Protecting Location Privacy with K-Confusing Paths Based on Dynamic Pseudonyms

Ken Mano
NTT Corporation
Kanagawa 243–0198, Japan
Email: mano.ken@lab.ntt.co.jp

Kazuhiro Minami
Institute of Statistical Mathematics
Tokyo 190–8562, Japan
Email: kminami@ism.ac.jp

Hiroshi Maruyama
Institute of Statistical Mathematics
Tokyo 190–8562, Japan
Email: hm2@ism.ac.jp

***Abstract*—As smart phones with a GPS receiver have been becoming popular recently, many people have realized the issue of protecting their location privacy. Previous research on location privacy mainly focuses on anonymization techniques for removing identifiable information from users' location traces. Although anonymized location data is useful to many applications, such as traffic monitoring, we can provide a new class of location-based services by utilizing path information of mobile users. In this paper, we present a dynamic pseudonym scheme for constructing alternate possible paths of mobile users to protect their location privacy. We introduce a formal definition of location privacy for pseudonym-based location data sets and show an efficient verification algorithm for determining whether each user in a given location data set has sufficient number of possible paths to disguise the user's true movements.**

I. Introduction

Nowadays a vast majority of people are using mobile devices equipped with a GPS receiver, and it thus becomes feasible to keep track of people's movements in a wide area by collecting GPS data from those mobile devices. Such a large volume of location data gives us a precise global view on people's mobility patterns, and we can thus support analytic location-based services, such as real-time traffic monitoring [5] and urban planning for future sustainable cities [15].

However, due to the significant concern for location privacy [1], the sharing of mobile users' location traces has been mostly restricted to anonymized data sets where users' identities are removed. We usually need to follow the practice of ensuring the k-anonymity [6], which degrades the granularity of location data to make sure that every location contains more than k people. Consequently, k anonymized data sets provide little information on users' mobility patters, which makes it difficult to link multiple data points produced by the same user.

There are, however, many situations where we can improve our analytic methods by considering users' mobility patterns. For example, Draffic [4] provides a statistical analysis of people's moving paths in sightseeing areas so that hotels and souvenir shops can take effective measures to draw more visitors and provide them with better services. Similarly, a shopping mall manager could allocate various stores in the mall such that customers' shopping experiences match their moving behaviors conveniently.

We, therefore, propose a new dynamic pseudonym scheme for constructing a location data set that retains users' path information while preserving their location privacy. Our basic approach is to exchange multiple users' pseudonyms only when they meet at the same location to eliminate the linkability of their pseudonyms before and after that exchange. Our privacy metrics requires that, at a given time t, every user has enough number of plausible paths heading towards K different locations.

To make this dynamic pseudonym-based scheme practical, we address an issue of multi-path inconsistencies among multiple users. Assuming that user's home locations are public knowledge available to an adversary, we find that all pseudonym exchanges cannot be effective; the adversary can detect global inconsistencies among multiple plausible paths taken by different users. It is thus not trivial to decide whether a given data set is safely publishable. We, therefore develop an efficient algorithm for determining whether it is possible to convert a given location data set into pseudonym-based data satisfying the (K, t)-privacy metrics.

The rest of the paper is organized as follows. We first introduce our system model for pseudonym-based location services in Section II and then define our privacy metrics in Section III. Next, we present a verification algorithm for a pseudonym-based location data set and discuss possible future work concerning the algorithm in Section IV and Section V, respectively. We cover related work in Section VI and finally conclude in Section VII.

II. System model

Figure 1 shows our system model for pseudonym-based location systems. We assume that a mobile user u_i carrying a GPS-enabled mobile devices periodically reports a triplet (u_i, l_k, t_k), which indicates that user u_i is in location l_k at time t_k. The pseudonym-based location server receives from multiple users their identifiable location data, replaces their identities with pseudonyms, and provides location-based content providers, such as traffic monitoring applications, with location data with pseudonyms.

We first introduce the following four sets U, P, L, and T to define our system model.

U: a set of m mobile users such that $|U| = m$.
P: a set of m pseudonyms such that $|P| = m$.
L: a set of symbolic locations.

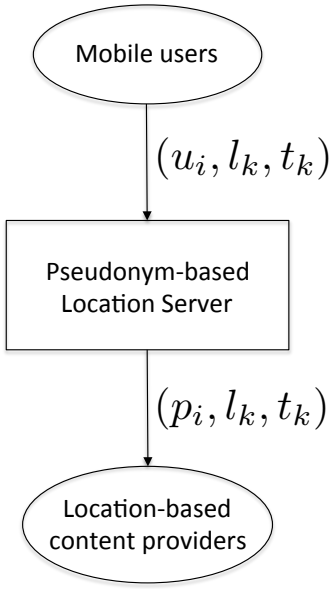

Fig. 1. System model. The pseudonym-based location server replaces a user's identity u_i with a pseudonym p_i before releasing location data to content providers.

T: a set of timestamps $\{0, 1, \ldots, t^*\}$ where t^* is the last timestamp.

We next define the following two functions.

Definition 1 (Location function W): The location function $W : U \times T \rightarrow L$ returns a location l of user u at time t.

Definition 2 (Pseudonym assignment function N): The pseudonym assignment function $N : U \times T \rightarrow P$ maps a user u at time t to a pseudonym p. We say that a user u owns a pseudonym p at time t if $N(u, t) = p$. For every time $t \in T$, the function $N_t(u) \equiv N(u, t)$ is a one-to-one function from U to P.

We now define a pseudonym-based data set parameterized by the functions W and N as the following set of triplets:

$$\{(p, l, t) \mid t \in T, u \in U, p = N(u, t), l = W(u, t)\}.$$

This data set represents the output from the pseudonym-based location server in Figure 1. In this paper, we consider a malicious content provider who legitimately obtains a data set from the pseudonym-based location server and tries to reveal the identity of a user corresponding to a certain pseudonym in the data set.

III. Pseudonym-based location privacy

To replace the user identity on a given moving path with the static pseudonym does not necessarily protect the user's location privacy. We take an approach of changing each user's pseudonym dynamically to prevent inference attacks using external knowledge about her home location.

A. Pseudonym exchanges

A user typically starts his moving path from her home and returns there again at the end. Therefore, if a user's home address is known to a malicious content provider, which is a well-agreed assumption in location privacy research [6], his moving path with the same pseudonym does not protect his location privacy; it is trivial to infer that the whole path belongs to the same user whose home address appears at both ends.

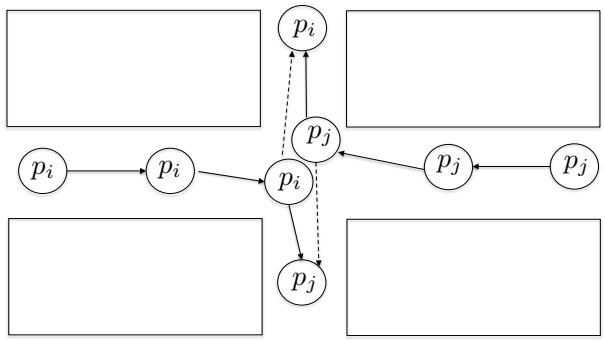

Fig. 2. Example pseudonym exchange. Two users exchange their pseudonyms p_i and p_j at the intersection. The solid lines denote each user's actual path while the dotted lines denote an alternate possible path.

Therefore, it is necessary to change pseudonyms dynamically to prevent the above attack. The basic idea is to divide a whole path of the same user into multiple segments with different pseudonyms such that it makes it infeasible to link any neighboring segments. However, when a user moves in an area where there is no other nearby users, it is straightforward to link two pseudonyms of the same user since we know that that user who is subjective to laws of physics cannot jump to a remote distant place in a short period.

To address this issue, we take an approach of exchanging multiple users' pseudonyms only when they meet at the same location (i.e., a mix zone [2]). Figure 2 shows an example of two users' exchanging their pseudonyms. Two users who own pseudonyms p_i and p_j respectively randomly exchange their pseudonyms when meeting at the intersection. Although the user who previously owns pseudonym p_i actually turns left at the corner, we consider the alternate path of turning right also possible. The other user similarly has the two possible paths after passing the intersection.

To consider only such valid pseudonym exchanges, we put the following constraint on the pseudonym assignment function N. For every pair of two different users $u, u' \in U$ and time $t > 0$, if $N(u, t-1) = p$ and $N(u', t) = p$, then $W(u, t-1) = W(u', t-1)$ holds. Intuitively, this constraint implies that if a user u' receives another user u's pseudonym p at time t, users u and u' must meet at the same location at the previous time $t-1$.

B. Multi-path consistency

If we consider the possible paths of a single user, whenever the user meets another user, we can add a new branch as a possible segment of the path. However, we assume in this paper that every user starts his path from his home location and then eventually returns there at the end of the path. Thus,

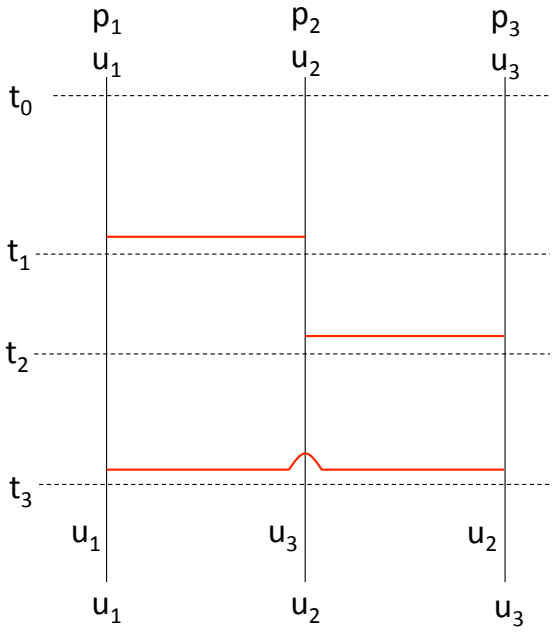

Fig. 3. An example of time-changing pseudonym assignments based on the ladder model.

we need to eliminate some possible branches if taking that direction makes it impossible for the user to return to his home location. Furthermore, even if one user u_i is able to return home with some possible path, another user u_j who exchange her pseudonym with u_i might lose her possible route to return home.

We elaborate this multi-path consistency issue with the ladder model in Figure 3. The ladder model represents a pseudonym assignment function N in a graphical way abstracting away each user's physical movements. Figure 3 shows an example ladder model for three users u_1, u_2 and u_3. The model denotes each pseudonym by a vertical line, and represents an encounter of multiple users by connecting those pseudonyms with a horizontal line. Assuming that time passes downward vertically, we specify the sequential order of users' meetings by the positions of horizontal lines.

Each pseudonym p_i is associated with a particular user at any give time t. Pseudonym p_1, p_2, and p_3 in Figure 3 are associated with users u_1, u_2, and u_3 respectively, both at the start and end times t_0 and t_3. If we construct user u_1's possible time-changing pseudonym assignments by exchanging pseudonyms, we obtain the following sequences:

1) $p_1 \to p_1 \to p_1 \to p_1$
2) $p_1 \to p_2 \to p_2 \to p_2$
3) $p_1 \to p_2 \to p_3 \to p_3$
4) $p_1 \to p_2 \to p_3 \to p_1$

If we consider the requirements that user u_1 owns pseudonym p_1 at times t_0 and t_3, we must eliminate sequences (2) and (3) leaving (1) and (4) as possible sequences of pseudonym assignments. However, if we take the pseudonym sequence (4), users u_2 and u_3 are forced to take the pseudonym sequences $p_2 \to p_1 \to p_1 \to p_3$ and $p_3 \to p_3 \to p_2 \to p_2$ respectively,

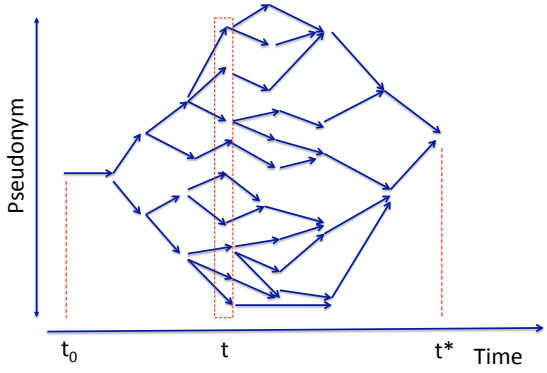

Fig. 4. Concept of (K,t)-pseudonym location privacy. We assume that any pseudonym sequence following arrows from left to right can be produced by a certain multi-path consistent pseudonym assignment function N.

violating their endpoint requirements. Thus, it turns out to be impossible for user u_1 to take the pseudonym sequence (4) above.

We should therefore consider possible pseudonym sequences of multiple users simultaneously to ensure that the resulting pseudonym assignment function N satisfy the following multi-path consistency requirement.

Definition 3 (Multi-path consistent function N): We say that, a given location function W, a pseudonym assignment function N is multi-path consistent if

1) $\forall u, u' \in U, \forall t \in T > 0 : N(u, t-1) = p \wedge N(u', t) = p \Rightarrow W(u, t-1) = W(u', t-1)$ and
2) $\forall u_i \in U : N(u_i, 0) = N(u_i, t^*) = p_i$.

C. (K,t)-pseudonym location privacy

We argue that the number of possible pseudonym sequences is not appropriate privacy metrics for pseudonym-based location services. Consider the situation where two users move together taking the same moving path. If those two users possibly exchange their pseudonyms at each time, we end up having exponential number of possible pseudonym sequences with respect to the length of time. Therefore, we rather use the number of pseudonyms at a given time t on possible pseudonym sequences satisfying the multi-path consistency requirement as our location privacy metrics. Figure 4 shows such multiple pseudonym sequences of user u_i. There is only a single possible pseudonym at the initial time t_0 and the last time t^*. On the other hand, user u_i can take multiple pseudonyms in the middle of those sequences. If user u_i can take more than or equal to K pseudonyms at a given time t, we say that user u_i satisfies (K,t)-pseudonym location privacy.

We now formally define the notion of (K,t)-pseudonym location privacy as follows.

Definition 4 ((K,t)-pseudonym location privacy): Given a user u_i, we say that a location function W satisfies (K,t)-pseudonym location privacy if there exist more than or equal to K pseudonym assignment functions $N_0, N_1, \ldots, N_K$ that are multi-path consistent such that every $N_l(u_i, t)$ for $l = 0$ to K outputs a distinctive pseudonym.

Time	p_1	p_2	p_3
t_0	$\{u_1\}$	$\{u_2\}$	$\{u_3\}$
t_1	$\{u_1, u_2, u_3\}$	$\{u_1, u_2, u_3\}$	$\{u_1, u_2, u_3\}$
t_2	$\{u_1, u_2, u_3\}$	$\{u_1, u_2, u_3\}$	$\{u_1, u_2, u_3\}$
t_3 (=t*)	$\{u_1\}$	$\{u_2\}$	$\{u_3\}$

Fig. 5. Example matrix A.

Time	Exchangeable pseudonyms
t_1	$\{p_1, p_2\}$
t_2	$\{p_2, p_3\}$
t_3	$\{p_1, p_3\}$

Fig. 6. Example list AM.

IV. Privacy evaluation algorithm

We present a privacy evaluation algorithm for computing how many possible pseudonyms each user u_i could have at a given time t. The algorithm takes two data structures $A[t,i]$ and $AM[t]$ as inputs. The matrix $A[t,i]$ contains a set of users who can possibly take a pseudonym p_i at time t. Initially, for all i, each field $A[t,i]$ contains the set of all users U except for $A[0,i]$ and $A[t^*,i]$, which only contains a user u_i. Figure 5 shows an example of matrix A. The list $AM[t]$ contains a set of pseudonyms that can be exchanged by their owner users at time t. The example AM in Figure 6 shows that pseudonyms p_1 and p_2 can be exchanged by time t_1.

Taking A and AM as inputs, the algorithm keeps updating the content of A propagating the constraints at the both ends and outputs the final A, with which we can check how many pseudonyms a given user u_i takes at time t.

Algorithm 1 is the main program, which iteratively calls two functions O and I until matrix A cannot be updated any more. The function O sequentially narrows down the entries $A[t,i]$ at time t by computing all the possible mappings from pseudonyms to users at time t using the mapping information at time $t-1$. The function I performs this task in the reverse order.

Algorithm 1 Main program.

```
1: while 1 do
2:     prevA ← A
3:     A ← O(A, AM)
4:     A ← I(A, AM)
5:     if A = prevA then
6:         break;
7:     end if
8: end while
9: return A
```

Algorithm 2 shows how the function O computes possible user-pseudonym mappings sequentially. The function O takes A and AM as inputs and updates A as follows.

The function *compPossibleMappings* in line 3 computes all the possible pseudonym-user mappings at time $t-1$ from A as follows:

$$compPossibleMappings(A,t) = \{(u_1,\ldots,u_n) \mid \forall i : u_i \in A[t,i] \wedge \forall i,j : u_i \neq u_j\}.$$

We represent such a mapping as a sequence of users. For example, mapping (u_1,u_2,u_3) means that u_1, u_2, and u_3 own pseudonyms p_1, p_2, and p_3, respectively. Line 3 stores such possible mapping in variable *psec* and computes all possible pseudonym-user mappings by applying all possible pseudonyms exchanges specified in $AM[t-1]$. The variable *seq* in line 4 maintains all the user-pseudonym mappings at time t while iterating the for loop on each pseudonym-user mapping at time $t-1$. The function *compCurrentSeqs* is formally defined as follows:

$$compCurrentSeqs(A,AM,t_1,t_2,pseq) = \{seq \mid seq \in exchangeable(pseq, AM[t_2]) \wedge \forall i : seq[i] \in A[t_1,i]\}$$

where the function *exchangeable* returns a list of possible pseudonym-user mappings derived from *pseq* considering a list of exchangeable pseudonyms in list $AM[t_2]$. Finally, line 6 updates matrix A by replacing ith row with the new row computed from the user-pseudonym mappings in *seq* using the function *replaceRow*. The outermost while loop iterates this operation sequentially from time $t = 1$ to t^*.

Algorithm 2 Function O for computing possible user-pseudonym mappings sequentially.

```
1: for t = 1 → t* do
2:     seq ← ∅
3:     for all pseq ∈ compPossibleMappings(A, t − 1) do
4:         seq ← seq ∪ compCurrentSeqs(A, AM, t − 1, t, pseq)
5:     end for
6:     A ← replaceRow(A, t, seq)
7: end for
8: return A
```

Similarly, Algorithm 3 shows how the function I computes possible user-pseudonym mappings in the reverse order.

Algorithm 3 Function I for computing possible user-pseudonym mappings in the reverse order.

```
1: for t = t* → 1 do
2:     seq ← ∅
3:     for all nseq ∈ compPossibleMappings(A, t) do
4:         seq ← seq ∪ compCurrentSeqs(A, AM, t − 1, t − 1, nseq)
5:     end for
6:     A ← replaceRow(A, t − 1, seq)
7: end for
8: return A
```

Example: Consider the matrix A in Figure 5 again. At time t_0, only mapping $(u_1,u_2,u_3) \rightarrow (p_1,p_2,p_3)$ is possible. Therefore, the function *compPrevSeqs* in line 3 returns the sequence (u_1,u_2,u_3), and that sequence is stored in variable *psec*. Next, the function *compCurrentSeqs* computes the possible mappings at time t_1. If we look up $AM[1]$ in Figure 6, we learn that pseudonyms p_1 and p_2 are exchangeable at time t_1. Thus, we obtain two possible mappings (u_1,u_2,u_3) and (u_1,u_3,u_2). This implies that $A[1,1] = \{u_1\}$, $A[1,2] = \{u_2,u_3\}$, and $A[1,3] = \{u_2,u_3\}$, and the function *replaceRow* takes care of this task.

V. Discussion and Future work

We plan to prove that it is possible to construct every multi-path consistent path function N from matrix A produced by the algorithm in Section IV. The proof should show that the algorithm should satisfy both the *soundness* and the *completeness* properties. The soundness property requires the algorithm not to produce any pseudonym assignment function N that is not multi-path consistent, and the completeness property requires it not to miss any valid assignment function N. We expect that simple inductive proofs would be sufficient for our purposes.

In addition, there are a few possible extensions of the algorithm. First, an adversary might know that some location in the middle of a user's path is associated with a particular user. For example, the adversary might know a user's office location at daytime. We can handle such additional external knowledge of an adversary with a minor modification of the algorithm. We just need to define an initial matrix A where some elements in A corresponding to known intermediate locations contains a single user. Second, it is desirable to keep longer path segments in a data set as long as that set preserves given privacy metrics. We plan to extend the current algorithm such that it determines the minimum number of items in an array AM that are necessary to achieve given privacy metrics. Third, we would like to consider a realistic, weaker assumption that an adversary only obtains a *partial* data set, which does not contain users' all path information. We expect that there is a better strategy to disguise the users' actual paths while satisfying the privacy metrics.

VI. Related work

Several researchers [7], [10], [11], [13], [14] propose fine-grained access-control schemes based on rules for protecting location privacy in pervasive environments. Here, their focus is to provide a flexible policy language for protecting identifiable location data of mobile users. Hengartner [7] supports access-control policies considering the granularity of location information and time intervals. Myles [13] provides a XML-based authorization language for defining privacy policies that protect users location information. Users must trust a set of validators that collect context information and make authorization decisions. Those schemes allows a user to define fine-grained access-control policies. Apu [11] provides users with an intuitive way of defining access control policies, which represent physical boundaries surrounding the users. However, no previous scheme considers the issue of inference based on the mobility patterns of users.

Location privacy has been studied heavily in the context of the anonymization and obfuscation of location data (See [12] for a comprehensive survey). The focus of research in this area is to ensure that no anonymized and/or obfuscated data is associated with an individual. For example, Gruteser [6] proposes a scheme that changes the granularity of location information to ensure that each location contains at least k users (i.e., k-anonymity).

Using pseudonyms is a promising way to make location data unlinkable to a particular user. Beresford and Stajano [2] first discuss the idea of dynamically changing pseudonyms in a mix zone where multiple people meet, in order to prevent an adversary from linking two pseudonyms of the same user. However, they only consider the situation where an adversary only has a local view of users' movements observing pseudonyms of entering or leaving the same mix zone. Hoh and Gruteser [8] present a path perturbation algorithm that adds noises to original location data so that each user can construct alternate possible paths by exchanging his pseudonym with those of other users when they meet at the same place. However, their scheme does not consider an adversary's external knowledge that can associates each user with a particular home location, as we assume in this paper. On the other hand, our scheme does not add noises to location data to increase the number of points where multiple users meet. Instead, our algorithm computes the number of all the combinations of users' valid alternate routes that satisfy the home location constraints.

Buttyán et al. [3] study the effectiveness of changing pseudonyms in the context of vehicular networks. They evaluate the linkability of consecutive pseudonyms assuming an adversary who can monitor the location traces of vehicles at a limited number of places. Hoh et al. [9] also consider the linkability of two consecutive points assuming their spatio-temporal correlation and propose a sampling scheme in space where location data is collected only when mobile users pass particular geographical areas. We are more concerned with the indistinguishability of a user's *global* paths rather the unlinkability of pseudonyms in *local* areas. Also, their adversary model is different from ours; that is, the adversary in our model can obtain location data with pseudonyms at any places although the adversary cannot physically see the movements of users in any limited area.

VII. Conclusions

In this paper, we present a dynamic pseudonym scheme for constructing confusing paths of mobile users to protect their location privacy. We introduce a formal definition of location privacy based on pseudonyms and show an efficient verification algorithm for determining whether each user in a given location data set has sufficient number of possible paths to disguise the user's true movements. Future work includes providing a correctness proof of the algorithm and showing the effectiveness of our psuedonym-based scheme with actual data sets.

Acknowledgments

This research is supported by the Strategic Joint Research Grant for NTT and Research Organization of Information and Systems (ROIS) and by the Grants-in-Aid for Scientific Research C, 11013869, of Japan Society for the Promotion of Science.

References

[1] Denise Anthony, Tristan Henderson, and David Kotz. Privacy in location-aware computing environments. *IEEE Pervasive Computing*, 6(4):64–72, 2007.

[2] Alastair R. Beresford and Frank Stajano. Location Privacy in Pervasive Computing. 2(1):46–55, January-March 2003.

[3] Levente Buttyán, Tamás Holczer, and István Vajda. On the effectiveness of changing pseudonyms to provide location privacy in vanets. In *Proceedings of the 4th European conference on Security and privacy in ad-hoc and sensor networks*, ESAS'07, pages 129–141, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] Dentsu draffic. http://itpro.nikkeibp.co.jp/article/JIREI/20121005/427881/.

[5] Google maps. http://maps.google.com/.

[6] Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of Mobisys 2003: The First International Conference on Mobile Systems, Applications, and Services*, San Francisco, CA, May 2003. USENIX Associations.

[7] Urs Hengartner and Peter Steenkiste. Access control to people location information. *ACM Transactions on Information and System Security (TISSEC)*, 8(4):424–456, 2005.

[8] Baik Hoh and M. Gruteser. Protecting location privacy through path confusion. In *Security and Privacy for Emerging Areas in Communications Networks, 2005. SecureComm 2005. First International Conference on*, pages 194 – 205, sept. 2005.

[9] Baik Hoh, Marco Gruteser, Ryan Herring, Jeff Ban, Daniel Work, Juan-Carlos Herrera, Alexandre M. Bayen, Murali Annavaram, and Quinn Jacobson. Virtual trip lines for distributed privacy-preserving traffic monitoring. In *Proceedings of the 6th international conference on Mobile systems, applications, and services (MobiSys)*, pages 15–28, New York, NY, USA, 2008. ACM.

[10] Jason I. Hong and James A. Landay. An architecture for privacy-sensitive ubiquitous computing. In *Proceedings of the 2nd international conference on Mobile systems, applications, and services (MobiSys)*, pages 177–189, New York, NY, USA, 2004. ACM.

[11] Apu Kapadia, Tristan Henderson, Jeffrey J. Fielding, and David Kotz. Virtual Walls: Protecting Digital Privacy in Pervasive Environments. In *Proceedings of the Fifth International Conference on Pervasive Computing (Pervasive)*, volume 4480 of *LNCS*, pages 162–179. Springer-Verlag, May 2007.

[12] John Krumm. A survey of computational location privacy. *Personal Ubiquitous Computing*, 13(6):391–399, 2009.

[13] Ginger Myles, Adrian Friday, and Nigel Davies. Preserving privacy in environments with location-based applications. *IEEE Pervasive Computing*, 2(1):56–64, January-March 2003.

[14] Vagner Sacramento, Markus Endler, and Clarisse de Souza. A privacy service for location-based collaboration among mobile users. *Journal of the Brazilian Computer Society*, 14(4):41–57, 2008.

[15] Tsuyoshi Seike, Hiroya Mimaki, Yusuke Hara, Ryo Odawara, Tomohiro Nagata, and Masayuki Terada. Research on the applicability of "mobile spatial statistics" for enhanced urban planning. *Journal of the City Planning Institute of Japan*, 46(3):451–456, 2011.