Profile Privacy Aware Framework for Static objects Participatory Sensing

Dorothy Kalui
School of Computer and Communication Engineering
University of Science and Technology Beijing
and
Meru University of Science and Technology- Kenya

Dezheng Zhang
School of Computer and Communication Engineering
University of Science and Technology Beijing

ABSTRACT

Mobile devices users in Participatory Sensing Systems (PSS) are required to collect information from their nearby data collection points (DCs). A query normally reveals the identity (id), location, and user profile (e.g., race domain). This information facilitates an adversary PS server to infer over time a comprehensive user location summary with a high degree of precision. Some privacy techniques in PSS have been suggested recently to provide user privacy protection. However, only a few techniques that consider trust in static objects but disregard profile information. For credibility of data, there is scarcely any service, which entails the user to prove that she is at a particular DC point at a certain time. Yet none of the position and time information achieved by nowadays mobile devices is reliable. In this paper, we propose an enhanced K-location privacy-aware framework for static objects in PS system. The experimental results demonstrate in our approach user a high degree of anonymity and reliability of collected data.

General Terms

Participatory Sensing, Entropy

Keywords

Privacy, Anonymity, location attacks, profile, Visibility

1. INTRODUCTION

Location-based services have become popular due to the growing use of mobile devices. User's exact location is necessary for high quality of services which brings the risk of being linked to identify the query issuer. A campaign based PS systems require coordinated efforts of participants to collect useful data and avoid unattended data collection points (DCs). Besides, PS applications rely completely on the users readiness to participate and submit to the system correct and up-to-date data. Each user capture sensed data from their assigned DC using a variety of sensors devices such as GPS, smart phones embedded in their own devices. Users share their collected data with a PS server, which processes the received data to monitor, or analyze some incidents or phenomena of common interest. For example, the GPS data gathered from people as they go about their daily lives provide understanding into public transportation system while individualized sensing offers sampling of happenings as experienced by users e.g., individualized medicine[23, 3]. A mobile user queries the PS system to know DCs that is assigned to him. After collecting required data, the user upload the data to a PS server in real time to be transmitted to appropriate consumers. However, the privacy issue is that the user's identity may be guessed by adversary, if the adversary combine the location information with the profile (i.e location and profile quasi identifiers). Further, trustworthiness is a concern due to users experience and interest in data contribution exercise[16, 10, 12, 25, 6]. For example a dishonest user can alter the observed data to report misleading incident or a malicious user report forged data. The privacy and honest issues are the key barrier to deploy the PS system successfully.

Table 1. Illustrating Identification Probability's possibilities in a cloaked region

ID	ASR Generalised Identification Prob	unique Prob
A	$\frac{2}{5},\frac{2}{5},\frac{1}{5}$	Yes ($\frac{1}{5}$)
B	$\frac{4}{5},\frac{1}{5}$	Yes ($\frac{1}{5}$)
C	$\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5}$	No
D	$\frac{2}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5}$	No
E	$\frac{3}{5},\frac{1}{5},\frac{1}{5}$	No

To protect the mobile user privacy, [22, 26, 15] considers the user profile in centralized environment. However, our PS system implements a P2P architecture. [11, 20, 12, 9, 10], use the location k-anonymization technique i.e., Partial Inclusivity Range Independence (PiRi) that cloaks a user among k-1 other users. Consequently, each query region elect a group leader to submit the query on their behalf. While [9, 10] each user state a location privacy condition that include profile as an element, the fact that user broadcast for peers with identical profile result to large size ASR or experience high waiting time. In their research authors [11, 20, 12] exclude user profiles. Yet an adversary server can easily obtain profile of a mobile user by watching him, if the user profile is quite unique from the rest in his ASR. An identification probability is unique if it is the only one with a probability distribution of $\frac{1}{K}$ in the ASR, table 1. This makes the user associated with the profile highly visible from the rest. Although existing work investigated interplay of trust and privacy of participants, privacy breaches of PS systems still remain as open challenges.

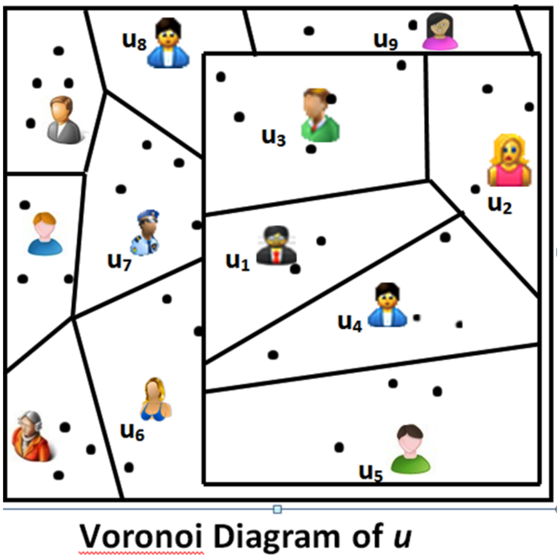

Fig. 1. K Location ASR

For example, in figure 1 given $K = 5$ at k=1, we consider users proximity of u_1, u_2, u_3, u_4, u_5. Note that the collection points are shown as black points and are assigned to each user already. When the cloaked region, i.e., the rectangle includes the five users is sent to the server together with the users profile, i.e., gender, the adversary can easily identify mobile user $u2$ by associating the profile knowledge *female* with the location information. Here, the probability distribution is $\frac{4}{5}$ male vs $\frac{1}{5}$ female. In this scenario, the K-anonymity cannot be guaranteed. We call such attack as *location attacks* since the adversary can identify a participant from other local peers in a cloaked region by knowing the location information.

More recently, trust systems have been deployed in participatory sensing. Several authors seek to estimate the trustworthiness of users' activities in order to assess the quality of their collected data. Existing trust-based frameworks [25, 6, 12] are not protected from location attacks based on visibility of user profile. With few [10, 9] that considers user profile assuming uniform probability distribution, giving rise to larger cloaked size and waiting time issues. The works [26, 15] study the location-based privacy issues considering the profile. However, the centralized techniques proposed cannot support the PS system.

In this paper, we propose enhanced K location approach, to support a typical PS systems called TAPAS [12]. Intuitively, mobile users with the lowest identification probabilities have a high degree of anonymity and a low visibility. [9, 10] studied PS system of moving users that allocates each DC point to one user and multiple users respectively. In the work [9], focus on uniform probability of location and profile information. Cloaked region are formed by users according to proximity and identical profile. [10] tackle trust by allocating multiple DCs to moving users with uniform probability, a peer broadcast to get K-1 users considering proximity and identical profile. This is similar idea in this paper, however differ with this work in that, cloaked region are formed by users based on proximity then thereafter, where applicable remove any user with unique profile information. In both [9, 10] a peer broadcast message is in the format [peer identity, profile choice, K anonymity, minimum area plus other personalized parameters] while in this paper, message format is [peer identity, K anonymity, minimum area]

The disadvantage of anonymized spatial region (ASR) that considers proximity and uniform probability of static user profile, is its large size that compromise quality of service with many negative hits in the candidate answer set. In [9, 10] authors deal with dynamic requests over static data objects, while in this paper our scope is limited to static requests over static data objects, making allocation of DC point one time event that is efficient to implement.

The main problem we try to focus is how to limit the visibility of the profile while ensuring both quality service and validity of collected data in a cloaked region that deal with static requests. Our approach enhanced K location is supported by PiRi technique and the fact that we retrieve DC points for cloaked region rather than individual user. With PiRi technique, only one user who queries for DC points and distribute accordingly to the other local peers. This means only the query forwarder is critical once a cloaked region is formed. Motivated by this idea we argue that we can achieve privacy together with quality service to maximize the accuracy of the locations by minimizing their cloaked areas. In our approach, query region are based on proximity only then we can prune unique users in a cloaked region after election of group leader to control profile visibility. The pruning of users does not affect the quality of service since the query forwarder has correct position of the local peers in corresponding cloaked region. Our anonymization approach focus is to improve privacy and the size of cloaked area, smaller cloaked areas indicates the PS server returns smaller answer sets.

The rest of this paper is organized as follows. Section 2 reviews the related work. In Section 3, we introduce preliminary studies of this paper. Section 4, we explain our anonymization approach. Section 5, presents the experimental results, Finally, Section 6, we conclude our study and discuss the future directions.

2. RELATED WORK

The location information of mobile users can be used by an adversary to identify the users physically. However, the location is required by most systems. Privacy issues make users reluctance to contribute therefore diminish the effects and relevance of sensing campaigns deployed at large scale, as well as restricting the benefits to the users. To control the risk that a users privacy might be compromised, mechanisms to preserve user privacy are mandatory. The most popular technique to protect privacy is the spatial cloaking

Some recent techniques in distributed environments [4, 5, 18], utilize complicated data structures to anonymize users via fixed communication system (e.g., base stations). Such architecture are costly in updating if users enter or exit frequently. [2, 11, 20, 12] propose an unstructured peer-to-peer networks where users can cloak their locations in a region by communicating with their neighbor peers without relying on a shared data structure. In this paper, we utilize the P2P spatial cloaking techniques to protect user's identity when issuing a query to PS server.

Few works [11, 20, 12, 14, 7, 13], have studied privacy issues for the PS systems. In a PS system, there are two phases namely, data contribution phase and the coordination phase.[14, 7, 13], discuss privacy issue in uploading the collected data to the server without revealing the identities of the users. Whereas [11, 20, 12] focus on how to secretly assign a set of data collection points to each participant. [11, 20, 12], PiRi cloaking technique is employed to preserve user privacy.In this approach, Only the sample of the elected leaders i.e., monitors, send queries to the server on behalf of the other mobile users and it shares query region results with the users.

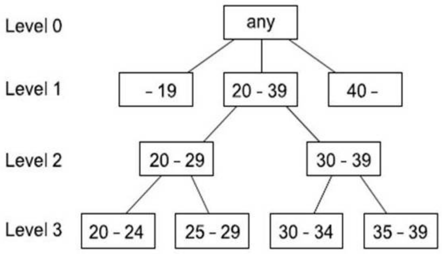

Fig. 2. Illustrating Taxonomy for profile domain "AGE"(Adapted from [15])

The studies[9, 10], the focus is uniform probability of location and profile anonymizaton. The ASRs are formed by users based on proximity and profile. [10] allocates multiple DC to moving users with uniform probability, a peer broadcast to get K-1 users considering proximity and identical profile. This is similar idea in this paper, however differ with this work in that, ASRs are formed by users based on proximity then thereafter, remove where applicable any user with unique profile information.

The disadvantage of ASR formed utilizing uniform probability especially for static requests is large sized cloaked region that compromise quality of service with many negative hits in the answer set. Second, a peer may fail to get enough users to meet their stated minimum anonymity levels. In [9, 10] authors deal with dynamic requests, while in this paper our scope is limited to static requests, making allocation of DC point one time event that is efficient to implement.

Some recent studies consider users' properties [22, 26, 15, 24, 10, 9] in a personalized privacy-protection mechanism. The idea of [26, 15], is to build a hierarchical taxonomy for each profile. Each user select the profile level in the hierarchy that correspond with her choice or value. The identification probability of a user is $\frac{1}{K}$ when there are K users in the query region. Figure 2 illustrate an example of taxonomy for Age domain. In the hierarchical tree for any cloaked region of users, we represent its subtree as all the nodes containing each user, with the highest node level being the root and the rest as leaves depending on the disclosure level.[15, 26] approach uses a simple data structure for anonymization that is efficient. The approaches,[22, 26, 15] use a centralized approach which is not appropriate for PS systems. Entropy metric is widely used to quantify the degree of anonymity [17, 8, 24, 9, 10]. Entropy-based approaches have been adopted [17, 8, 24] to achieve k-anonymity for users in LBS. [17] uses entropy in false locations. Users protect their location privacy by reporting their exact locations with a set of fake locations, termed dummies. We cannot use it directly for the PS system because of its expensive infrastructural requirements. [24] develops entropy based classification algorithm that uses personalized users attributes to measure anonymity. It is totally different from our work because in traditional data publishing an adversary can simply match users quasi-identifiers and their external information (e.g.telephone directory), and track them. It contradicts our problem because visibility of the profile is an important threat to users' privacy. [9, 10] employ entropy in dynamic requests were users state a privacy requirement policy making all users in ASR have identical profile resulting to uniform probability distribution. However, some of the techniques employ pseudonyms to provide privacy which not be enough.

In [12, 10], authors introduce redundancy of data collected to improve data reliability by positioning more than one user in each DC point. The authors ensure collected data is valid if the results shared by the users in each DC is the same. The work assume static requests and dynamic request [12, 10] respectively. In this report, we borrow the idea in [12, 10], our work differs with [10] first, in the way we anonymize profile, at broadcast stage, we only consider proximity then once cloaked region is formed, we prune non identical profile from the rest. Second, our work deal with static requests. In [12] the authors consider proximity only in anonymization while addressing static requests. This gives rise to privacy leaks caused by visible users profile in case unique from the rest in a cloaked region.

3. PRELIMINARIES

As discussed in Section 2, we start by using the P2P PiRi cloaking technique to address the privacy problem in participatory sensing.The idea of TAPAS approach (see [12, 20]) is based on the fact that the range queries sent to the PS server have duplication. A user communicates with his neighboring peers via multi-hop routing to find at least $K-1$ other peers. A distributed voting mechanism, is used to elect monitors where by the user who scores highest votes is declared the monitor. Therefore, instead of each user issuing a separate query, only a group of monitors ask queries from the server on behalf of the rest of users and share their results with those who have not send any query. The monitors are elected by majority votes by peers. Thus, the goal of the PiRi approach is to assign to each user those DC-points which are closer to that user than to any other participant, without revealing users identity.

This requires to compute secretly reverse k-nearest DC points to a participant in order to guarantee every DC point have k participants. Each user specify his own privacy level (i.e.,m) and area minimum (i.e.,A). After satisfying the m-anonymity requirement, the user extends the cloaked region to A to satisfy the minimum area privacy requirement. To incorporate a trust parameter k, every DC-point should be assigned to a minimum k participants. Here, the query is to retrieve DC points for an ASR rather than the individual peers. The monitors receive the query results from PS server, then using location information of local peers forward to each. However, [12] does not protect the users' privacy from the location attacks as shown in figure 1. In this paper we extend the work[12, 20] to our context to provide location privacy of mobile users.

A hierarchical taxonomy structure is used in the anonymization procedure[15], for example, Figure.2. It organizes users based on selected node for each profile. A node contain users with similar privacy specifications. New leaf nodes are created according to users specification if not matching existing nodes. Each user profile is represented independent in the hierarchical tree. The nodes are specified by disclosure level $(1, 2, \cdots, n)$.

DEFINITION 1. ***Profile Set***.
For a given ASR v, the profile set of v is defined as the set of profile value for all the its users i.e. $\mathcal{A}_v := \{A_{u_i} : u_i \in v\}$. For one profile case A_{u_i} is simply an element, and for multi profile case A_{u_i} is a tuple namely $A_{u_i} = (A_{u_{i_1}}, A_{(u_{i_2}}, \cdots, A_{u_{i_n}})$ $\forall i, j \quad i \neq j$ $A_{u_i} = A_{u_j}$ iff every component of A_{u_i} equals with that of A_{u_j}.

DEFINITION 2. ***Separation of a profile set.***
A Separation of a profile set $\mathcal{A}$ which is denoted $S(\mathcal{A})$ is set family of profile subset $\mathcal{A}_i s \subseteq \mathcal{A}$ where for each $\mathcal{A}_i$ all of its elements(namely profile values for users) are same, and for each different pair of i,j their elements must be different with each other. i.e.

$$S\mathcal{A} := \{\mathcal{A}_i : \mathcal{A}_i \subseteq \mathcal{A}\}$$

each element S of $\mathcal{A}_i$ are the same, $\forall i, j \quad i \neq j \quad A_i \neq A_j\}$. And obviously

$$|\mathcal{A}_v| = \sum_{i=1}^{|S(A)|} |\mathcal{A}_i|$$

.

DEFINITION 3. ***Some properties.***
Especially if $|S(\mathcal{A}_i)| = 1$, we declare that $\mathcal{A}_v$ is not separable,otherwise $\mathcal{A}_v$ is separable. It is declared that $S(\mathcal{A})$ is a fair separation iff

$$|\mathcal{A}_i| = |\mathcal{A}_j| \equiv C \qquad \forall i, j \quad i \neq j$$

Fair separation is when all ASRs have same number of K users.

EXAMPLE 1. ***Example 1.***
Consider the age profile, we have a ASR with 5 users, and the level of age of them being $\{1, 1, 0, 0, 2\}$ respectively, (refer figure 2). Then we have

$$\mathcal{A}_v = \{1,1,0,0,2\} \qquad S(\mathcal{A}_v) = \{(1,1),(0,0),2\}$$

.

DEFINITION 4. ***Entropy of a ASR v.***
To quantify the degree of anonymity, [21] define the measure of the uncertainty that an adversary can identify user by observing the users in the target area.
Given a ASR v the entropy of which is defined as the average entropy of all sets in

$$S(\mathcal{A})E(v) = \frac{1}{n} \sum_{\mathcal{A}_i \in S(\mathcal{A}_v)} \log |\mathcal{A}_i|, \qquad n := |S(\mathcal{A})|$$

DEFINITION 5. ***Anonymity.***
The anonymity refers to the ability of a mobile user collecting data and uploading it in real-time without being identified with probability greater than $\frac{1}{K}$, if there are K users.

DEFINITION 6. ***Visibility.***
A profile is visible if the degree of anonymity is very low such that an adversary can identify the user. An adversary can observe some of the users profile without them knowing. In Fig. 1 in which cloaked region formed by u_1, u_2, u_3, u_4, u_5 is observable by profile female.

3.1 System Model

Our architecture has two parts, users (data collectors) and PS server, integrated with a campaign administrator. An adversary can attack the PS system and obtain all the information stored in the system. Since the adversary can successfully obtain the profile of the users, it can distinguish the query issuer from other users who are also located in the anonymized query region by looking at the users.[7] since the users most of times send their queries from the same locations (office, home), which can be identified through physical observation, and triangulation(i.e., to determine user location point by measuring angles to it from known points at either end of a fixed baseline) and so on. For simple we consider the PS server is the adversary. The assignment of DC-points is an event that can be performed offline. There exists a pseudonymiser between the mobile users and the PS system to conceal the users' identity.

All the DC-points are stored in the PS server together with embeded privacy processor. The campaign-defined k value represent the number of users assigned to each DC-point. The k parameter indicate user confidence level where a DC point is assigned to $k > 1$ ASR. Since the adversary knows the anonymization procedure used in the system, it can identify out some users successfully from the other users in the set of ASR using their profile. Local peers share their locations. The anonymizer has confidence with the majority of the users but not trust in the PS.

4. OUR APPROACH

Local peers carry out elections of a leader among themselves followed by cloaking their locations among other $m - 1$ users. The leaders send their ASRs with different anonymity levels to the server. The system assigns the DC-points to the participants in each ASR.

A times users profile information vary with only one user being associated with a particular value unique from the rest, then its a privacy problem.

We consider anonymization solution, the optimal solution and the K location solution (i.e., PiRi cloaking technique). The optimal solution represents the ideal situation. In fig 3 further explain our basic idea. We use PiRi cloaking technique to provide users protection and submit query regions with users profile. However, since the PS server has profile of users, the achieved privacy level is much less. As a result, the entropy drops significantly labeled entropy before pruning. If we enhance the PiRi approach by pruning the unique users and then we measure again the entropy. By removing the unique users associated with high visibility in the ASRs, the entropy increase. The k location anonymized ASRs are left without observable profile thus profile anonymized. This make them attain higher anonymity levels. High levels of entropy indicate high degree of anonymity thus preferred. Pruning a user from an ASR does not interfere service quality. The PS server, its retrieves DC points according to the ASR. It does not retrieve DC points for each mobile user. *The server uses the location of an ASR to assign DC points to users.* In addition, the ASR query monitors know the actual position of each local peer and they receive results from the server. They send the corresponding results to each of its local peers.

To demonstrate the problem that we tackle, we include closeness in terms of Euclidean distance and pruning non identical user in assignment of DC-points as shown in Figure 4. The server allocates each user a group of DC-points closer to her than any other user by constructing of the Voronoi diagram of data collection (DC) points. Note that the collection points are shown as black points and are assigned to each user already. For example, a user state privacy requirement K=5, the cloaked region is divided into several five Voronoi cells, consisting of u_1, u_3, u_4, u_5 and one cell without user. The user in this cell has been pruned to avoid being identified by PS server with unique profile ((female) refer figure 1).

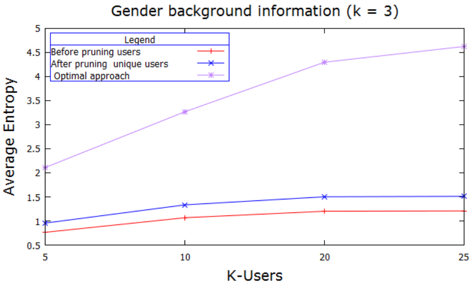

Fig. 3. PiRi technique vs Visibility, (gender profile)

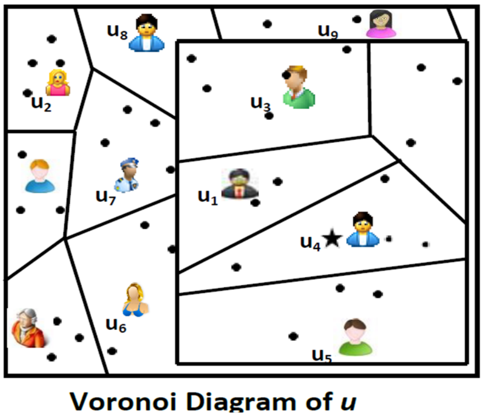

Fig. 4. Illustrating enhanced cloaked region in assignment of DC-points to the users

Table 2. Illustrating k location only (gender,age profile domain)

ASR	profile	location attacks
ASR1 u_1,u_2,u_3,u_4,u_5	[M],[F],[M],[M],[M]	(F)u_2
ASR2 u_2,u_3,u_4,u_5,u_8	[F],[M],[M],[M],[M]	(F)u_2
ASR3 u_1,u_2,u_3,u_4,u_5	[20],[24],[25],[39],[26]	(39)u_4
ASR4 u_2,u_3,u_4,u_5,u_8	[24],[25],[39],[26],[23]	(39)u_4

Our approach, enhanced location anonymity (enhanced LA) is based on pruning of unique users in location anonymized cloaked region to eliminate their profile visibility and thus, restrict location attacks, as in Figure 4.

Enhanced LA approach is based on two major phases:

Step1, Broadcast for peers based on proximity, then form query region that utilize closeness in terms of Euclidean distance only.

In table 4, when we anonymize location only, ASR1, ASR2 ASR3 and ASR4 have unique user. The unique users u_2 (female) and u_4 (age) are potential to location attacks respectively. If we submit the ASRs to the PS server, u_2 and u_4 are identified. These are users visible to an adversary by observation of their profile information, thus having low degree of anonymity.

Step 2, deals with enhancement of PiRi technique i.e profile anonymization. Here we prune any user in corresponding ASR associated with non identical profile information.

Table 3. Illustrating enhanced anonymized ASRs (gender,age profile domain

ASR	profile anonymized
ASR$_{bv1}$u$_1$,u_3,u_4,u_5	[M],[M],[M],[M]
ASR$_{bv2}$u$_3$,u_4,u_5,u_8	[M],[M],[M],[M]
ASR$_{bv3}$u$_1$,u_2,u_3,u_5	[20-24],[25-29],[20-24],[25-29]
ASR$_{bv4}$u$_2$,u_3,u_5,u_8	[20-24],[25-29],[25-29],[20-24]

On contrary, in table 4 ASR_{bv1}, ASR_{bv2}; ASR_{bv3} and ASR_{bv4} are anonymised then *enhanced* by removing the unique user. The submitted ASRs have no unique user. Therefore, *We achieve true K location anonymization by pruning unique users from each ASR if any*. If we enhance PiRi technique, two ASRs possibilities are the outcome, one, uniform probability distribution, second, non uniform probability distribution without unique user. The details of the profile anonymization approach can be found in [15], we extend to our context. For example to generalize users age profile, users select their preference from the nodes shown in fig. 2. Motivation for selecting the root or near root of profile is that: At the root node or near the root node the visibility is low. In other words, the degree an adversary can identify any user is minimum. At the root node or near the root node user identity is concealed. If the node is near the leaf level, the privacy can not be well protected.

Our aim is to achieve the maximum entropy, i.e., the highest uncertainty to identify a user in a cloaked region. An identification probability of a user $\frac{1}{K}$ when there are K users in the cloaked region result to maximum entropy $log|A_v|$. At maximum entropy the degree of anonymity is highest for each ASR users indicating minimum visibility by an adversary.

We take the following steps to restrict visibility of users profile to protect them from the location attacks. Our entropy based algorithm uses personalized users profile to measure anonymity.

Step 1:Compute sum of probability distribution of profile domain:

We aim to eliminate unique probability. If all the users have the same probability of being identified by an adversary; this ASR is reported. Otherwise, we go to the identification of unique users (step 2).

In Algorithm 1, line 1 shows this step.

Step 2: Identification of unique users in each ASR:

Line 2, to identify unique users in each ASR, we check the sum p_i of each profile domain. If theres probability distribution of $\frac{1}{k}$ non identical from the rest in the ASR, its a unique one. This means the user associated with that profile is different from the rest. The profile is visible to an adversary. If no unique user in ASR, then it is reported. Otherwise, we go to the pruning phase (step 3).

Step 3 :Pruning the users with the unique profile:.

Line 3 prune the users with the highest visibility profile. We prune the users with unique profile of $\frac{1}{k}$ because it is non identical from the rest in the ASR.

Pruning a user from an ASR does not interfere with service quality. The PS server, retrieves DC points according to the ASR. It does not retrieve DC points for each mobile user. The server uses the location of an ASR to assign DC points to users. In addition, the ASR query monitors know the actual position of each local peer and they receive results from the server. They send the corresponding results to each of its local peers.

Therefore, our ASR become profile anonymized inclusive. For example in fig.4 we show enhanced location anonymized ASR comprised of users u_1, u_3, u_4, u_5. Note that the collection points are shown as black points and are assigned to each user already.

Step 4: Compute the entropy of profile domain:

In line 4 we compute the entropy of ASRs that has users (using definition 4).

Our algorithm outputs the sets of ASRs highest entropy= E(v), line 5.

Algorithm 1: Enhanced K location anonymization

```
input  : Identification probabilities $p_i$, ASRs with sets of users
         supplied by anonymizer, privacy requirements $(m)$
output: Sets of ASRs maximum $E(v)$
begin
    repeat
        for each $v \in ASR$ do
1           Calculate the sum of probability for each profile domain
2           Discover the set of unique profile sets,
                $\mathcal{B} = \{\mathcal{A}_s : |\mathcal{A}_s| = 1, \mathcal{A}_s \in$
            S(A$_s$
3           prune all users in each set in B
4           Calculate the entropy $E(v)$using definition 4
5           Return the sets of ASRs E(v)
        end
    until No unique profile set could be found
end
```

Remark 1: For definition 3 if all of the profiles in $\mathcal{A}_v$ are the same, then $E(v)$ is degenerated to $\log |A_v|$, which we will show in theorem 7 that this is the ideal case for privacy.

THEOREM 7. *For step 1-3 in algorithm 1, their entropy is denoted $E(v)_1, E(v)_2, E(v)_3$ respectively then we have $E(v)_2 \leq E(v)_3 \leq E(v)_1$.*

For $E(v)_1$ and $E(v)_3$ the equality holds only when v is exactly the ideal case described in remark 1, while for $E(v)_2$ and $E(v)_3$ the equality holds only when removal does not triggered.

Proof step 1: E(v)$_1 = \log |\mathcal{A}_v|$

step 2 : $E(v)_2 = \frac{1}{n} \sum_{(\mathcal{A}_i \in S(\mathcal{A}_v))} \log |\mathcal{A}_i|$

step 3 : $E(v)_3 = \frac{1}{(n-l)} \sum_{(\mathcal{A}_i \in S(\mathcal{A}_v) \log |\mathcal{A}_i| \neq 0)} \log |\mathcal{A}_i|$

$= \frac{1}{(n-l)} \sum_{(\mathcal{A}_i \in S(\mathcal{A}_v)} \log |\mathcal{A}_i| = \frac{n}{(n-l)} E(v)_2 \geq E(v)_2 E(v)_3$

$= \frac{1}{(n-l)} \sum_{(\mathcal{A}_i \in S(\mathcal{A}_v) \log |\mathcal{A}_i| \neq 0)} \log |\mathcal{A}_i|$

$\leq \max_{(\mathcal{A}_i \in S(\mathcal{A}_v) \log |\mathcal{A}_i| \neq 0)} \log (|\mathcal{A}_i|) =$

$\log \left(\max_{(\mathcal{A}_i \in S(\mathcal{A}_v)} \log |\mathcal{A}_i| \neq 0) |\mathcal{A}_i| \right)$

$\leq \log |\mathcal{A}_v| = E(v)_1$

From (3)

$E(v)_3 = E(v)_2$ iff $n - l = n$ i.e. no removal is triggered.

From(4)

$E(v)_2 = E(v)_3$ iff $\max_{(\mathcal{A}_i \in S(\mathcal{A}_v) \log |\mathcal{A}_i| 0)} |\mathcal{A}_i| = |\mathcal{A}_v|$ which means $\mathcal{A}_v$ is not separable. And every elements in $\mathcal{A}_v$ are the same.

DEFINITION 8. ***Attack resistance***

Given a ASR v, it is concluded that v is attack resistant iff.

P($u_i \in v$) $= P(u_j \in v) \quad \forall i, j \quad u_i \in v, u_j \in v, u_i \neq u_j$.

$P(E) :=$ *the possibility E is to be discovered by an attacker.*

THEOREM 9. *v is attack resistant $\Leftrightarrow$ $S(\mathcal{A}_v)$ is fair.*

(1)$\Leftarrow$.

P(u$_i \in v | \mathcal{A}_(u_i) \in \mathcal{A}_i) = \frac{1}{(|A_i|)} \equiv \frac{1}{C} \equiv \frac{1}{|\mathcal{A}_j|} = P(u_i \in v | \mathcal{A}_(u_i) \in \mathcal{A}_i), \quad \forall i, j \quad u_i \in v, u_j \in v$. which implies equation (8)

(2)$\Rightarrow$.

P(u$_i \in v) - P(u_j \in v) = \frac{(|\mathcal{A}_i| - |\mathcal{A}_j|)}{|\mathcal{A}_i||\mathcal{A}_j|}, \quad u_i \in \mathcal{A}_1$

$P(u_i \in v) - P(u_j \in v) = 0 \Rightarrow |\mathcal{A}_i| = |\mathcal{A}_j|$.

Hence i, j are arbitrary chosen this then implies

$|\mathcal{A}_i| = |\mathcal{A}_j| \quad \forall i, j \quad u_i \in v, u_j \in v$.

THEOREM 10. *$\mathcal{A}_v$ is not separable $\Rightarrow$ $S(\mathcal{A}_v)$ is a fair separation.*

Proof:

This property obviously holds in that if $\mathcal{A}_v$ is not separable $S(\mathcal{A})$ only has one element, i.e. there is only one subset of profile Set where the profile value for users are the same.

THEOREM 11. *$\mathcal{A}_v$ is not separable $\Rightarrow$ $E(v)$ reaches its maximum.*

Proof:

If $\mathcal{A}_v$ is separable

{

$$E'(v) = \frac{1}{n} \sum_{(\mathcal{A}_i \in S(\mathcal{A}_v))} \log |\mathcal{A}_i| \leq \log(\max_{(\mathcal{A}_i \in S(\mathcal{A}_v) \log |\mathcal{A}_i| 0)}$$

—A$_i$|)} $\leq \log |\mathcal{A}_v| = E(v)$.

Corollary 1:

$\mathcal{A}_v$ is not separable $\Rightarrow$ v is attack resistant and v has the maximum extent of invisibility.

Remark 2

From Corollary 1 , it is concluded that our approach has one, the maximum of entropy which responds to a maximum extent of invisibility and every ASR v can resist attack in that every user in v shares same profile values. Two, though no user vulnerable to location attacks, the entropy is not maximum. This is because the $\mathcal{A}_v$ may be separable and $S(\mathcal{A}_v)$ may not be balanced.

Furthermore, if the effect of the diversity of profile is ignored and randomly choose users to form a ASR v then such v will not be attack resistant in that different profile value may reduce the entropy for S(A).

5. RESULTS

We evaluate our approach for different experimental setups. First we discuss our experimental settings. Next, we present our experimental results.

5.1 Experimental Settings

We perform three sets of experiments to evaluate the scalability and anonymity levels of our proposed entropy based algorithm. We use three performance measure, CPU time, communication cost and wastage in terms of excess DC point allocated to each user. We measure the performances of our approach with respect to varying numbers of users from 200 to 500 in terms of the CPU time.

In order to measure the anonymity level, we propose the entropy metric to evaluate how the PS server can associate the submitted queries to the query locations in PS system. The optimal solution is k anonymity which is maximum entropy, all user are represented by one uniform profile domain. The K location is the baseline approach which share the same idea as our proposed algorithm 1. They both allocate multiple DCs to each user.

In our experiments, the task of the PS system is to ask participants to gather set of photos from 200 locations (DC points) in Beijing (capital city of China).

In our experiments, the profile data is generated from the adult dataset (census information) [1]. We utilize profiles: age and gender. The first 500 records of this dataset is used to sample 500 profiles. The DC-points are randomly selected. The users' locations are generated randomly too. We set the default number of the participants to be 20 and vary the number from 20 to 500, usually a limited number of users take part in a PS campaigns [19, 11]. The degree of the anonymity m for each participant varies from 5 to 25, the default value is 5. We assign 2 to 5 participants to each DC point with a default, $k = 3$. We set the transmission range for queries to be 250 meters. We run 100 cases and report the average of the results.

5.2 Experimental Results

5.2.1 Changing scalability. The experiments for evaluating the scalability is carried out by varying the number users from 20 to 500 with $k = 3$. We report CPU time in figure 5. The figure shows the influence of the number of users on CPU time. The CPU time increases in all cases due to the number of profile information of each user. With many users, the algorithm costs much time in enhanced LA. Additional time is used in removing the users associated with unique profile information in the ASRs. Observe two opposing effects (i) with large of users the number of non identical users in ASRs reduce hence less time in pruning. (ii) More time is needed to process the many ASR since with large users high number of ASR is formed. The optimal approach and baseline performs better in all cases since no pruning involved.

In fig. 6 show Communication cost increase with more users. As the figure shows, the number of messages slightly increases in most cases. In a denser network, more communication is required among the peers to perform their queries. We observe the equal increase in enhanced LA, K-location. The optimal is slightly less than the other two approaches. In all the three approaches this communication overhead is due to the P2P communication for preserving the privacy. The cost of cloaking is the same, the slight difference is caused by the position of querying user who broadcast for other K-1 peers.

5.2.2 Changing anonymity levels. Figure 7 is the CPU time required. It shows the effect of increasing K when the number of users is fixed 500. As K increases, the cloaking cost increases, since more nodes have to be traversed to find enough users. It can be seen that performance deteriorates at higher K in all cases. With higher K, the three approaches take more time to find the users to satisfy stated anonymity levels. At bigger K, our enhanced LA performs worse. The removals increase since a user is pruned by one profile can trigger other user removal and this process became recursively.

Fig. 8 shows the impact of changing K on the communication overhead. The figure shows that the number of messages increases with an increase in K. This is because as K grows, more communication is required among the peers and sharing DCs in corresponding ASR. There is constant increase in both enhanced LA and K location because both employ similar peer search based on proximity and similar PiRi. The performance of enhanced LA and K location is similar in all cases.

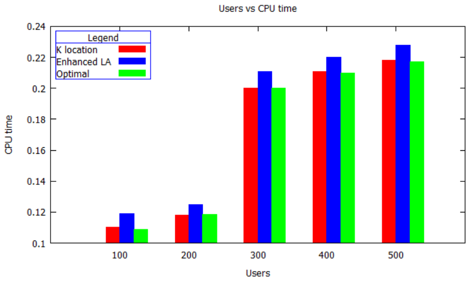

Fig. 5. CPU runtime

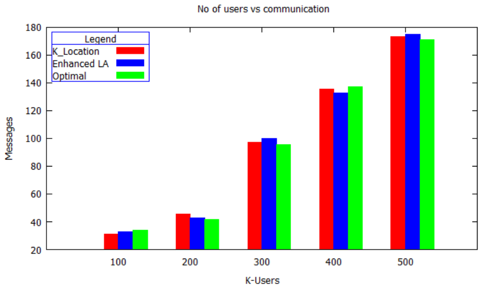

Fig. 6. Communication cost

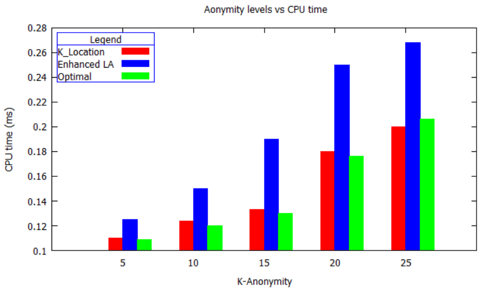

Fig. 7. CPU runtime

5.2.3 Privacy vs K-anonymity. In the experiments for privacy, we compute entropy with different privacy requirements from 5 to 25 and total users 500 (multi-profiles) shown in Figure 9. The entropy of optimal scheme is maximum in all cases. The performance of enhanced LA is better than K-location. For the K-location approach entropy drops significantly with multi-profiles. This is because in K-location (PiRi technique) there are more visible users some by gender and others by age. Since an adversary PS server may have

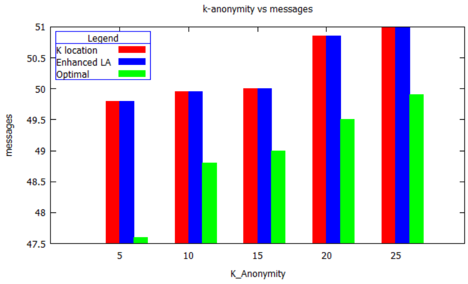

Fig. 8. Communication cost

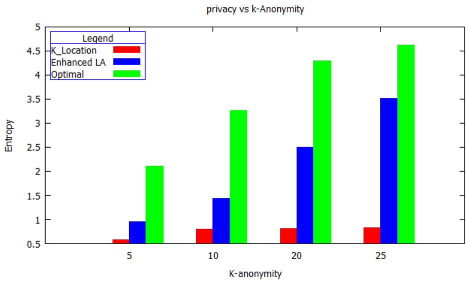

Fig. 9. Average Entropy for Multi-Profiles

such profile information, several unique users are identified. Thus, it is hard to efficiently assure the desired k-anonymity. The drop represents the number of users identified by an adversary in K location. For instance at K=10, approximately 3.2 to 0.7, a difference of 2.5.
The enhanced LA entropy increases with large k since the visible separable sets reduce. This can also be attributed to the fact that one user can be in several ASRs thus similar profile reducing visibility. The stronger entropy in enhanced LA approach is a result of none observable sets of users profile information. The performance of the optimal solution in all cases is always better since all the candidates users have the same identification probability to be identified as query issuer.

6. DISCUSSION

Our main observation from our experiments is that with slight increase in terms of CPU cost and no extra communication cost, enhanced LA can improve privacy levels by up to 2.5 times over K location method. Moreover, our experiments showed that with enhanced LA improves by increasing the anonymity levels (e.g., entropy increases to 4.6 with K = 25). However, as the number of mobile users grows, we see an increase in the of communication cost e.g., communication cost increases to 171 with 500 users in both enhanced LA and K location. This shows that our proposed enhanced LA approach performs better than K location in campaigns which have large number of users with higher privacy stated levels. Therefore, proposed approach appropriate for campaign based PS systems.

7. CONCLUSION AND FUTURE WORK

In this paper, we propose the new approach that uses entropy metric to quantify the anonymity degree in PS systems. We design enhanced LA privacy aware framework for the PS systems. With some additional computational cost, high privacy of static users is achievable. By incorporating multi profile domains of users makes the profile inclusive privacy anonymity approach more versatile. Our approaches can guarantee k anonymity all the times. In the future work, we aim to study query processor that can handle user movement in specific direction. This is to ensure privacy to mobile users driving different direction as well as allocating direction based data collection points.

8. REFERENCES

[1] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

[2] C. Y Chow, M. F. Mokbel, and Liu. Spatial cloaking for anonymous location-based services in mobile peer-to-peer, environments. *GIS.*, 15:351–380, 2009.

[3] Cisco and its affiliates. Global mobile data trafic forecast. *Cisco visual networking index*, 2013.

[4] C Cornelius, A Kapadia, D Kotz, D Peebles, M Shin, and N Triandopoulos. Anonysense: privacy-aware people-centric sensing. *MobiSys* ., pages 211–224, 2008.

[5] BCM Fung, K. Wang, R. Chen, and P.S Yu. Privacy-preserving data publishing: a survey of recent developments. *ACM Comput Surv*, 42(4):1–14, 2010.

[6] K. L. Huang, S. S. Kanhere, and W. Hu. On the need for a reputation system in mobile phone based sensing. *Ad Hoc Networks*, 12.

[7] KL Huang, SS Kanhere, and W Hu. Towards privacy-sensitive participatory sensing. *PERCOM*, pages 1–6, 2009.

[8] T. Jiang, H. J. Wang, and Y.-C. Hu. Preserving location privacy in wireless lans. *ACM MobiSys.*, page 246, 2007.

[9] D. Kalui, X. Guo, D. Zhang, Y. Xie, and Z. Yang. Personalized privacy aware framework for moving objects in participatory sensing. pages 191 – 198. IEEE Computer Society, 2015.

[10] D. Kalui, X. Guo, D. Zhang, Y. Xie, and X. Zhang. Trust assurance privacy aware framework for moving objects in participatory sensing. pages 246–251. IEEE, 2016.

[11] L Kazemi and C Shahabi. A privacy-aware framework for participatory sensing. *SIGKDD Explorations*, 13(1):43–51, 2011.

[12] L. Kazemi and C. Shahabi. Trustworthy privacy-aware participatory sensing. *Knowl Inf Syst*, 37(3):105– 127, 2012.

[13] K.Puttaswamy, R.Bhagwan, and VN.Padmanabhan. Anonygator: Privacy and integrity preserving data aggregation. *Middleware*, pages 763–774, 2010.

[14] Hu Ling and C Shahabi. Privacy assurance in mobile sensing networks: go beyond trusted servers. *PerCom Workshops*, pages 613–619, 2010.

[15] Masanori Mano, Xi Guo, Tingting Dong, and Yoshiharu Ishikawa. Privacy preservation for location based services based on attribute visibility. *VLDB CEUR Workshop Insta*, 908(4):33–41, 2006.

[16] Hayam Mousaa, Sonia Ben Mokhtara, Omar Hasana, Osama Younesb, Mohiy Hadhoudb, and Lionel Bruniea. Trust management and reputation systems in mobile participatory sensing applications:a survey. *Computer Networks the International Journal of Computer Telecommunications Networking*, 90.

[17] B Niu, Q Li, X Zhu, G Cao, and Hui Li. Achieving k-anonymity in privacy-aware location-based services. *Infocom.*, 978:799–3360., 2014.

[18] Amna Qureshi, David Megas, and Helena Rif-Pous. Framework for preserving security and privacy in peer-to-peer content distribution systems. *Expert Systems with Applications*, 42:1391 – 1408, 2015.

[19] F. Restuccia, S. K. Das, and J. Payton. Incentive mechanisms for participatory sensing: Survey and research challenges. *ACM Trans. Sen. Netw*, 12:1 – 38, 2016.

[20] C Shahabi and L Kazemii. Towards preserving privacy in particpatory sensing. *PerCom*, pages 328– 331, 2011.

[21] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical*, 27:379–423, 1948.

[22] H. Shin, V. Atluri, and J. Vaidya. A profile anonymization model for privacy in a personalized location based service environment. *International Conference on Mobile Data Management (MDM)*, pages 73–80, 2008.

[23] Sameer Tila. Real-world deployments of participatory sensing applications:current trends and future directions. *Sensor Networks*.

[24] Bo Wang and Jing Yang. Personalized algorithm based on entropy classification. *Computational Information Systems*, 8(1):259–266, 2012.

[25] X. O. Wang, W. Cheng, P. Mohapatra, and T. Abdelzaher. Enabling reputation and trust in privacy-preserving mobile sensing. *IEEE Transactions on Mobile Computing*, 99.

[26] X. Xiao and Y. Tao. Personalized privacy preservation. *ACM SIGMOD.*, pages 229–240, 2006.

[1] [2].

[1] k, the number of nearest neighbors(DC points) is different from K, the anonymity levels

[2] To retrieve the DC points for users in ASR, we use the approach as proposed in [12]