Preserving Anonymity of Recurrent Location-based Queries

Daniele Riboni, Linda Pareschi, Claudio Bettini
DICo - University of Milan
riboni,pareschi,bettini@dico.unimi.it

Sushil Jajodia
CSIS - George Mason University
jajodia@gmu.edu

***Abstract*—The anonymization of location based queries through the generalization of spatio-temporal information has been proposed as a privacy preserving technique. We show that the presence of multiple concurrent requests, the repetition of similar requests by the same issuers, and the distribution of different service parameters in the requests can significantly affect the level of privacy obtained by current anonymity-based techniques. We provide a formal model of the privacy threat, and we propose an incremental defense technique based on a combination of anonymity and obfuscation. We show the effectiveness of this technique by means of an extensive experimental evaluation.**

I. Introduction

Location based services (LBS) are Internet services that provide information or enable communication based on the location of users and/or resources at specific times. They are often designed to answer spatio-temporal nearest-neighbor or range queries issued from mobile devices, taking as one of the parameters the current location as identified through positioning technologies like GPS, cell tower triangulation, or WiFi positioning. Several commercial LBS like assisted car navigation, friend-finder, and proximity marketing are currently available. The success and popularity of these services will partly depend upon the privacy preserving technologies that will be designed and offered to final users. Indeed, compared with privacy issues in database publication, the spatio-temporal information contained in each user request, and the recurrence of requests in time, forces the consideration of new privacy threats and the design of specific defense techniques.

The general privacy threat consists in the acquisition by an adversary of the association between an individual's identity and her private information. In some cases, location at a specific time, as included in a request, is considered private; in other cases the service invoked or the specific parameters are considered private, and location and time may be used by the adversary to re-identify the issuer. The actual threats do not depend only on the nature of private information; a careful specification of the adversary model in terms of which requests he may acquire, and which external knowledge he may have access to, is a precondition to the identification of the privacy threats, and to the design of defense techniques. In this paper we illustrate a privacy threat in LBS due to the ability of the adversary to acquire requests issued by multiple users, in the same time granule as well as in different time granules. An example is illustrated in Section II along with the specification of the adversary model. In particular, we show that even if each request has been anonymized with state of the art techniques, the adversary can still associate private information with specific individuals with a high probability. The attack is based on the observation that users tend to issue LBS requests with parameters influenced by their personal profile, including personal data like nationality, age, gender, and more importantly their interests. While profile data can evolve in time, it is a rather slow process and this is reflected in the persistence of the same or similar service parameters in a subset of the requests issued at different times by each user. We illustrate a specific method an adversary can use to update, upon observing the requests issued at each time granule, his knowledge about the probability of each user to be associated to certain service parameters. This knowledge refinement, coupled with the ability of an adversary to restrict the set of potential issuers of each request based on location information as used in previous work [1], [2], [3], leads to a dangerous privacy threat not previously recognized in the literature.

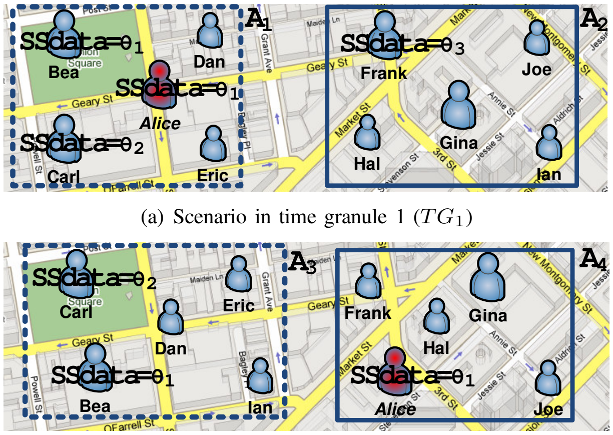
Related work can be divided in two main streams. Obfuscation-based defenses aim at obfuscating the private information in each request so that even if the issuer is identified, the adversary cannot recognize the specific private values associated with the original issuer's request. These techniques have been mostly applied in the case location and time are considered private, as in [4]. Anonymity-based defenses aim at preserving the anonymity of the issuers so that an adversary is not able to associate private information present in the requests with a specific individual. The defenses transform the so-called *quasi-identifier* information in requests so that the issuer becomes indistinguishable in a sufficiently large group of users (called *anonymity set*). Usually, service parameters are considered the data to be protected, and location information is considered a quasi-identifier, since the adversary may obtain information from external sources about the presence of a specific individual in the location from which the request was issued. A common technique is the generalization of the location to an area in order to include at least k potential issuers that become part of the anonymity set, enforcing k-anonymity. Most proposed techniques have considered anonymization of requests in isolation, i.e., ignoring the possibility of the adversary to correlate requests at different times [1], [2], [3], [5], as well as requests by different users. Only a few approaches consider the threats involved in dynamically acquiring requests (often called *historical attacks*), as we do in this paper; the threats involved in the recognition of traces of requests by the same (anonymous) issuer have been considered in [6], [7], [8], [9] and defenses have been proposed. Traces are supposed to be recognized by comparing pseudo-identifiers in requests or by spatio-temporal reasoning. Our work differs in two aspects: a) the threat we consider occurs even if no trace is recognized, b) we consider the effects on the composition of anonymity sets due to concurrent requests by multiple users with the same request parameters. To our knowledge this last aspect has been ignored in all previous work in LBS privacy except in a preliminary work of ours [10], and in a more recent paper [11], and has close relationship with the *diversity* problem identified in database publication [12]. Finally, we should mention that techniques based on private information retrieval have also been proposed for LBS [13] and they may be applied both for obfuscation and anonymity, since exchanged data is encrypted; however, their practical applicability seems limited both in terms of supported queries, and in terms of computational costs.

The contributions of this paper can be summarized as follows: (i) We formalize a previously unrecognized privacy threat in LBS due to correlation between concurrent request by multiple users, as well as to incremental refinement of adversarial knowledge along the service history; (ii) We propose a novel defense technique protecting from the identified threat; (iii) We present an experimental evaluation in a profile-based proximity marketing scenario.

In Section II we formalize the adversary model and illustrate the threat with an example. In Section III we formally define the adversarial inference method. In Section IV we propose a defense technique that is experimentally evaluated in Section V. Section VI concludes the paper.

II. Adversary's model and motivating example

As in several related works, our reference scenario includes a trusted server (*LTS*) which is aware of the actual location of users. This assumption is not far from reality, since most of us rely on a mobile operator for mobile communications, that is aware of our approximate position. The *LTS* acts as a proxy, by filtering and generalizing each user's request before it is forwarded to the service provider (SP) which is considered untrusted. Each service request r is logically divided into three parts: $IDdata$, $STdata$, and $SSdata$, containing user identification data, location and time of request, and service parameters, respectively. We refer to the set of possible values of *SSdata* as $\Theta = \{\vartheta_1, \ldots, \vartheta_n\}$, and we assume that Θ can be represented as a taxonomy. The *LTS* transforms each request r into a request r', by dropping $IDdata$ and generalizing the value of $STdata$, and possibly of $SSdata$ too. The adversary's model considered in this paper is based on the following *context assumptions*:



(a) Scenario in time granule 1 (TG_1)

(b) Scenario in time granule 2 (TG_2)

Figure 1. Motivating example

- ○ The generalization algorithm adopted by the *LTS* is publicly known;
- ○ We assume that the *LTS* works at a given time granularity, so that at each time-granule a group of generalized requests is forwarded to the SP. We assume that only one request per time granule can be issued by each user.
- ○ The adversary may obtain the generalized requests issued in one or more time granules. We refer to this context assumption as C_{MH} (Multiple-issuer Historical case).
- ○ The adversary may observe or obtain from external sources the position of specific individuals at given times. As in related work, we make a worst case assumption C_{ST} that considers complete location knowledge about potential issuers.
- ○ Correlation of requests at different time granules can only be done by analyzing *SSdata*. In principle, traces of requests made by the same individual can also be recognized on the basis of spatio-temporal reasoning or pseudo-identifiers included in requests. However, algorithms to deal with this case have been previously proposed [9], and can be seamlessly integrated with the one proposed in this paper.

Note that in this work we assume that the adversary has no specific prior knowledge about the association between individuals and sensitive service parameters (e.g., "Alice is interested in vegetarian restaurants"). Hence, his *prior knowledge* is modeled according to the following definition.

Definition 1 (Prior knowledge)**.** *The prior knowledge of the adversary is a function $K_{pri} : U \to \Upsilon$ in which U is the* *set of users,* $\Upsilon = \{(p_1, \ldots, p_n) | \sum_{1 \leq i \leq n} p_i = 1\}$ $(0 \leq p_i \leq 1)$ *is the set of possible probability distributions of values on the sensitive attribute* SSdata*, and for all users in* U*,* $\Upsilon = \{(\frac{1}{n}, \ldots, \frac{1}{n})\}$.

After observing generalized requests issued at time granule TG (and possibly also in time granules preceding TG) the adversary may compute his *posterior knowledge*, which is modeled according to the following definition.

Definition 2 (POSTERIOR KNOWLEDGE)**.** *The posterior knowledge of the adversary is a function* $K_{pos} : U \times \mathcal{TG} \rightarrow \Upsilon$ *in which* U *is the set of users,* $\mathcal{TG}$ *is a set of time granules, and* $\Upsilon = \{(p_1, \ldots, p_n) | \sum_{1 \leq i \leq n} p_i = 1\}$ $(0 \leq p_i \leq 1)$ *is the set of possible probability distributions of values on the sensitive attribute* SSdata *computed after observing the requests issued in* TG *and in previous time granules.*

Note that the above definition is very general. An inference method to actually compute the posterior knowledge K_{pos} is presented in Section III. On the basis of K_{pos}, the goal of the adversary is to reconstruct the association between a user u and the sensitive service parameter ϑ included in her request issued at TG. For instance, by observing that, according to $K_{pos}(u, TG)$, the probability of ϑ for u is considerably higher than the one for other users in U, the adversary may conclude that u issued a request having private value ϑ. Various profile-based proximity services are prone to this kind of privacy threats. The following example considers the case of a proximity marketing service.

Example 1. *Consider a proximity marketing service that proactively provides location-aware advertisements about sales on items belonging to a set of interest categories. Each registered user periodically communicates her current location to the service provider in order to receive advertisements. However, since the service provider is untrusted, users communicate to the service only part of their interest categories, while they do not report the ones involving sensitive information such as health status, religious beliefs, and political affiliations. However, advertisements regarding the latter categories can be obtained on-demand by issuing anonymous queries in which the user's location is generalized by the* LTS*, and containing the category of interest (a value in* $\{\vartheta_1, \vartheta_2, \ldots, \vartheta_{12}\}$*).*

Suppose that during TG_1 *a user Alice issues a request for sales regarding items of category* ϑ_1*. By joining location information in requests issued at* TG_1 *with the one communicated by its users, the adversary identifies two anonymity sets* A_1 *and* A_2 *(corresponding to users depicted in Figure 1(a)), both having cardinality* 5*. In our example, two of the three requests issued from users in* A_1 *(including Alice) ask for* ϑ_1 *and one for* ϑ_2*. Hence, the adversary can infer that the probability that Alice issued a request for* ϑ_1 *is* $\frac{2}{5}$*, while it is* $\frac{1}{5}$ *for* ϑ_2*. Next, suppose that the adversary can observe also requests issued at* TG_2*, including the one issued by Alice for* ϑ_1*. Once again, the adversary can recognize two anonymity sets* A_3 *and* A_4 *of cardinality* 5*, corresponding to the users depicted in Figure 1(b). During the lapse of time between* TG_1 *and* TG_2 *users have changed their positions. With regard to Alice's anonymity set* A_4*, the adversary can observe that the set of requests issued by users in* A_4 *is composed of a single request having private value* ϑ_1*. Consequently, the adversary can notice that the presence of Alice in a given anonymity set is correlated with a frequency of the private value* ϑ_1 *that is higher than the average frequency of the same value in the whole set of requests. Hence, he can conclude that probably Alice issued requests for* ϑ_1*.*

III. DERIVING POSTERIOR KNOWLEDGE

In this section we formally model the derivation of posterior knowledge in the historical multiple-issuers case. The following notation is necessary:

- $A_C(r')$ is the anonymity set of potential issuers of request r' identified on the basis of r' and of context C. For instance, if r' is the request issued by Alice during TG_1 (Example 1), $A_C(r') = \{$Alice, Bea, Carl, Dan, Eric$\}$.
- $R(A) = \{r'_1, \ldots, r'_n\}$ is the set of generalized requests issued by users in anonymity set A; in particular, $\forall r'_1, r'_2 \in R(A) \; : \; r'_1.STdata = r'_2.STdata$. For instance, if A is the anonymity set identified above (i.e., $A = A_C(r')$), $R(A)$ is the set composed of requests issued by Alice, Bea and Carl during TG_1.
- $\Theta(R) = \{\vartheta_1, \ldots, \vartheta_l\}$ is the set of values of *SSdata* included in the set R of generalized requests. For instance, if R is the set of requests identified above (i.e., $R = R(A)$), $\Theta(R) = \{\vartheta_1, \vartheta_2\}$.
- $m_{\vartheta, R}$ is the number of requests in R which include the *SSdata* ϑ; this value is called the *multiplicity* of ϑ in R. For instance, if $R = R(A)$ as above, the multiplicity of ϑ_1 in R is $m_{\vartheta_1, R} = 2$.
- Given posterior knowledge $K_{pos}(u, TG) = (p_1, \ldots, p_n)$, we denote by $K^{(i)}_{pos}(u, TG)$ the probability associated to the i-th sensitive value, i.e., $K^{(i)}_{pos}(u, TG) = p_i$. Similarly, given $K_{pri}(u) = (p_1, \ldots, p_n)$, $K^{(i)}_{pri}(u) = p_i$.

Intuitively, the probability that a user u issued one of the requests at time TG_n with parameter ϑ is influenced by the frequency of observation of the same parameter in the requests in $R(A)$ for each anonymity set A including u at $TG_1, \ldots, TG_n$. The higher is the frequency, the more it is probable that u issued a request with parameter ϑ. However, in most cases the cardinality of $R(A)$ is smaller than the cardinality of A, since service users do not continuously issue requests. Therefore, when the adversary computes his posterior knowledge based on requests issued in a given TG, he must consider the possibility that the user did not issue requests in TG. The following definition models the adversary's inference method under $\widetilde{C} = C_{MH+ST}$.

Definition 3 (INFERENCE METHOD)**.** *Given the context $\widetilde{C}$, an ordered set of time granules $\mathcal{TG} = \{TG_1, \ldots, TG_m\}$, a set of requests R issued at TG_m, a user $u \in U$, the set $\Theta = \{\vartheta_1, \ldots, \vartheta_n\}$ of SSdata, the inference method to derive the posterior knowledge at TG_n under $\widetilde{C}$ consists in the computation of: $K_{pos}(u, TG_m) = (p_1, \ldots, p_n)$, where for each $i \in \{1, \ldots, n\}$:*

$$p_i = \begin{cases} K_{pos}^{(i)}(u, TG_{m-1}) & \text{if } \nexists r \in R : u \in A_{\widetilde{C}}(r) \\ \beta_i + (1-\alpha) \cdot K_{pos}^{(i)}(u, TG_{m-1}) & \text{otherwise} \end{cases}$$

where $K_{pos}^{(i)}(u, TG_0) = K_{pri}^{(i)}(u)$, $\beta_i = \dfrac{m_{\vartheta_i, R(A)}}{|A|}$, $\alpha = \dfrac{|R(A)|}{|A|}$, and A is the anonymity set the user u belongs to (if such anonymity set exists).

Intuitively, if user u does not belong to any anonymity set at TG_m (first case in the formula of Definition 3), the adversary does not acquire any new information about u. Hence, his posterior knowledge regarding u at TG_m does not change with respect to the one at TG_{m-1}. In particular, if u never belonged to an anonymity set throughout $\mathcal{TG}$, the adversary's posterior knowledge corresponds to his prior knowledge $K_{pri}^{(i)}(u)$. On the contrary (second case), if u belongs to an anonymity set A she is the potential issuer of a request $r \in R(A)$. The actual probability that u issued one request in $R(A)$ is $\alpha \in [0, 1]$; hence, we call this parameter the *learning rate* of the adversary. Given a sensitive value ϑ_i, the parameter β_i accounts for the probability that u issued a request at TG_m having that sensitive value (first addend in the formula). The second addend $(1-\alpha)$ accounts for the probability that u did not issue a request at TG_m; under this hypothesis, the posterior knowledge $K_{pos}^{(i)}(u, TG_{m-1})$ at TG_{m-1} is taken into account.

Proposition 1. *$K_{pos}(u, TG_m)$ computed by the inference method illustrated in Definition 3 is a probability distribution. It follows that the inference method illustrated in Definition 3 computes the adversary's posterior knowledge.*

Example 2. *Continuing Example 1, we show how the adversary computes his posterior knowledge about the association of user Alice and sensitive value ϑ_1 after observing requests issued at TG_1 and TG_2. Recall that the cardinality of the set Θ of* SSdata *is 12. At the first time granule TG_1, for each user the adversary's prior knowledge K_{pri} is modeled by the uniform distribution $(\frac{1}{12}, \ldots, \frac{1}{12})$. Hence, according to Definition 3, $K_{pos}^{1}(Alice, TG_1) \simeq 0.43$. After observing requests issued at time granule TG_2, the adversary's posterior knowledge is $K_{pos}^{1}(Alice, TG_2) \simeq 0.54$. Hence, after TG_2 the value that associates Alice to ϑ_1 is considerably higher than the value for the other users belonging to the same anonymity set as Alice* $(0.54$ *vs* $0.27)$.

Algorithm 1: HMID algorithm

```
Input: $k$ - minimum $k$-anonymity level; $\widetilde{C}$ - attack context; $P_i$ - list
       of potential issuers at $TG_i$; $\mathcal{R}_i$ - requests issued at $TG_i$;
       $tc_1, \ldots, tc_L$ - $t$-closeness levels for each level of
       generalization of SSdata; MaxST - max level of
       generalization admitted for STdata.
Output: $R'_i$ - set of anonymized requests.

HMID( $\widetilde{C}, P_i, \mathcal{R}_i, k, tc_1, \ldots, tc_L$, MaxST )
begin
    $R'_i$ := $\emptyset$
    $P_i$ := HilbertOrdering($P_i$, location)
    repeat
        forall level $j = 1, \ldots, L$ of generalization of SSdata do
            int $n := k$
            $A_j$ = first $n$ users in $P_i$
            while $MBR(A_j) \leq MaxST$ and
            $t\text{-}cl(R(A_j), j, \mathcal{R}_i) \geq tc_j$ and $P_i \neq \emptyset$ do
                $n := n + k$
                $A_j$ = first $n$ users in $P_i$
            $QoS_j$ := $QoS(A_j, R(A_j), j)$
        if no $A_j$ exists that satisfies $tc_j$ then
            $A$ := group users until: MBR($A$) $> MaxST$ or
            $A = P_i$
            $\mathcal{R}_i := \mathcal{R}_i \setminus R(A)$ ; $P_i := P_i \setminus A$
        else
            $\overline{j}$ := level of generalization s.t. $QoS_j$ is maximum
            $P_i := P_i \setminus A_{\overline{j}}$
            $R(A_{\overline{j}})$ := Anonymize($A_{\overline{j}}, R(A_{\overline{j}})$)
            $R(A_{\overline{j}})$ := Obfuscate($R(A_{\overline{j}}), \overline{j}$)
            $R'_i := R'_i \cup R(A_{\overline{j}})$
    until $\mathcal{R}_i = \emptyset$ or $P_i = \emptyset$
    return $R'_i$
end

$t$-$cl$ ($R, j, \mathcal{R}_i$ )
begin
    $D$ := PDF($R$, SSdata)
    $D'$ := PDF($\mathcal{R}_i$, SSdata)
    return $KL(D, D')$
end
```

IV. Defense technique

In order to measure the success of privacy attacks, as well as of defenses against them, it is necessary to define the criteria by which the adversary can choose the *SSdata* ϑ to be associated with a user u. If the adversary chooses the correct value, the attack is successful. For the sake of this paper we adopt a criterion γ, which consists in comparing $\omega_n(\vartheta_i, u) = K_{pos}^{(i)}(u, TG_n)$ at time granule TG_n with the average value $\overline{\omega}_n(\vartheta_i, U) = \frac{\sum_{u \in U} K_{pos}^{(i)}(u, TG_n)}{|U|}$ computed at time granule TG_n in the considered population of service users U. Experimental evidence (reported in Section V) shows that this attack criterion is very effective. However, our defense technique can be also applied to different criteria. We call *confidence* Ω_n the function:

$$\Omega_n(\vartheta_i, u) = \begin{cases} 0 & \mathit{if}\, \overline{\omega}_n(\vartheta_i, U) = 0 \\ \frac{\omega_n(\vartheta_i, u)}{\overline{\omega}_n(\vartheta_i, U)} & \mathit{otherwise} \end{cases}$$

According to criterion γ, the value ϑ chosen by the adversary is the one having maximum confidence:

$$\Omega_n(\vartheta, u) = \max_{\vartheta_i \in \Theta} \{\Omega_n(\vartheta_i, u)\}.$$

HMID: defending with anonymity and obfuscation: As for any other defense technique, the objective of our technique, called *historical multiple-issuers defense (HMID)*, is to guarantee the necessary level of privacy while maximizing the usefulness of the data. To this aim, HMID adopts both anonymity (obtained by generalizing *STdata*) and obfuscation (obtained by generalizing *SSdata*). Its specific goal is to find the combination of the generalization levels for *STdata* and *SSdata* that maximizes the data quality while enforcing the required privacy level.

For the sake of LBS requests, data quality can be naturally measured as a function of the generalization level of user's location and of request parameters in anonymized requests. However, different applications may have different requirements that determine their actual quality of service (QoS). For instance, some services need very precise location information, while being quite tolerant with respect to the generalization of service parameters. On the other hand, for other services accurate users' location is not strictly required, while service parameters are the most prominent data. HMID copes with this aspect by supporting the definition of any kind of function $\mathcal{L}_{QoS}$ to determine the QoS resulting from requests generalization.

The *privacy leak (pl)* determined by an attack at a given time granule can be measured as the percentage of users that are correctly associated with their *SSdata* by an adversary based on context $\widetilde{C}$ and criterion γ. Hence, we define the level of privacy $\mathcal{L}_p$ as: $(1 - pl)$. The desired level of privacy is guaranteed by enforcing k-anonymity coupled with a variant of the t-closeness technique originally proposed by Li *et al.* [14] for privacy protection of microdata released from databases. K-anonymity ensures that, based on $\widetilde{C}$, the issuer of each generalized request r is indistinguishable in an anonymity set A of at least k potential issuers. However, as shown in Example 1, k-anonymity is insufficient when the adversary may observe multiple requests issued in the same time granule. Indeed, in that case he may derive the association between a user and a request based on the *SSdata* in that request, and on the distribution of *SSdata* in the history of requests originated from the anonymity sets including that user. Hence, considering the whole set of requests issued in a time granule, our t-closeness variant aims at counteracting this kind of adversarial inference by smoothing the differences among the distribution of *SSdata* in requests originated from the different anonymity sets. In particular, for each anonymity set A we ensure that the distance between the distribution of *SSdata* in requests originating from A and the distribution of *SSdata* in the whole set of requests issued during the same time granule is below a threshold t. Given a privacy threshold h ($0 < h < 1$), the value of t sufficient to guarantee $\mathcal{L}_p \geq h$ is experimentally estimated; in general, a different value of t must be used for each *SSdata* generalization level. We measure the difference between the two distributions using the well known Kullback-Leibler (KL) divergence. If an anonymity set satisfies k-anonymity but does not fulfill our t-closeness variant, HMID adds more potential issuers to it (by further generalization of request location), until the required level of t-closeness is reached; if that level cannot be enforced, requests originating from that anonymity set are discarded, and their issuers are informed. In most cases the number L of levels in the hierarchy of *SSdata* is quite limited. Hence, HMID tries all the possible levels of *SSdata* generalization, coupled with the finest-grained generalization of *STdata* that satisfy both k-anonymity and our t-closeness variant, in order to find the combination of *SSdata* and *STdata* generalization levels that maximizes $\mathcal{L}_{QoS}$. As in most related works, for efficiency reasons we adopt a heuristic algorithm in order to group users in anonymity sets. In particular, as proposed in [15] we adopt a strategy based on the Hilbert [16] space-filling curve. The Hilbert space-filling curve is a function that maps a point in a multi-dimensional space into an integer; with this technique, two points that are close in the multi-dimensional space are also close, with high probability, in the one-dimensional space obtained by the Hilbert transformation. As it can be evinced from its pseudo-code (reported in Algorithm 1), the complexity of HMID is $O(L \cdot \frac{|U|^2}{k})$. Since the dominant factor is U, an optimization consists in partitioning – based on location – the whole set U of users into a number of smaller subsets, and in applying HMID independently to every such set considering the set of requests originating from it.

Algorithm: For each time granule TG_i, based on the sets $\mathcal{R}_i$ of requests and P_i of potential issuers, the algorithm returns a set of anonymized requests R'_i.

At first (line 4), the algorithm orders users in P_i according to their index obtained from the application of the Hilbert space filling curve on their current location. Then (lines 6 to 12), for each level j of possible *SSdata* generalization, a growing set A_j of users is grouped according to the Hilbert ordering until the minimum generalization level of *STdata* (computed as the minimum bounding rectangle including every user in A_j) satisfying both k-anonymity and t-closeness is reached. The corresponding level QoS_j of QoS is then computed.

If it does not exist an *SSdata* generalization level satisfying both k-anonymity and t-closeness (lines 13 to 15), requests are discarded and their potential issuers are removed

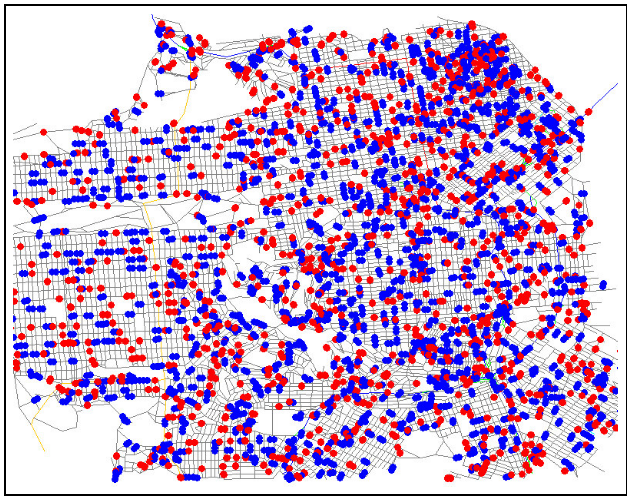


Figure 2. A snapshot of pedestrians' and drivers' positions

from P_i. Otherwise (lines 17 to 21), the generalization level $\overline{j}$ of *SSdata* maximizing the QoS is chosen. The *SSdata* in requests originating from anonymity set $A_{\overline{j}}$ are generalized at level $\overline{j}$, while *STdata* in the same requests are generalized by the minimum bounding rectangle including the location of every user in $A_{\overline{j}}$. Original requests originating from $A_{\overline{j}}$ are removed from $\mathcal{R}_i$, and the corresponding generalized requests are included in R'_i. The algorithm continues until no other request remains in $\mathcal{R}_i$.

V. Experimental evaluation

In this section we experimentally evaluate our defense technique in terms of enforced level of privacy and achieved data quality.

Experimental setup: Experiments were performed on synthetic data obtained using the moving object generator described in [17]. The simulation models a population of 50,000 persons moving in the San Francisco area, from a random starting point to a random destination, during a time period of 200 minutes (each one corresponding to a single time granule TG_m). A snapshot showing the position of part of the users in a time granule is shown in Figure 2. The dimension of the considered area is about 100km^2, with an average density of 500 persons per km^2. This density was the highest we could obtain with the used generator to model 200 time granules. Note that this density is lower than the real one in a urban area; when considering a higher density, we expect the resulting generalized areas to be proportionally smaller than the ones obtained in our experiments. Persons are equally divided into pedestrians (that move at an average speed of 4 km/h) and people using public transportation (average speed of 20 km/h), and update their location at the *LTS* every one minute.

The population is further divided into a group of *active users* of the proximity marketing service (i.e., users issuing at least one anonymous query during the length of our simulation; 20% of the whole population), and a group of *idle users*. Each active user is randomly associated with one of the 12 possible *SSdata* contemplated in our motivating example; each request contains the *SSdata* of its issuer. We have performed the experiments under 3 different conditions: *i)* low frequency of requests (Freq.1: each active user has a probability ranging from 25% to 0.016% of issuing a request at a given time granule), *ii)* medium frequency of requests (Freq.2: from 75% to 6%), and *iii)* high frequency of requests (Freq.3: from 100% to 12.5%). In the following we compare HMID with different defense techniques from adversary's posterior knowledge acquired under context $\widetilde{C}$ based on requests issued at time granules $\mathcal{TG} = \{TG_1, \ldots, TG_{200}\}$. The goal of defense techniques is to keep the $\mathcal{L}_p$ higher than 0.8 (i.e., at each time granule the adversary has less that 20% probability of correctly identifying the *SSdata* of a user).

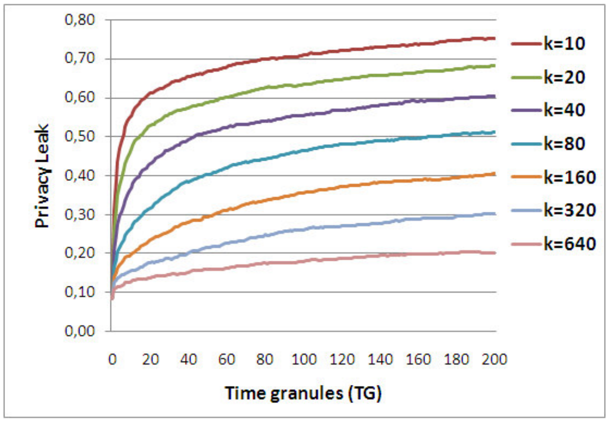


Figure 3. k-anonymity: privacy leak

We measure by means of the parameter $\mathcal{L}_{QoS}$ the level of QoS deriving from the transformations of service requests introduced by the defense techniques. To estimate the QoS we consider the information loss $\mathcal{IL}_{SS}$ and $\mathcal{IL}_{ST}$ (having values from 0 to 1) deriving from *SSdata* and *STdata* generalization, respectively. Formally, $\mathcal{L}_{QoS} = (1 - \mathcal{IL}_{SS}) \cdot (1 - \mathcal{IL}_{ST})$. In particular, in a first set of experiments we measure $\mathcal{IL}_{SS}$ adopting the information loss metrics introduced in [18]; we measure $\mathcal{IL}_{ST}$ by a function linearly growing from 0 (perimeter of the generalized location is 0) to 1 (perimeter greater or equal to 6Km). We call this metric $\mathcal{L}_{QoS_1}$.

Defense based on k-anonymity: In the first set of experiments we evaluated the application of a standard k-anonymity technique to protect against attacks under $\widetilde{C}$. In this experiment, we adopt the Hilbert ordering to arrange users in anonymity sets. We have performed the experiments with different values of k. Results are shown in Figure 3 and Table I, and show that this technique is not well-suited to

k	**20**	**40**	**80**	**160**	**320**	**640**
Area (Km^2)	0.03	0.08	0.19	0.44	0.97	2.05
Perimeter (m)	620	1001	1579	2439	3694	5456

Table I
k-ANONYMITY: LOCATION GENERALIZATION

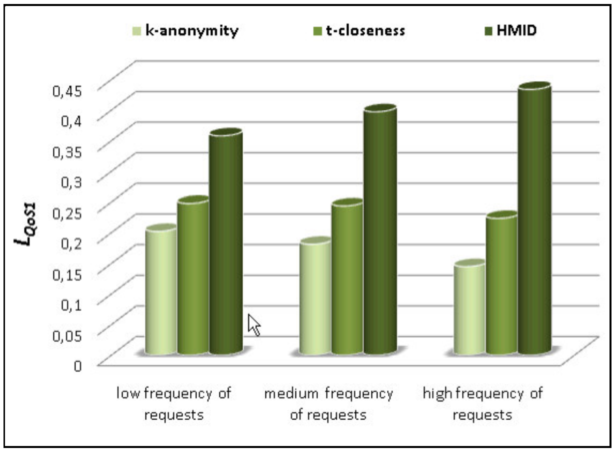


Figure 4. Comparison based on QoS ($\mathcal{L}_{QoS_1}$)

Freq.1	**Perimeter (Km)**	**Area (Km^2)**	**% non-gen.**	**% gen.1-lev.**	**% gen.2-lev.**
k-an.	5,48	2,06	100%	0%	0%
t-cl.	5,26	2,00	100%	0%	0%
HMID	3,57	1,09	39%	38%	23%

Freq.2	**Perimeter (Km)**	**Area (Km^2)**	**% non-gen.**	**% gen.1-lev.**	**% gen.2-lev.**
k-an.	5,72	2,23	100%	0%	0%
t-cl.	5,35	2,10	100%	0%	0%
HMID	2,96	0,86	32%	26%	42%

Freq.3	**Perimeter (Km)**	**Area (Km^2)**	**% non-gen.**	**% gen.1-lev.**	**% gen.2-lev.**
k-an.	6,16	2,57	100%	0%	0%
t-cl.	5,55	2,24	100%	0%	0%
HMID	2,30	0,58	18%	24%	58%

Comparison in terms of: request frequency; perimeter and area of generalized location; % of requests with generalized *SSdata*.

Table II
GENERALIZATION (HMID WITH $\mathcal{L}_{QoS_1}$).

the considered attack (Definition 3). Indeed, the minimum k required to keep the *privacy leak* below 0.2 (k=640) leads to generalized areas too wide to guarantee a satisfactory quality of service ($2.2\,\mathrm{km}^2$, with an average perimeter of $5.7\,\mathrm{km}$; see also Figure 4). The *privacy leak* grows considerably when using smaller levels of k. For instance, in order to keep the average generalized location area below $1\,\mathrm{km}^2$ a value of $k \leq 320$ must be chosen; this value corresponds to a *privacy leak* greater than 0.3.

Defense based on k-anonymity and t-closeness: This technique is similar to HMID, with the only difference that obfuscation of *SSdata* is not allowed. In these experiments the level of t sufficient to guarantee the required privacy level ($\mathcal{L}_p \geq 0.8$) is empirically estimated, and a minimum level of anonymity $k = 20$ is chosen. Experimental results (Figure 4, label *t-closeness*) show that, given the same privacy level, this technique slightly outperforms the baseline k-anonymity technique in terms of $\mathcal{L}_{QoS_1}$.

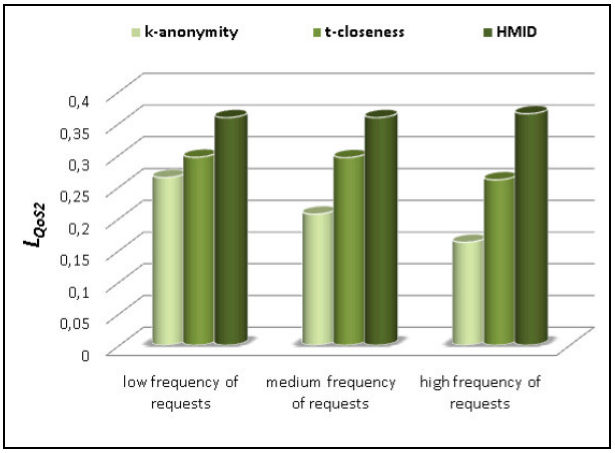


Figure 5. Comparison based on QoS ($\mathcal{L}_{QoS_2}$)

HMID technique: In the last set of experiments we evaluated the HMID technique. We empirically chose the levels of t-closeness for three levels of *SSdata* obfuscation: non-generalized *SSdata*, generalized one level (from 12 to 6 *SSdata*), and generalized two levels (from 12 to 3 *SSdata*). The chosen t-closeness levels were sufficient to guarantee $\mathcal{L}_p > 0.8$. Experimental results (Figure 4) show that HMID outperforms the other ones in terms of QoS while enforcing the same level of privacy $\mathcal{L}_p$. A deeper analysis of the results is shown in Table II. In particular, HMID leads to smaller average perimeters and areas with respect to the other techniques. The percentage of requests with generalized *SSdata* depends on the frequency of requests.

In order to evaluate the robustness of HMID with respect to different QoS metrics we performed a further set of experiments using different functions for $\mathcal{IL}_{SS}$ and $\mathcal{IL}_{ST}$. In particular, in this set of experiments we assigned a proportionally growing information loss to growing levels of *SSdata* generalization. Hence, $\mathcal{IL}_{SS}$ is 0 if the service parameter is not generalized; $\mathcal{IL}_{SS}$ is $\frac{1}{3}$ if it is generalized one level; it is $\frac{2}{3}$ if it is generalized two levels. With regard to $\mathcal{IL}_{ST}$, we set no information loss if the perimeter of the generalized location is less than 2Km; information loss grows logarithmically from 0 to 1 until the perimeter is up to 6Km; it is 1 for perimeters larger than 6Km. We call the combination of these metrics $\mathcal{L}_{QoS_2}$. Experimental results are reported in Figure 5 and Table III, and show that HMID is robust with respect to different QoS metrics (possibly determined by the specific requirements of different services).

VI. CONCLUSION AND FUTURE WORK

In this paper we addressed privacy issues for recurrent location-based queries. We showed that if an adversary may observe multiple concurrent requests, and similar requests are issued several times by the same issuers, the distribution of different service parameters in the requests can

Freq.1	Perimeter (Km)	Area (Km^2)	% non-gen.	% gen.1-lev.	% gen.2-lev.
k-an.	5,25	1,90	100%	0%	0%
t-cl.	5,28	2,02	100%	0%	0%
HMID	3,88	1,18	48%	36%	16%

Freq.2	Perimeter (Km)	Area (Km^2)	% non-gen.	% gen.1-lev.	% gen.2-lev.
k-an.	5,72	2,23	100%	0%	0%
t-cl.	5,33	2,07	100%	0%	0%
HMID	3,03	0,86	34%	24%	42%

Freq.3	Perimeter (Km)	Area (Km^2)	% non-gen.	% gen.1-lev.	% gen.2-lev.
k-an.	6,16	2,57	100%	0%	0%
t-cl.	5,63	2,30	100%	0%	0%
HMID	2,71	0,74	25%	27%	47%

Table III
GENERALIZATION (HMID WITH $\mathcal{L}_{QoS_2}$).

significantly affect the level of privacy obtained by current anonymity-based techniques. We formalized this kind of privacy threats, we proposed a defense technique based on a combination of anonymity and obfuscation, and we showed that this technique outperforms ones based on k-anonymity and on a variant of t-closeness in terms of quality of service while enforcing the required privacy level.

Future research directions include the extension of our formal model and defense techniques to other possible context assumptions; in particular, the ability of an adversary to have specific prior knowledge about the association among classes of users and sensitive request parameters. On the other side, the worst case assumption of the adversary having access to complete location information may be relaxed to more realistic cases.

ACKNOWLEDGMENTS

This work has been partially supported by the Italian Ministry of University and Research under grant PRIN-2007F9437X (project *ANONIMO*), and by the National Science Foundation (NSF) under grant N. CNS-0716567.

REFERENCES

[1] M. Gruteser and D. Grunwald, “Anonymous usage of location-based services through spatial and temporal cloaking.” in *Proc. of the 1st International Conference on Mobile Systems, Applications and Services*. The USENIX Association, 2003.

[2] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, “Preventing location-based identity inference in anonymous spatial queries,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 12, pp. 1719–1733, 2007.

[3] B. Gedik and L. Liu, “Protecting location privacy with personalized k-anonymity: Architecture and algorithms,” *IEEE Transactions on Mobile Computing*, vol. 7, no. 1, pp. 1–18, 2008.

[4] M. L. Yiu, C. S. Jensen, X. Huang, and H. Lu, “SpaceTwist: Managing the Trade-Offs Among Location Privacy, Query Performance, and Query Accuracy in Mobile Services,” in *Proceedings of the 24th International Conference on Data Engineering (ICDE ’08)*, 2008, pp. 366–375.

[5] D. Riboni, L. Pareschi, and C. Bettini, “Shadow attacks on users’ anonymity in pervasive computing environments,” *Pervasive and Mobile Computing*, vol. 4, no. 6, pp. 819–835, 2008.

[6] A. R. Beresford and F. Stajano, “Location privacy in pervasive computing,” *IEEE Pervasive Computing*, vol. 2, no. 1, pp. 46–55, 2003.

[7] C. Bettini, X. S. Wang, and S. Jajodia, “Protecting privacy against location-based personal identification.” in *Proc. of the 2nd workshop on Secure Data Management (SDM)*, ser. LNCS, vol. 3674. Springer, 2005, pp. 185–199.

[8] T. Xu and Y. Cai, “Location anonymity in continuous location-based services,” in *Proc. of ACM International Symposium on Advances in Geographic Information Systems*. ACM Press, 2007.

[9] S. Mascetti, C. Bettini, X. S. Wang, D. Freni, and S. Jajodia, “*ProvidentHider*: an Algorithm to Preserve Historical k-Anonymity in LBS,” in *Proceedings of the 10th International Conference on Mobile Data Management (MDM ’09)*. IEEE Computer Society, 2009.

[10] C. Bettini, S. Jajodia, and L. Pareschi, “ Anonymity and Diversity in LBS: a Preliminary Investigation,” in *Proc. of the 5th Int. Conf. on Pervasive Computing and Communication (PerCom)*. IEEE Computer Society, 2007.

[11] F. Liu and K. Hua, “Query l-diversity in location-based services,” in *To appear in the Proc. of the First International Workshop on Mobile Urban Sensing (MobiUS)*. IEEE Computer Society, 2009.

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-Diversity: Privacy Beyond k-Anonymity,” in *Proceedings of ICDE 2006*. IEEE Computer Society, 2006.

[13] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, “Private queries in location based services: Anonymizers are not necessary,” in *Proc. of SIGMOD*. ACM Press, 2008.

[14] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity.” in *ICDE*. IEEE, 2007, pp. 106–115.

[15] G. Ghinita, P. Kalnis, and S. Skiadopoulos, “PRIVE: anonymous location-based queries in distributed mobile systems,” in *Proceedings of the 16th International Conference on World Wide Web*. ACM, 2007, pp. 371–380.

[16] A. R. Butz, “Alternative Algorithm for Hilbert’s Space-Filling Curve,” *IEEE Trans. Comput.*, vol. 20, no. 4, pp. 424–426, 1971.

[17] T. Brinkhoff, “A Framework for Generating Network-Based Moving Objects,” *GeoInformatica*, vol. 6, no. 2, pp. 153–180, 2002.

[18] X. Xiao and Y. Tao, “Personalized privacy preservation,” in *SIGMOD ’06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM Press, 2006, pp. 229–240.