\mathbb{I} = \{1, \ldots, 9\}$ and $f(x_i) = a_i$, then the weighted model becomes our scenario. This implies that we can regard the priority as a counting rule of an attribute, which is, we count an attribute x_i a_i times.

Therefore, we construct a new priority-aware coefficient $P(A, B)$ based on Ochiai similarity coefficient and the weighted sample. For two attribute sets $X = \{x_1, x_2, \ldots, x_{n_1}\}$ and $Y = \{y_1, y_2, \ldots, y_{n_2}\}$, the corresponding priority vectors are $A = (a_1, a_2, \cdots, a_{n_1})$ and $B = (b_1, b_2, \cdots, b_{n_2})$, let $S = X \cap Y = \{s_1, \ldots, s_q\}$. First we generate two counting sets:

$$\begin{aligned} X' &= \{x_1, x_1+1, \ldots, x_1+a_1, \\ &\quad x_2, x_2+1, \ldots, x_2+a_2, \ \ldots \\ &\quad x_{n_1}, x_{n_1}+1, \ldots, x_{n_1}+a_{n_1}\}, \end{aligned} \tag{6}$$

$$\begin{aligned} Y' &= \{y_1, y_1+1, \ldots, y_1+b_1, \\ &\quad y_2, y_2+1, \ldots, y_2+b_2, \ \ldots \\ &\quad y_{n_2}, y_{n_2}+1, \ldots, y_{n_2}+b_{n_2}\}. \end{aligned} \tag{7}$$

Then

$$|X'| = \sum_{i=1}^{n_1} a_i, \ |Y'| = \sum_{i=1}^{n_2} b_i. \tag{8}$$

Next we find the intersection of X' and Y',

$$\begin{aligned} X' \cap Y' &= \{s_1, s_1+1, \ldots, s_1+c_1, \\ &\quad s_2, s_2+1, \ldots, s_2+c_2, \ \ldots \\ &\quad s_q, s_q+1, \ldots, s_q+c_q\}, \end{aligned} \tag{9}$$

where every $c_i = \min\{a_i, b_i\}$ for $i = 1, \ldots, q$. Then

$$|X' \cap Y'| = \sum_{i=1}^{q} c_i = \sum_{i=1}^{q} \min\{a_i, b_i\}. \tag{10}$$

Provided (8) and (10), now we apply the Ochiai similarity coefficient to the two counting sets X' and Y' given by (6) and (7), and define it to be our new similarity function:

$$\begin{aligned} P(A, B) &= O(X', Y') \\ &= \frac{|X' \cap Y'|}{\sqrt{|X'| \cdot |Y'|}} \\ &= \frac{\sum_{i=1}^{q} \min\{a_i, b_i\}}{\sqrt{\sum_{i=1}^{n_1} a_i} \cdot \sqrt{\sum_{i=1}^{n_2} b_i}}. \end{aligned} \tag{11}$$

The range of the priority-aware similarity coefficient (11) is from 0 to 1, which are corresponding to the cases "no common attribute at all" for 0 and "same attributes, same priorities" for 1.

Algorithm 1: Part of Enhanced Version on Responder Side

Input : when receives $\langle h_p(x_i)^{K_A}, a_i^{k_B} \rangle$ from *Alice*
Output: $n, P(A, B)$ to *Alice*

computes $(h_p(x_i)^{K_A})^{K_B}$;
sends $\langle h_p(x_i)^{K_A}, (h_p(x_i)^{K_A})^{K_B} \rangle$ to *Alice*;
computes k'_B to decrypt a_i;
creates $L_3 = \langle (h_p(x_i)^{K_A})^{K_B}, a_i \rangle$;
compares $(h_p(x_i)^{K_A})^{K_B}$ in L_3 with $(h_p(s_i)^{K_A})^{K_B}$ in L_2;
for those $(h_p(x_i)^{K_A})^{K_B}$ in L_2, computes $c_i = \min\{a_i, b_i\}$;
$P(A, B) = \frac{\sum_{i=1}^{q} c_i}{\sqrt{\sum_{i=1}^{|X|} a_i} \cdot \sqrt{\sum_{i=1}^{|Y|} b_i}}$, where $q = |S|$;
if $(P(A, B) < t)$ **then**
    terminates;
else
    replies *Alice* with $P(A, B)$
end

4.2.2 P-match$^+$

To achieve Privacy Level II, we change some procedures from step vi in our basic version.

Algorithm 1 shows the changes on *responder* side in our *P-match*$^+$. When *Bob* receives $\langle h_p(x_i)^{K_A}, a_i^{k_B} \rangle$ from *Alice*, he computes $(h_p(x_i)^{K_A})^{K_B}$ and sends them back to *Alice* together with $h_p(x_i)^{K_A}$. He can also decrypt a_i by computing k'_B. Then he creates a list $L_3 = \langle (h_p(x_i)^{K_A})^{K_B}, a_i \rangle$ and compares the first element in L_3 with $(h_p(s_i)^{K_A})^{K_B}$ in L_2. Followed, he computes $c_i = \min\{a_i, b_i\}$ for those $(h_p(x_i)^{K_A})^{K_B}$ in L_2, the priority-aware Ochiai Coefficient can be computed as $P(A, B) = \frac{\sum_{i=1}^{q} c_i}{\sqrt{\sum_{i=1}^{|X|} a_i} \cdot \sqrt{\sum_{i=1}^{|Y|} b_i}}$, where $q = |S|$. If $P(A, B) < t$, the algorithm terminates, otherwise, *Bob* replies $P(A, B)$ to *Alice*.

Algorithm 2: Part of Enhanced Version on Initiator Side

Input : when receives $\langle h_p(x_i)^{K_A}, (h_p(x_i)^{K_A})^{K_B} \rangle$ from *Bob*
Output: The number of common attributes with *Bob*

creates $\langle x_i, (h_p(x_i)^{K_A})^{K_B} \rangle$;
compares $(h_p(x_i)^{K_A})^{K_B}$ with $(h_p(y_i)^{K_B})^{K_A}$;
if $((h_p(x_i)^{K_A})^{K_B} = (h_p(y_i)^{K_B})^{K_A})$ **then**
    $s_i = x_i$ and $S = S \cup \{s_i\}$;
else
    terminates;
end
outputs $|S|$;

While on the *initiator* side, *Alice* needs to do some computation work to obtain the number of the common attributes, the details are shown in Algorithm 2. Based on the received message $\langle h_p(x_i)^{K_A}, (h_p(x_i)^{K_A})^{K_B} \rangle$ from

Bob, the algorithm creates a list of $\langle x_i, (h_p(x_i)^{K_A})^{K_B}\rangle$, and obtains the set of common attributes by comparing $(h_p(x_i)^{K_A})^{K_B}$ in this list with $(h_p(y_i)^{K_B})^{K_A}$, which is received before. It terminates if there is no match found by computing $(h_p(x_i)^{K_A})^{K_B} = (h_p(y_i)^{K_B})^{K_A}$, otherwise, outputs $|S|$ to *Alice*.

4.3 Efficient Version

For both *P-match* and *P-match*$^+$, they can achieve their designed goals at the cost of system performance. Since the heavy cryptographic operations, such as the keyed hash functions and exponentiation operations, are hard to perform in current mobile devices, the performance drops dramatically when the number of attributes goes large. We thus design a more efficient version, *E-match*, which employs a Bloom filter **[?]** instead of the heavy exponentiation operations in an honest-but-curious environment. A Bloom filter is a space-efficient probabilistic data structure which is used to test whether an element is a member of a set.

4.3.1 Initialization

Suppose that the public database is $R = \{r_i\}_{i=1}^n$ consisting of attributes of all users. We let $\{h_i(\cdot)\}_{i=1}^l$ be a family of hash functions with $h_j(r_i) \in [1, \lambda]$ for an attribute $r_i \in R$, and $\mathcal{H}$ be a large public pool of hash functions of such h_i with each indexed by a unique identifier. An empty λ-bit Bloom filter is an array of λ bits, all setting to 0 bits. To add an element r_i to the λ-bit Bloom filter, we set all the bits in $h_j(r_i)$ positions 1, $1 \le j \le l$. To check whether an r_i is some user's attribute, we verify whether all the bits in positions $h_j(r_i)$ are 1. If not, r_i is not this user's attribute; otherwise, r_i is his/her attribute with a probability determined by n, λ and l.

We label each $r_i \in R$ by the 2-tuple $\{j, r_i(j)\}_{j=1}^{\kappa}$, where $r_i(j) = r_i + j - 1$ is the counting function of r_i. Then the database R is extended to the set $\{i, \{j, r_i(j)\}_{j=1}^{\kappa}\}_{i=1}^n$. Moreover, this extended set can be identified with the indexed set $R' = \{\{i, j, r_i(j)\}_{j=1}^{\kappa}\}_{i=1}^n$. If *Alice* has an attribute set $X = \{x_i\}_{i=1}^{n_1}$ with the priority set $\{a_i\}_{i=1}^{n_1}$, we can assign her a personal set $S_A = \{\{i, j, x_i(j)\}_{j=1}^{a_i}\}_{i=1}^{n_1}$, where $x_i = r_{i'}$ for an attribute $r_{i'} \in R$, and $a_i \in \{1, 2, \cdots, \kappa\}$ is the priority of x_i. The same technique can be applied to *Bob*, we get $S_B = \{\{i, j, y_i(j)\}_{j=1}^{b_i}\}_{i=1}^{n_2}$, where y_i is in *Bob*'s attribute set Y, $|Y| = n_2$, b_i is the priority of y_i. Denote $q_1 := |S_A|$, $q_2 := |S_B|$, then we have $q_1 = \Sigma_{i=1}^{n_1} a_i = |X'|$, $q_2 = \Sigma_{i=1}^{n_2} b_i = |Y'|$ with X' given by Equation 6, and Y' given by Equation 7.

We pair every l a random number l', $1 < l' < l$, and publish the 2-tuple (l, l') to all users. Now we can use a λ-bit Bloom filter to check how many attributes are in *Alice*' set X with her priorities, and how many are in *Bob*'s set Y with his priorities, so that one of them can learn the similarity level.

4.3.2 E-match

We then present our proposed *E-match* shown in Fig. 3 in details.

(i) *Alice* sends a request to *Bob*.
(ii) *Bob* agrees.
(iii) *Alice* randomly chooses $\{h_i\}_{i=1}^l \subset \mathcal{H}$ with indexes denoted by $\mathcal{H}_A$. *Alice* adds each $\{i, j, x_i(j)\}$ in S_A into a λ-bit Bloom filter array, denoted by BF_A, with l' different random hash functions in $\mathcal{H}_A$ and $(l - l')$ random hash functions out of $\mathcal{H}$ (Computation offline above). *Alice* sends $\mathcal{H}_A$ and BF_A to *Bob*.
(iv) *Bob* counts the number of 0-bits in BF_A, denoted by d_1, then he adds every $\{i, j, y_i(j)\}$ in S_B to BF_A using $\mathcal{H}_A$ to get a λ-bit Bloom filter array, denoted by BF_B. He counts the number of 0 bits BF_B, say, d_0, and computes

$$P^*(A, B) = \frac{\sqrt{l}[q_2 - \lambda(\ln d_1 - \ln d_0)]}{l'\sqrt{\lambda q_2(\ln \lambda - \ln d_1)}}, \tag{12}$$

Bob compares $P^*(A, B)$ with his pre-defined threshold t, so that he decides whether to match *Alice* or not.

5 Security Analysis

In this section, we prove that our proposed schemes can achieve the required privacy level in turns.

5.1 Analysis of the Basic Scheme

Theorem 1. *P-match ensures Privacy Level I if the commutative encryption function is secure.*

Proof. For *Alice*. *Bob* encrypts the hash value of his attributes using his secret keys K_B and k_B, then sends the messages ($h_p(y_i)^{K_B}$ and $\langle h_p(x_i)^{K_A}, (a_i^{k_A})^{k_B}\rangle$) to *Alice* side. As mentioned in Section 3.4, the commutative encryption function is secure, so it is computationally impossible to *Alice* to obtain any of the attributes or the corresponding priorities of *Bob*. When the protocol ends, what *Alice* is able to get is a similarity value, which, in general, indicates a rough similarity. However, it is impossible to deduce any personal information, such as the number of common attributes or the corresponding priorities, from the similarity value.

For *Bob*. As a *responder*, *Bob* learns more information than *Alice*. It is reasonable because a *responder* has the weaker motivation to start an attack. In *P-match*, even if this kind of attacks happens, the commutative encryption function and cryptographic hash function can guarantee that what *Bob* gets are only the common attributes and the corresponding priorities of *Alice*, since all encryption function are injective, i.e. $(h_p(x_i)^{K_A})^{K_B} = (h_p(y_i)^{K_B})^{K_A}$ if and only if $x_i = y_i$, it holds the similar conclusion for the corresponding priorities. So, *Bob* knows nothing about *Alice* except the common attributes and priorities so that he can compute the similarity value on his own side. □

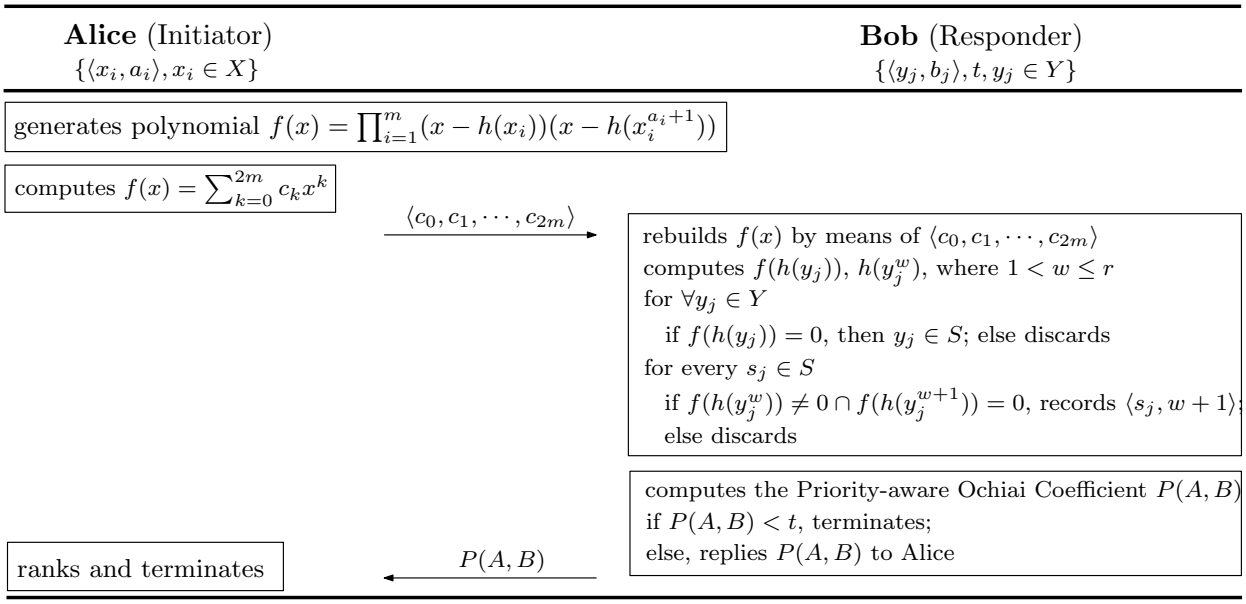


Fig. 3: Efficient Private Matching Scheme

5.2 Analysis of the Enhanced Version

Theorem 2. *P-match$^+$ ensures Privacy Level II if the commutative encryption function is secure.*

Proof. Similar to the proof of Theorem 1, the commutative encryption function and keyed hash function provide end users with a secure channel, it means that only the one who has the secret key can decrypt the message if he has at least one common attribute with the other user. Here, the following two possible threats are our focal points in this match, 1) *Alice* may illegally input her attributes as well as the priorities; 2) *Bob* tries to learn extra information from the received messages from *Alice* by adjusting his threshold.

For case 1), the more *Alice* inputs attributes x_i with the higher priorities a_i, the more personal information of *Bob* she gets easily and possibly. We note that for the priority-aware Ochiai coefficient, the denominator of (11) has the factor $\sqrt{\sum_{i=1}^{|X|} a_i}$, which goes large quickly if either $|X|$ or any $|a_i|$ becomes large. Meanwhile, the numerator $\sum_{i=1}^{q} \min\{a_i, b_i\}$ and the other term $\sqrt{\sum_{i=1}^{|Y|} b_i}$ in the denominator are preserved for the same responder *Bob*. Hence, if *Alice* is a malicious user, the priority-aware Ochiai coefficient between her and any other responder *Bob* decreases dramatically so that *Bob* terminates the protocol according to his pre-defined threshold, and *Alice*'s attack fails.

For case 2), to get more extra information from *Alice*, *Bob* does not want to terminate the protocol, so he must lower his threshold t to proceed Algorithm 1. But in P-match$^+$ *Alice* does not send *Bob* anything except $h_p(x_i)^{K_A}$, $a_i^{k_A}$ and $a_i^{k_B}$ in the beginning phase. Since the encryption function $f_k(x) = x^k$ is pre-image resistant, so *Alice*' information which is known to *Bob* is not change at all, whether *Bob* adjusts his threshold. That means *Bob* cannot get more information from *Alice* by lowering the threshold t.

Considering the proof of Theorem 1, we obtain the security for the enhanced version P-match$^+$. □

5.3 Analysis of the Priorities

Agrawal *et al*. [17] proved that, it computes impossible to map x_i^k to x_i without knowing the secret key k under the Decisional Diffie-Hellman hypothesis (DDH). Specifically, for fixed values of i and j, $\langle x_i, f_k(x_i), y_j, f_k(y_j)\rangle$ is indistinguishable from $\langle x_i, f_k(x_i), y_j, z\rangle$, where $f_k(x) = x^k \mod p$. That means, *Bob* cannot map x_i^k back to x_i if he does not know the value of k, which is *Alice*'s secret key and only known to *Alice*. Based on these conclusions, each user's attributes are safe. Now we consider the security on the priorities. For the two similarity functions $T(A, B)$ and $P(A, B)$, note that *Alice* knows the common attributes and *Bob*'s corresponding priorities if $T(A, B) = 1$ and $|X| = |X \cap Y|$ in the basic version, and if $P(A, B) = 1$ in the enhanced version. Otherwise, when the two conditions do not hold, we have the following results.

Theorem 3. *In P-match, if Alice has only one attribute, she can confirm a common attribute and the corresponding priority of Bob iff. there is only one common attribute between them. Otherwise, Alice knows nothing if she has more than one attribute.*

Proof. When *Alice* and *Bob* have only one common attribute with priorities a and b, respectively, based on *P-match*, she can easily compute b from $T(A, B) = \frac{a \cdot b}{a^2+b^2-a \cdot b}$, where a and $T(A, B)$ are known to *Alice*. The other direction is trivial. Now we assume *Alice* has more than one attribute. Since in *P-match* we only take the common attributes and their priorities into account,

and we have proved that *P-match* ensures Privacy Level I if the commutative encryption function is secure in Theorem 1, which means *Alice* knows nothing if the encryption function is secure. We note that the attributes and corresponding priorities are encrypted by two different secret keys, and the security of the attribute part was proved by Agrawal *et al*. in [17], then *Alice* knows nothing. □

Theorem 4. *In P-match$^+$, Alice can confirm a common attribute as well as the corresponding priority of Bob iff. she has only one attribute which is also the only one common attribute. Otherwise, Alice would know at most one common attribute and its corresponding priority of Bob.*

Proof. The proof for the first part is similar to Theorem 3. Since in this enhanced version, Privacy Level II can be ensured if the commutative encryption function is secure, which means, *Alice* knows the common attributes. If *Alice* has at least two attributes and they have the common attributes of number $q \geq 2$, from the expression of $P(A,B)$, *Alice* only knows the ratio $\frac{\sum_{i=1}^{q}\min\{a_i,b_i\}}{\sqrt{\sum_{i=1}^{m}b_i}}$. However, this ratio means nothing if *Alice* does not know the right priority of every common attribute. □

As stated above, we see that there is other potential attack that a malicious user inputs only one attribute with the corresponding priority to learn other's secret information. However, there are many practical ways to figure out this problem, such as setting a rule to limit the minimum number of input. Thus, it is omitted here.

5.4 Analysis of the E-match

The security of the *E-match* is based on the probability of false positives for the λ-bit Bloom filter. The accuracy of the *E-match* is given by the following two results.

Theorem 5. *Via our protocol, Bob can know the expected values of counts $q_1 = |S_A|$ and $q' := |S_A \cap S_B|$, by*

$$q_1^* = \frac{\lambda(\ln\lambda - \ln d_1)}{l}, \tag{13}$$

and

$$q'^* = \frac{lq_2 + \lambda(\ln d_0 - \ln d_1)}{l'}, \tag{14}$$

where d_1 is the number of 0-bits in BF_A, d_0 is the number of 0-bits in BF_B.

Proof. According to [**?**], the distribution for 0-bits in a λ-Bloom filter can be regarded as a binomial distribution. When the length λ is large enough, it approximates a normal distribution asymptotically. We now suppose that λ is large enough. For a fixed bit in BF_A, the probability that it is set to 0 by adding one element $x_i(j)$ with l hash functions is $(1-\frac{1}{\lambda})^l$, the probability that it is set to 0 by adding all elements $x_i(j)$ with l hash functions is

$$(1-\frac{1}{\lambda})^{lq_1} \approx e^{-\frac{lq_1}{\lambda}}.$$

Thus we have $e^{-\frac{lq_1}{\lambda}} = \frac{d_1}{\lambda}$. Solving this equation leads to q_1^* given by the equation(13).

For a fixed bit in BF_A, the probability that it is set to 0 by adding q' common element with l' hash functions is $(1-\frac{1}{\lambda})^{l'q'}$, the probability that it is set to 0 not by adding q' common element with l' hash functions is $(1-\frac{1}{\lambda})^{lq_1-l'q'}$. Thus for a bit in BF_B, the conditional probability that it is set to 0 is

$$(1-\frac{1}{\lambda})^{lq_2+lq_1-l'q'} \approx e^{-\frac{lq_1+lq_2-l'q'}{\lambda}}.$$

Then we have

$$e^{-\frac{lq_1+lq_2-l'q'}{\lambda}} = \frac{d_0}{\lambda}.$$

Solving this equation, and combining with the Equation (13) lead to q'^* in the Equation (14). If we note that the Equation (11) of Ochiai coefficient implies that

$$P(A,B) = \frac{q'}{\sqrt{q_1 q_2}},$$

The Equation (12) is obtained from the Equation (13) and Equation (14). □

Theorem 6. *In the E-match, let $q_2 = |S_A|$, then, with the same notations shown in Theorem 5, we have*

$$q_1^* \sim \mathcal{N}[q_1, \frac{\lambda}{l^2}(e^{\frac{lq_1}{\lambda}} - 1)], \tag{15}$$

and

$$q'^* \sim \mathcal{N}[q', \frac{\lambda}{(l')^2}(e^{\frac{lq_1}{\lambda}} + e^{\zeta} - 2 - \zeta)] \tag{16}$$

with $\zeta = \frac{lq_1+lq_2-l'q'}{\lambda}$.

Proof. From [**?**] we know that when λ is large enough,

$$d_1 \sim \mathcal{N}[\mu_1(q_1), \sigma_1^2(q_1)] \tag{17}$$

where

$$\mu_1(q_1) = \lambda(1-\frac{1}{\lambda})^{lq_1} \approx \lambda e^{-\frac{lq_1}{\lambda}},$$

$$\sigma_1^2(q_1) = \lambda(1-\frac{1}{\lambda})^{lq_1}[1-(1-\frac{1}{\lambda})^{lq_1}] \approx \lambda e^{-\frac{lq_1}{\lambda}}(1-e^{-\frac{lq_1}{\lambda}}).$$

To apply the special version of the central limit theorem stated in the Theorem 6 in [**?**], we let $\lambda \to \infty$, $lq_1 \to \infty$ while $\frac{lq_1}{\lambda}$ is fixed. Regard q_1 as a variable, and note that $\mu_1(q_1)$ is monotonically decreasing, then μ_1 has an inverse function, denoted by g_1. Based on the Theorem 6 in [**?**] and the Equation (17), we know that

$$g_1(d_1) \sim \mathcal{N}[g_1(\mu_1(q_1)), \delta_1^2(q_1)], \tag{18}$$

where $\delta_1^2(q_1) = \sigma_1^2(q_1)(g_1'(\mu_1(q_1)))^2$. Since $g_1(d_1) = q_1^*$, and

$$g_1'(\mu_1(q_1)) = \frac{1}{\mu'(q_1)} = -le^{-\frac{l}{q_1}}, \tag{19}$$

we thus obtain the result in Equation (15).

To prove (16), we let

$$q_0'^* = \frac{lq_1 + lq_2 - \lambda(\ln\lambda - \ln d_0)}{l'}. \tag{20}$$

Then the same technique as used in [**?**] results in

$$q_0'^* \sim \mathcal{N}[q', \frac{\lambda(e^{\zeta}-1-\zeta)}{(l')^2}]. \tag{21}$$

Since the distributions of q_1^* and $q_0'^*$ are both normal distributions, the distribution of the 2-tuple $(q_1^*, q_0'^*)$ is a multivariable normal distribution. Note that $q'^* = q_0'^* + \frac{l}{l'}q_1^* - \frac{l}{l'}q_1$, thus from the Equations (15), and (21) we have the Equation (16). □

The estimation of the *E-match* is given by the following theorem.

Theorem 7. *In E-match, let q_1^* and q'^* be defined by the Equations* (13) *and* (14)*, respectively. Then for ϵ_1, ϵ_2 such that*

$$\epsilon_1 q_1 \geq \frac{\left(\lambda(e^{\frac{lq_1}{\lambda}} - 1)\right)^{\frac{1}{2}}}{l}$$

and

$$\epsilon_2 q' \geq \frac{\left(\lambda(e^{\frac{lq_1}{\lambda}} + e^{\zeta} - 2 - \zeta)\right)^{\frac{1}{2}}}{l'},$$

we have

$$Pr(|q_1^* - q_1| \leq \epsilon_1 q_1) \geq 1 - p_1, \quad Pr(|q'^* - q'| \leq \epsilon_2 q') \geq 1 - p_2, \quad (22)$$

where

$$p_1 \geq \frac{\lambda(e^{\frac{lq_1}{\lambda}} - 1)}{\epsilon_1^2 q_1^2 l^2}, \quad p_2 \geq \frac{\lambda(e^{\frac{lq_1}{\lambda}} + e^{\zeta} - 2 - \zeta)}{\epsilon_2^2 (q'l')^2}.$$

Proof. Combining the Equations (15), (16) with Chebyshev's inequality leads to the result in Equation (22). □

We qualify the priority-aware attribute by the Shannon entropy, which is a common measurement of uncertainty. In the *E-match*, only *Alice* sends her information to other users, so that we only examine *Alice*'s set S_A, where $S_A \subseteq R'$. On *Bob*'s side, he only knows κn before the protocol, and he will know λ, l and l' after the protocol. He will compute the expected value of $|S_A|$ but will not know any explicit attribute in S_A. Considering the size of R' and S_A, there are $\frac{(\kappa n)!}{q_1!\,(\kappa n - q_1)!}$ choices of $S_A \subseteq R'$. Among those choices, although some sets are counted repeatedly, *Bob* does not know which sets are the same since he cannot know any of the priority a_i. That means, the $\frac{(\kappa n)!}{q_1!\,(\kappa n - q_1)!}$ candidate sets are all equally unknown in *Bob*'s eyes. We replace q_1 by q_1^*. Thus the uncertain attribute of *Alice* to *Bob* can be estimated in bits by

$$\mathbf{E}^* = \log_2 \frac{(\kappa n)!}{q_1!\,(\kappa n - q_1)!}.$$

Moreover we have the following result for the entropy.

Theorem 8. *In the E-match, suppose that BF_A is the Bloom filter constructed by Alice using l' hash functions in $\mathcal{H}_A$ based on her priority-aware personal set S_A. Then after sending BF_A and $\mathcal{H}_A$ to Bob, her remaining privacy information of S_A against Bob is*

$$\mathbf{E} = q_1 \mathbf{E}[i, j], \quad (23)$$

where

$$\begin{aligned} \mathbf{E}[i,j] &= \sum_{x=1}^{\kappa n} \binom{\kappa n}{x} P^x (1-P)^{(\kappa n - x)} \log_2 x, \\ P &= \sum_{i=1}^{l} \binom{l}{i} p^i (1-p)^{l-i}, \\ p &= 1 - e^{-\frac{lq_1}{\lambda}}. \end{aligned}$$

TABLE 2: Experiment setting

	Cancer	Music	Football	Tennis	Cooking
Alice	8	4	1	3	2
Bob	7	-	2	-	-
Charles	1	9	4	2	1
David	9	8	-	6	-
Emmy	-	2	9	1	1
Frank	8	3	-	-	-

Proof. We refer to Theorem 2 in [**?**] for the privacy-preserving spatiotemporal matching. The result in Equation (23) can be obtained if we consider each possible location cell (cID) to be each possible interest $r_i[j]$, and note that the size of the whole interest pool R' is κn. □

6 Performance Evaluations

In this section, we first design an experiment to verify the correctness of our proposed protocols. Then we analyze the system complexity and show our experimental results.

6.1 Experiment of Correctness

To verify our work, we design a simple experiment by letting the *initiator Alice* do the matching with several candidates (5 users in our experiment) in vicinity. For simplicity, we assign 5 attributes for each user, they can set the priorities on each individual from 1 to 9 randomly. As a result, each user may have dozens of combinations on these attributes. We choose a snapshot of these priorities, which is shown in Table 2. The notation "-" means having no interest at all. Then we compute the matching results of several existing work.

The best match of *Alice* in FindU [8] is *Charles*, cause they have 5 common attributes. Algorithm in [9] considers the difference of priorities on each common attributes, as a result, it is hard to choose the best candidate from *Bob* and *David*, since the differences on these attributes are: *Alice* to *Bob*, $[1, 4, 1, 3, 2]$; and *Alice* to *David*, $[1, 4, 1, 3, 2]$. While in *P-match*, the similarities with the nearby users are 0.9667, 0.3972, 0.8243, 0.2316, and 0.9870, respectively. Definitely, we prefer *Frank* as the best match. Because either *Alice* or *Frank* is more interested in *Movie* and *Music* in their common attributes, even though they have only these two common attributes. We also compute the similarity values of *P-match*$^+$, the results are 0.1122, 0.0905, 0.1138, 0.0547 and 0.1314, respectively. The best match is still *Frank*. Similarly, the matching result in *E-match* can also be computed by the priority-aware Ochiai similarity coefficient with a probability determined by λ and l, l', then the matching results are same with *P-match*$^+$. For instance, if we let $\lambda = 400$, $l = 12$, $l' = 11$, *E-match* chooses *Frank* as the best candidate for *Alice*, and computes the exact similarity coefficients with the probability of 0.85.

Remark. The range of *Bob* and *David* in the experiment above is different in *P-match* and *P-match*$^+$. This is because we consider the intersection of every two users' attributes as the sample in *P-match* and the union in *P-match*$^+$. That means, Tanimoto similarity ignores the number of the common attributes and it only computes the similarity of the priorities on the same attributes. Meanwhile, Ochiai similarity computes the number of the common attributes and the priorities simultaneously in the union attributes set. In the experiment above, *Bob* and *Alice* have 2 common attributes while *David* and *Alice* have 3 common attributes, but the differences on common attributes are the same. So *Bob* is before *David* in *P-match* and after *David* in *P-match*$^+$.

6.2 Complexity Analysis

To discuss the complexity of our schemes, we analyze the online/offline computation overhead and the communication cost from both the *initiator* and *responder* sides. The computation cost is measured by counting the keyed hash functions and exponentiation operations, since these operations are always resource-consuming in mobile devices. h represents a keyed hash function, such as SHA-256 or SHA-512, while mul_1, $exp_1$and exp_2 means 1024-bit multiplication, 1024-bit and 2048-bit exponentiation operations, respectively. The communication cost is evaluated by computing the transmitted and received bits. We compare our work with algorithms in [20] and [24], since the former algorithm considers the malicious behavior in private matching as our work, and [24] tries to offload the computation overhead in existing secure two-party computation, which is well used in secure private matching problem. We assume that there are several mobile users in vicinity, and each user holds m attributes, where every attribute has a priority value from $[1, \kappa]$. Table 3 shows the theoretic analysis in details. Our schemes have lower computation cost, especially the online parts.

6.3 Experiment Setup

To study the feasibility of our algorithms, we first evaluate the time taken for generating SHA-256, SHA-512, exp_1, exp_2,1024-bit and 2048-bit safe primes, respectively. We then implement our proposed algorithms on a Thinkpad laptop (the cryptography library is Crypto++) with 1.82 GHz CPU, 4 GB RAM, and Windows 7 32-bit Professional to simulate the performance, which is used for the offline computation. We also implement our schemes on two SAMSUNG Nexus S smartphones with 1 GHz Cortex-A8 processor, 512MB RAM, Android v2.3.6, and Bluetooth v2.1. Each result in our experiments is an average of 1000 runs.

6.4 Experiment Results

Table 4 and 5 show the mean, maximum, minimum, medium and standard deviation of time consumption of SHA-256, exp_1, exp_2 and generating safe primes with 1024-bit and 2048-bit, respectively. We can clearly see the better performance provided by the laptop since the powerful computing capability. For example, when we generate a 1024-bit safe prime, averagely, it needs to consume 156.37 ms on the laptop and 582.28 ms on the Nexus S, respectively. Fortunately, it needs not to be worried in our work since we can put these work in the offline computation phase.

TABLE 4: Time consumption of different operations on laptop

Operation	Mean	Max	Min	Median	Std
SHA-256 (μs)	2.13	2.5	2.1	2.1	0.048
SHA-512 (μs)	10.6	14	10.5	10.6	0.21
exp_1 (μs)	340.6	483	338	339	7.13
exp_2 (μs)	756.8	986	752	755	12.69
Prime-1024 (ms)	156.37	178	134	156	12.60
Prime-2048 (ms)	1545.27	1663	1413	1546	74.26

TABLE 5: Time consumption of different operations on Nexus S (ms)

Operation	Mean	Max	Min	Median	Std
SHA-256	18.79	20	18	19	0.83
SHA-512	22.17	24	21	22	1.10
exp_1	39.17	75	20	39	9.05
exp_2	59.94	110	35	58	15.31
Prime-1024	582.28	650	525	582	37.52
Prime-2048	7090.46	7175	6518	6892	219.36

We then show some evaluation results on the offline/online computation cost, communication cost and the execution time, respectively. Specifically, the offline communication cost means the operations which can be pre-computed without supply from other entities. The online computation cost represents the operations that to be computed in real time. The communication cost indicates the transmitted data in bits and the execution time stands for the total time consumption to perform a private matching procedure between users, including both the online computation cost and the data transmitting time between users.

6.4.1 The impact of m

Fig. 4 shows the evaluation results ($\kappa = 10$) of the impact of varying m. We first test the offline computation cost when the number of attribute m is changing from 20 to 200. Fig. 4a indicates that both of *P-match* and *E-match* have better performance than existing work [20], [24]. The computation cost in [20] is high since there are too many exp_1s employed in their schemes. *E-match* outperforms than *P-match* by the reason of the utilization of *poly*$^-$s instead of the exp_1s. The offline cost can be computed before the regular computations, so this part does not impact the execution time.

Fig. 4b compares the online computation cost of all the protocols in the $\log 10$ scale for varying m. It makes

TABLE 3: Comparison of Matching Algorithms

Protocols	Party	Offline Comp.	Online Comp.	Comm.trans (in bits)
[20]	Initiator	$(2m+2m^2)exp_1$, $(2m)h$	$(m+m^2)exp_1$, $(m)h$	$3m \cdot 1024$
	Responder	$(m+m^2)exp_1$, $(2m)h$	$(2m)exp_1$	$4m \cdot 1024$
[24]	Initiator	$(2rm)exp_1$, $(rm)exp_2$	$(rm)exp_1$, $(2rm)exp_2$	$rm \cdot 2048$
	Responder	—	$(2rm+1)exp_1$, $(2rm+1)exp_2$	$rm \cdot 2048$
P-match	Initiator	$(2m+1)exp_1$, $(m)h$	$(2m)exp_1$	$4m \cdot 1024$
	Responder	$(2m+1)exp_1$, $(m)h$	$(3m)exp_1$	$2m \cdot 1024$
E-match	Initiator	$(2m)h$, $1poly^+$	—	1024
	Responder	$(rm)h$, $((r-1)m)mul_1$	$(rm)poly^-$	32

sense that this part is sensitive to the execution time, so we aim to offload the online computation to offline as much as possible. We can clearly see the efficiency of our protocols over others. For users in *P-match*, they just need to perform $2m$ exp_1s on the *initiator* side and $3m$ exp_1s on the *responder* side. While in *E-match*, it decreases the computation cost significantly by validating the polynomial with several potential solutions. The online cost of the protocols in [20], [24] are much higher since they utilize several exp_1s and exp_2s in their processes.

Fig. 4c shows the communication cost between entities. Not surprisingly, the result of each protocol increases smoothly with the increasing m. Our *E-match* shows a better performance on the communication cost through replacing the complicated exchanging phases between entities with a single polynomial. The *initiator* only needs to transmit some simple parameters of a specific polynomial to exchange both the attributes and the corresponding priorities. For example, *E-match* needs to consume 50.28Kb on bandwidth when $m = 100$. This is easy for Bluetooth v2.1, since the transmission rate can achieve approximately 900Kb/s in our experiments, which means that we only need to spend 55.87 milliseconds to finish all the transmissions.

Fig. 4d provides the total execution time of all the algorithms. The execution time in our experiments mainly includes the online computation time and the information transmission time. Comparing with the schemes in existing work [20], [24], our *P-match*, *P-match*$^+$ and *E-match* degrade the execution time. When we look into our protocols more specifically, to get the common attributes securely, an *initiator* needs more time to complete the computation in *P-match* and *P-match*$^+$. However, it is obvious that all of our proposed protocols can be finished within about 600ms in all simulated sceneries. For example, when the number of attributes $m = 200$, protocols in [24] and [20] need 20.52 and 20.66 seconds to complete the matching phase for each user, while our *P-match* and *P-match*$^+$ require 0.33 and 0.42 seconds, respectively. This value in our *E-match* is 197.86 milliseconds, which is more practical for mobile users.

6.4.2 The impact of κ

In Fig. 5, we show some evaluation results ($m = 100$) of the impact of the varying κ. Fig. 5a indicates the changes on the offline computation cost with the varying κ. The evaluation results show that our *P-match* and *E-match* outperform other schemes in all the tested κs. [20] and *P-match* are steady in terms of various κ. Meanwhile, [24] and *E-match* are impacted by the changing κ. The reason is that, [24] needs extra exp_2s to transform ℓ_1 distance into ℓ_2 distance by performing Johnson-Lindenstrauss embedding, and*E-match* need to re-build the polynomial $f(x)$ by computing each possible priority value on the *responder* side.

Fig. 5b shows the offline computation cost of the schemes. The performance of [24] is the worst one when $\kappa \geq 8$, due to a number of utilizations on the exp_2s while less on the exp_1s. *E-match* outperforms than all other schemes since it did not employ the heavy operations such as exp_1s and exp_2s.

Fig. 5c demonstrates the general trends of the communication cost of different schemes. We can see [20] and *P-match* are stable with increasing κ. For example, when $m = 100$, the communication cost are 0.68 Mb in [20] and 0.59 Mb in *P-match*. However, in [24], it is heavily impacted by the increasing κ, the communication cost exceeds 3.91 Mb when $\kappa = 10$. This situation is changed a lot in our proposed *E-match*, it is not stable with the increasing κ, nevertheless, it does not change too much, for example, the communication cost is about 50.28 Kb when $\kappa = 10$, which is very comfortable for Bluetooth v2.1.

In Fig. 5d, the evaluation results clearly show the advantages of our proposed schemes. For instance, when $\kappa = 10$, compared with 5.40 seconds in [20] and 9.54 seconds in [24], our *P-match* and *P-match*$^+$ need 161.54 and 217.81 milliseconds, respectively. And in our *E-match*, the execution time is only 29.01 milliseconds to achieve the same goal.

6.4.3 Implementation Results on SAMSUNG Nexus S Smartphones

To validate the usability of our proposed protocols, we implement them on SAMSUNG Nexus S smartphones to test the performance. Fig. 6 shows some selected results on smartphones, which may be a litter different from the simulation results on laptop, the reason can be found in Table 4 and 5. Generally speaking, the main differences between them are the online computation cost and the execution time.

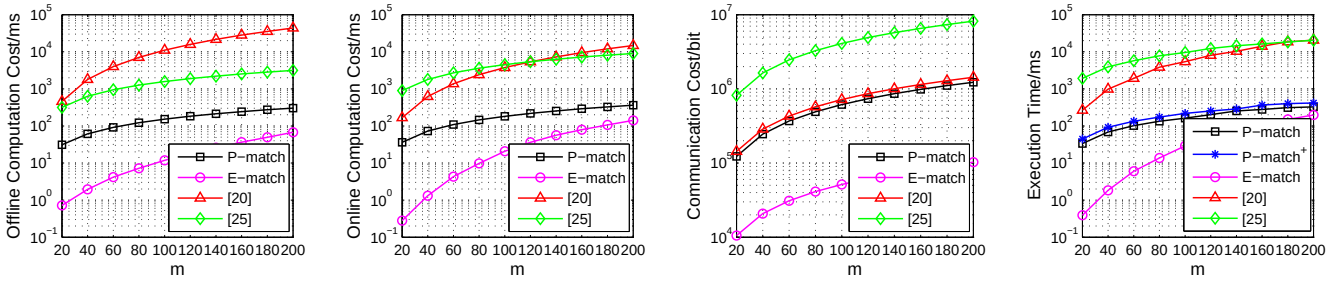


Fig. 4: Impact of the number of attributes m on (a): offline computation cost; (b): online computation cost; (c): communication cost; (d): the execution time, $\kappa = 10$

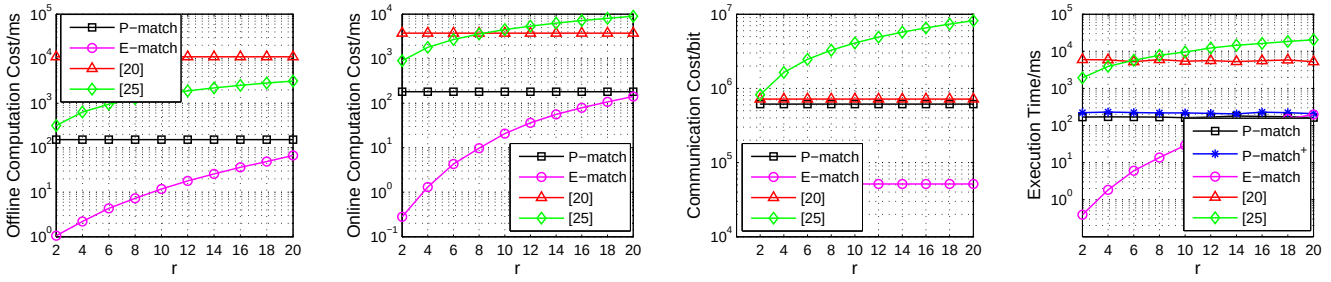


Fig. 5: Impact of the upper bound of priorities κ on (a): offline computation cost; (b): online computation cost; (c): communication cost; (d): the execution time, $m = 100$

Fig. 6a shows the online computation cost on the Nexus S with the varying m. The performance of [20] and [24] are quite similar with each other, however, they need several tens of seconds or more to do the online commination, which cannot be accepted by mobile users. Our proposed *P-match* and *E-match* perform better than others. Specifically, *E-match* is the best since it only needs to verify the polynomial by several possible results, which is quite simple for modern smartphones.

In Fig. 6b, we show the changes on the execution time of different schemes. Generally, the execution time is heavily related with the online computation time, and it increases with the increasing m. Not surprisingly, the performance of our *E-match* is much better than others.

Next, Fig. 6c indicates the online computation cost on the Nexus S with the changing κ when m is set to 100. Similar with the evaluation results of online computation cost on the laptop, the performance of [20] and [24] are unacceptable for mobile users since they need several minutes to finish the matching phase. Our *P-match* cannot be affected by the changing κ, however, it still needs 15.67 seconds to do the matching. While in our *E-match*, it is lightweight and only consumes 103.6 millisecond to complete the same phase when $\kappa = 10$.

In Fig. 6d, the changes on the varying κ bring less effect on the execution time for the smartphones in different schemes. Our *E-match* outperforms than others significantly. For instance, when $\kappa = 10$, the execution time of [20] and [24] are 584.46 and 757.60 seconds, respectively. The results of our *P-match* and *P-match*$^+$ are 17.41 and 23.47 seconds, respectively. While in our *E-match*, it only needs 186.48 millisecond to process the matching with others, which is absolutely efficient.

6.5 Energy Consumption

We also compute and compare the energy consumption of our scheme with others. The most energy consumed operations for modern smartphones are local computation, display and network transmission [25]. In our work, since we did not use heavy graphics, we pay much attention on two main factors, local computation cost and network transmission cost. We use the energy consumption model [26] $E_{computing} = P_{comp} \cdot T_{comp} + 0.3167T_{run}$ to estimate local computation cost, where P_{comp} represents the CPUs power consumption, T_{comp} means the time spent for computation and T_{run} indicates the total protocol run time. For a smartphone with 1 GHz CPU, we choose $P_{comp} \approx 0.38w$ [25]. The energy consumption model of the network transmission cost is based on [27]: $E_{network} = n_t \cdot E_t + n_r \cdot E_r$, where n_t and n_r are the transmitted and received data in bytes, and $E_t \approx 4.8\mu J$ is transmitting energy per byte, $E_r \approx 6.7\mu J$ is the receiving energy per byte. For simplicity, we omit the initial connection establishment energy since it is common in all schemes. Then our energy consumption model can be denoted as:

$$\begin{aligned} E &= E_{computing} + E_{network} \\ &= P_{comp}T_{comp} + 0.3167T_{run} + n_tE_t + n_rE_r. \end{aligned} \quad (24)$$

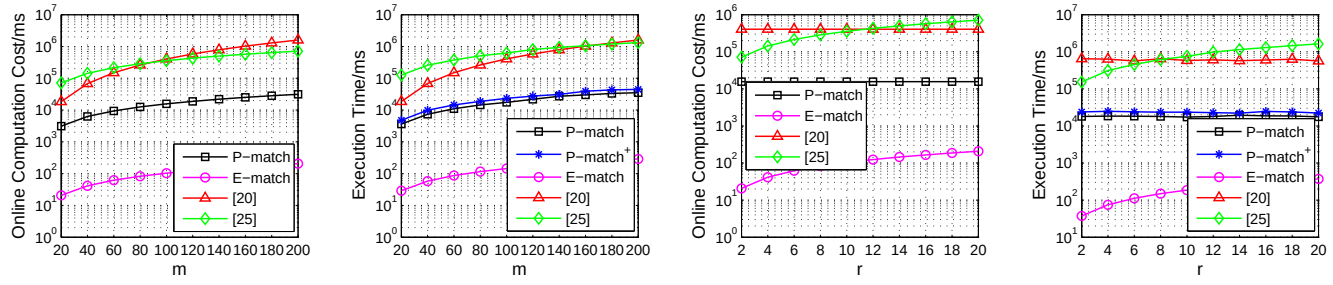


Fig. 6: Impact of number of attributes m on (a): online computation cost; (b): the execution time, $\kappa = 10$, impact of the priorities κ on (c): online computation cost; (d): the execution time, $m = 100$

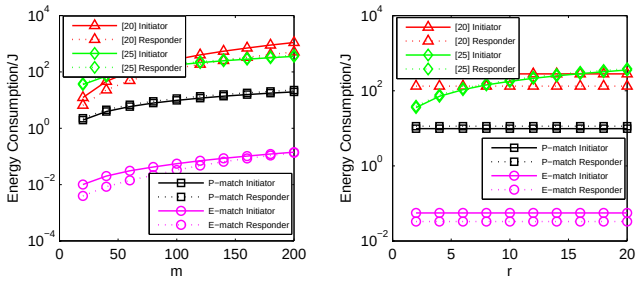


Fig. 7: Energy consumption

Fig. 7a shows the comparison of energy consumption on each side of the protocols. It is clear that our protocols consume less power than other schemes. For example, when the number of attributes $m = 100$ and $\kappa = 10$, the *initiator*s in the protocols [20], [24] need to consume $280.18J$ and $178.00J$, respectively. While in our *P-match*, it consumes lower energy of $9.81J$ to achieve the same goal. Finally, in our efficient version *E-match*, it degrades the energy consumption from both the computation and network transmission aspects, as a result, it consumes $55.60mJ$ and $33.00mJ$ on the *initiator* and the *responder* sides, respectively. We can conclude that our protocols are very practical in terms of energy consumption.

Fig. 7b indicates the impact of κ on the energy consumption, the number of attributes is fixed to 100. Since the similar result with Fig. 7a, we only point out that the energy consumption of our *E-match* is increasing slowly with the varying κ.

7 Conclusion

We propose a Priority-aware Private Matching problem to satisfy the requirements of our real social life for the first time. Comparing to existing work, the matching processes of our *P-match*s consider both the number of common attributes and the corresponding priorities. To avoid possible attacks from both *initiator* and *responder*, we then construct a priority-aware Ochiai similarity function in our enhanced version. Finally, an efficient version called *E-match* is also proposed to decrease the cost. The followed security analysis and performance evaluation show the correctness and efficiency of our algorithms. Our future work is to deploy our *P-match* and *E-match* into a large scale of real mobile environment to test the performance.

Acknowledgment

The preliminary work is accepted by IEEE MASS 2013 [28]. This work was supported by National Natural Science Foundation of China under Grant 61272457.

References

[1] E. Noah. (2011, Nov.) Mobile social networking shows promise, but rich media has higher engagement. [Online]. Available: http://www.emarketer.com/Articles

[2] N. Eagle and A. Pentland, "Social serendipity: mobilizing social software," *Pervasive Computing, IEEE*, vol. 4, no. 2, pp. 28 – 34, jan.-march 2005.

[3] L. P. Cox, A. Dalton, and V. Marupadi, "Smokescreen: flexible privacy controls for presence-sharing," in *Proc. of ACM MobiSys 2007*, pp. 233–245.

[4] J. Manweiler, R. Scudellari, and L. P. Cox, "Smile: encounter-based trust for mobile social services," in *Proc. of ACM CCS 2009*, pp. 246–255.

[5] A.-K. Pietiläinen, E. Oliver, J. LeBrun, G. Varghese, and C. Diot, "Mobiclique: middleware for mobile social networking," in *ACM workshop on Online social networks 2009*, pp. 49–54.

[6] W. Dong, V. Dave, L. Qiu, and Y. Zhang, "Secure friend discovery in mobile social networks," in *Proc. of IEEE INFOCOM 2011*, pp. 1647 –1655.

[7] R. Lu, X. Lin, X. Liang, and X. Shen, "A secure handshake scheme with symptoms-matching for mhealthcare social network," *Mobile Networks and Applications*, vol. 16, no. 6, pp. 683–694, Dec. 2011.

[8] M. Li, N. Cao, S. Yu, and W. Lou, "Findu: Privacy-preserving personal profile matching in mobile social networks," in *Proc. of IEEE INFOCOM 2011*, pp. 2435 –2443.

[9] R. Zhang, Y. Zhang, J. Sun, and G. Yan, "Fine-grained private matching for proximity-based mobile social networking," in *Proc. of IEEE INFOCOM 2012*, pp. 1969 –1977.

[10] M. J. Freedman, K. Nissim, and B. Pinkas, "Efficient Private Matching and Set Intersection," in *Proc. of LNCS EUROCRYPT 2004*, pp. 1–19.

[11] L. Kissner and D. Song, "Privacy-preserving set operations," in *Proc. of CRYPTO 2005*, pp. 241–257.

[12] Y. Sang and H. Shen, "Efficient and secure protocols for privacy-preserving set operations," *ACM Trans. Inf. Syst. Secur.*, vol. 13, no. 1, pp. 9:1–9:35, Nov. 2009.

[13] G. Ateniese, E. De Cristofaro, and G. Tsudik, "(if) size matters: size-hiding private set intersection," in *Proc. of ACM PKC 2011*.

[14] Z. Yang, B. Zhang, J. Dai, A. Champion, D. Xuan, and D. Li, "E-smalltalker: A distributed mobile system for social networking in physical proximity," in *Proc. of IEEE ICDCS 2010*, pp. 468 –477.
[15] X. Liang, X. Li, K. Zhang, R. Lu, X. Lin, and X. Shen, "Fully anonymous profile matching in mobile social networks," *Selected Areas in Communications, IEEE Journal on*, to appear.
[16] J. Manweiler, R. Scudellari, Z. Cancio, and L. P. Cox, "We saw each other on the subway: secure, anonymous proximity-based missed connections," in *Proc. of ACM HotMobile, 2009*.
[17] R. Agrawal, A. Evfimievski, and R. Srikant, "Information sharing across private databases," in *Proc. of ACM SIGMOD 2003*, pp. 86–97.
[18] J. Vaidya and C. Clifton, "Secure set intersection cardinality with application to association rule mining," *J. Comput. Secur.*, vol. 13, no. 4, pp. 593–622, Jul. 2005.
[19] M. von Arb, M. Bader, M. Kuhn, and R. Wattenhofer, "Veneta: Serverless friend-of-friend detection in mobile social networking," in *Proc. of IEEE WIMOB,*.
[20] E. D. Cristofaro, J. Kim, and G. Tsudik, "Linear-complexity private set intersection protocols secure in malicious model," in *Proc. of LNCS ASIACRYPT 2010*, pp. 213–231.
[21] A. C. Yao, "Protocols for secure computations," in *Proc. of IEEE SFCS 1982*.
[22] A. Ochiai, "Zoogeographical studies on the soleoid fishes found japan and its neighboring regions. ii," *Bull. Jap. Soc. sci. Fish.*, vol. 22, no. 9, pp. 526–530, 1957.
[23] L. Alan.H., "A proof of the triangle inequality for the tanimoto distance," *J. Math. Chem.*, vol. 26, pp. 263–265, 1999.
[24] S. Rane, W. Sun, and A. Vetro, "Privacy-preserving approximation of l1 distance for multimedia applications," in *Proc. of IEEE ICME 2010*, pp. 492 –497.
[25] R. Mittal, A. Kansal, and R. Chandra, "Empowering developers to estimate app energy consumption," in *Proc. of ACM Mobicom 2012*, pp. 317–328.
[26] A. Carroll and G. Heiser, "An analysis of power consumption in a smartphone," in *Proc. of ACM USENIXATC 2010*, pp. 21–21.
[27] A. Rahmati and L. Zhong, "Context-for-wireless: context-sensitive energy-efficient wireless data transfer," in *Proc. of ACM MobiSys 2007*, pp. 165–178.
[28] B. Niu, X. Zhu, T. Zhang, H. Chi, and H. Li, "P-match: Priority-aware friend discovery for proximity-based mobile social networks," in *Proc. of IEEE MASS 2013*.