A COLLABORATIVE PROTOCOL FOR PRIVATE RETRIEVAL OF LOCATION-BASED INFORMATION

David Rebollo-Monedero, Jordi Forné, Laia Subirats

Telematics Engineering Dept., Technical University of Catalonia
C. Jordi Girona 1-3, E-08034 Barcelona, Spain

Agustí Solanas, Antoni Martínez-Ballesté

UNESCO Chair in Data Privacy, Dept. of Computer Engineering and Mathematics, Rovira i Virgili University
Av. Països Catalans 26, E-43007 Tarragona, Spain

ABSTRACT

Privacy and security are paramount for the proper deployment of location-based services (LBSs). We present a novel protocol based on user collaboration t**o** privately retrieve location-based information from an LBS provider. Our approach neither assumes that users or the LBS can be completely trusted with regard to privacy, nor relies on a trusted third party. In addition, user queries, containing accurate locations, remain unchanged, and the collaborative protocol does not impose any special requirements on the query-response function of the LBS. The protocol is analyzed in terms of privacy, network traffic, and LBS processing overhead. We show that our proposal provides exponential scalability in the probability of guaranteed privacy breach, at the expense of a linear relative network cost.

KEYWORDS

Location-based services; private information retrieval; location privacy; trusted third parties; untrusted user collaboration

1. INTRODUCTION

The opening up of enormous business opportunities for location-based services (LBSs) is a result of the recent advances in wireless communications and positioning technologies. 3G technology makes mobile wireless communications faster than ever, and highly accurate positioning devices using GPS technology are widely accessible to the general public [7]. Due to the massive use of these technologies [2], an unprecedented amount of data is fleetingly traveling through high-speed networks from all over the world. Some of these data refer to users' private information such as their locations and preferences, and it should be handled carefully. The improper management of users' private data is a matter of considerable public concern and it could decelerate the deployment of LBSs. Location privacy and users' security are of paramount importance. If privacy and security issues are guaranteed, LBSs will become one of the most important representatives of the information and communications technologies (ICTs) in the 21st century.

The way LBSs are accessed by users is changing rapidly. The simplest form of information exchange in an LBS involves a user and an LBS provider *P*. The former sends a simple query *Q* containing some sort of identification information ID, their location L and a request for information *I* that the user wants to retrieve from *P*. Thus, a simple query sent from *U* to *P* may be Q ={ID_U, *L*,*I*}={ID_U,x_U, y_U, "Where is the nearest Italian restaurant?"}. Fig. 1 depicts this communication scheme.

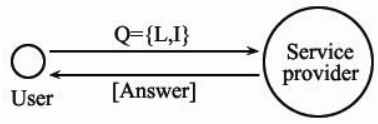


Fig. 1. Simple communication scheme between an LBS user and an LBS provider.

When users send their current locations to the LBS, they are not always guaranteed that the LBS will manage their data honestly and will refrain from any misuse. Consequently, more sophisticated mechanisms for location-based information retrieval are needed, which must protect the users' privacy. Most of the solutions proposed in the literature to address the LBS privacy problem are based on trusted third parties (TTPs), i.e., entities which fully guarantee the privacy of their users. Although this approach is widely accepted, it simply moves users' trust from LBS providers to intermediate entities. By doing so, LBS providers are no longer aware of the real locations and identities of their users; trust, and by extension power are handed over to intermediate entities such as psedonymizers and anonymizers. The problem is that users are not necessarily satisfied about completely trusting intermediate entities or providers, especially after the recent scandals related to the disclosure of personal data by this kind of trusted entities. See Fig. 2 for a graphical representation of TTP-based schemes.

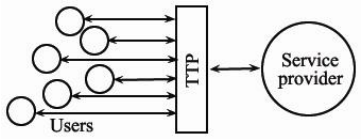


Fig. 2. Communication scheme between an LBS user, an intermediate trusted entity and an LBS provider.

The main difference between the simple communication scheme depicted in Fig.1 and the TTP-based one [6,14] is that in the latter the set of intermediate entities can be expected to be smaller than the number of service providers. Therefore, intermediate entities can be well known and the risk of trusting a dishonest entity is reduced. Pseudonymizers are the simplest intermediate entity between LBS users and providers. Anonymizers [1,10,13,14] are the most sophisticated option in TTP-based location privacy. However, many users would prefer to trust nobody and, consequently, TTP-free schemes [3,4,5,11,12] enter the arena. These represent a substantial change of paradigm. See Fig. 3 for a graphical representation of a TTP-free scheme. Instead of trusting a third party, users collaborate to protect their privacy. Moreover, there is no need to trust the users one collaborates with.

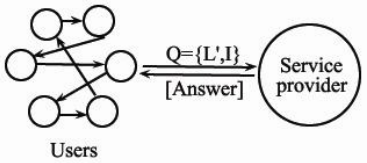


Fig. 3. Communication scheme between a set of collaborative LBS users and an untrusted LBS provider. Location information may not be the real one (L), but a perturbed one (L'). No TTP is used.

1.1 Contribution and plan

In this paper, we present a novel approach to privately retrieve location-based information from an untrusted LBS provider. Our method does not rely on TTPs but on the collaboration among multiple users to achieve privacy, despite the fact that users may not be completely trusted. Instead of defining common cloaking areas into which users become anonymous or sharing a perturbed bogus location, our method mixes queries from many users and prevents LBS providers from knowing which query refers to user. One of the main strengths of the protocol is that it benefits from an exponentially decreasing probability of guaranteed privacy breach, at the expense of only linearly increasing relative communication costs, with respect to the size parameters of the trellis. We provide a theoretical analysis of the probability of coincidental privacy breach. Furthermore, we carry out Monte Carlo simulations to verify the theoretical results and to investigate policy modifications of our protocol.

The rest of the paper is organized as follows. Sec. 2 presents our protocol for private, TTP-free, location-based information retrieval through user collaboration. The theoretical analysis of this protocol in terms of privacy is developed in Sec. 3 while the Monte Carlo simulation analysis is developed in Sec. 4. Finally, conclusions are drawn in Sec. 5.

2. A COLLABORATIVE PROTOCOL FOR PRIVACY IN LBSs

In this section we present a collaborative protocol that enables a number of users to interact with an LBS in a way that protects the privacy of their queries and replies. This is achieved in spite of two assumptions. First, it is assumed that neither the LBS nor other cooperating users can be completely trusted regarding the disclosure or a user's private information. Secondly, both the queries and the replies contain accurate information that may not be perturbed. Sec 2.1 makes our assumptions more precise. The privacy protocol proposed is described in Sec. 2.2, which relies on the existence of a cooperative structure of users.

2.1 Assumptions

In the following, we describe the assumptions on which our collaborative privacy protocol has been built:

- Users are allowed to cooperate but no party can be completely trusted, thus no TTP is available.
- Queries sent to the LBS must be kept private and accurate, thus they may not be perturbed. In particular, noise may not be added to the users' accurate location information to protect their privacy.
- The privacy protocol must be completely transparent to an arbitrary query-response function implemented by the LBS. This prevents, for instance, the use of cryptographic mechanisms operating on the assumption of a reduced response space, or a lookup table implementation of the query-response function.
- Knowledge of the user ID is inherent to the communication system, and no form of anonymization is possible, through a TTP or otherwise. IDs may neither be shared nor exchanged among users.
- Communication between any two parties is confidential and authenticated.
- Messages exchanged between users may be encrypted for the LBS to further strengthen confidentiality. In practice, this would require that the LBS participate in any collusion against a user's privacy.

Clearly, the last two hypotheses may be satisfied by the existence of a public key infrastructure (PKI), not necessarily online. The very last requirement, in particular, could be fulfilled by encrypting messages with the public key of the LBS.

Finally, we shall assume the existence of a secure mechanism by means of which users may organize themselves and adhere to a privacy protocol involving certain message exchanges. Particularly, we shall assume that there is a way to create and efficiently maintain collaborative structures, which is robust against denial-of-service attacks.

The creation and maintenance of the trellis structure is detailed in [9]. Creating and maintaining the ad hoc network structure needed for our protocol has been shown to be feasible with a small number of nodes. This may be sufficient in practical applications because our proposal does not need a large number of participants, due to the exponentially low likelihood of privacy breach. However, an interesting challenge arises from the fact that the protocol may be improved by devising a completely secure and more efficient mechanism to create and maintain collaborative structures, and to enforce the privacy protocol presented, in particular against denial of service attacks, and for large-scale structures [8].

We shall see that the LBS receives a list of queries, some of them forged, together with a list of subscriber IDs. This allows billing systems based either on flat fees or on the number of queries submitted, while preventing billing based on privacy-sensitive properties such as query length.

2.2 Query Permutation on a Trellis of Users

Consider first the simplest case when a single user must access an LBS. One way to ensure that the LBS is unable to completely ascertain the user's actual information interests is for the user to accompany his queries with forged ones. Unfortunately, this may represent a significant overhead in terms of network traffic and LBS processing.

To preserve privacy at a reasonable network and LBS processing cost, we propose the following protocol, based on query permutation in a trellis of users. More specifically, users form a trellis of m rows and n columns as shown in Fig. 4.

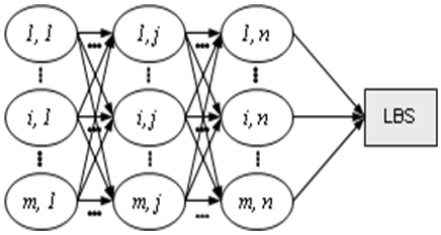


Fig.4. Query permutation on a trellis of users.

In this setting, only users in the first column generate forged queries and send them along with their authentic queries to users in the second column. In general, as illustrated in Fig. 5, user (i,j) in row i and column j receives permuted queries from users in column j-1 when j>1, or forges queries when j=1. Next, the user adds their own query, permutes the resulting list, and finally, splits it and sends each part to different users in the following column j+1 if j<n, or to the LBS if j=n. The choices regarding the permutation of the list, its splitting, and the users the parts are sent to, may be random.

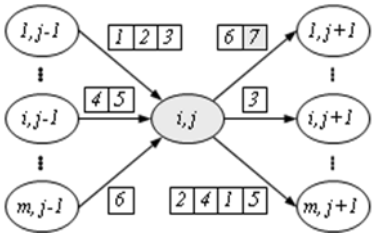


Fig.5. User (i,j) adds their own query to those received from the previous column, permutes the resulting list, splits it and sends the parts to users in the next column.

For 1<j<n, in order to guarantee that the privacy of user (i,j) be completely compromised regardless of how queries are split and transmitted, it seems that 2m users, namely all users from the previous column and the next one, must collude. However, m+1 users are enough if we are satisfied with the coincidental configuration where all users in column j-1 collude with the user in column j+1 which happened to receive the query.

2.3 Query forwarding policies

Intuition suggests that the random query forwarding policy may be improved by enforcing maximum diffusion of queries from a user to the column of recipients. For example, if (i,j) has at least m queries, sending at least one to each of the users in j+1 may have a positive impact in terms of probability of privacy breach. We shall see this is the case in the experiments reported in section 4. Such policy guarantees that nodes in j+1 do not need to generate forged queries to protect their privacy against users in j+2. On the other hand, nodes in the first column would still need to forge m-1 queries each, to enforce this throughout the trellis.

More generally, it is possible to carry out the protocol described in this section in slightly different ways, with consequent variations in performance in terms of privacy and number of messages. For example, for

$1<j<n$, user (i,j) could split its own query into portions sent to all users in column $j+1$. In this way, a single malicious user in the next column does not suffice for a complete privacy breach. These portions should be properly tagged in order for the LBS to recombine them. However, if all users in column $j+1$ are malicious, they could keep track of the recombination tags to discard incomplete groups of query portions coming from (i,j), in order to compromise the user's privacy. At the other extreme, an alternative to reduce the number of messages would consist in sending the entire query list to a single, randomly chosen user from the next column.

3. THEORETICAL ANALYSIS OF PRIVACY AND COST

In this section we analyze the trellis structure described in Sec. 2.2 in terms of two contrasting aspects. On the one hand, we consider the usefulness of this structure to preserve the users' privacy and carry out simulations for two simple policies. On the other, we study the overhead cost in regard to network connections, traffic and LBS processing.

3.1 Privacy

Interested in a conceptual, preliminary analysis, we shall simply assume that users disclose their lists of queries and are willing to collude with other users to compromise a given user's privacy, with identical probability $1-t$, independently from each other, conditioned on the event that the LBS is willing to act maliciously as well. Loosely speaking, t is the probability that a user can be trusted, given that the LBS cannot. Conditioning on the event that the LBS is malicious makes the computation identical regardless of whether queries are encrypted for the LBS, and yields slightly simpler expressions, omitting a constant factor, namely the probability that the LBS acts maliciously. More realistic scenarios could of course be better characterized by more complex probability models. Finally, our privacy analysis focuses only on queries, rather than replies, due to the similarity of the alternative analysis.

Privacy is not completely compromised as long as a list of queries contains at least one query in addition to the user's. Provided that users in the first column of the trellis of Sec. 2.2 submit forged queries, any group of colluding users will be unable to ascertain authentic queries, at least without further statistical analysis.

For $j>1$, consider the case when user (i,j)'s query is known to a group of users colluding with each other and the LBS. The probability that this situation is guaranteed to happen regardless of how query lists are split and transmitted, requires collusion of the 2m users in columns $j-1$ and $j+1$, or merely the m users in column $j-1$ if $j=n$. This probability of *guaranteed complete privacy breach* (GCPB) is $p_{GCPB}=(1-t)^{2m}$ for $1<j<n$, and $(1-t)^m$ for $j=n$. By definition p_{GCPB} is a probability conditioned on the event that the LBS acts maliciously, in cooperation with the group of colluding users.

It is shown in [9] that under the mild assumption of symmetry and random query forwarding, the probability of *coincidental complete privacy breach* (CCPB) is, for $1<j<n$,

$$p_{PCCPB} = (1-t)\sum_{b_1=0}^{m}\cdots\sum_{b_{j-1}=0}^{m}\left(\prod_{k=1}^{j-1}\binom{m}{b_k}\right)t^{\sum_{k=1}^{j-1}b_k}(1-t)^{m(j-1)-\sum_{k=1}^{j-1}b_k}\times$$

$$\times\left(1-\frac{\prod_{k=2}^{j-1}b_k}{m^{j-1}}\right)^{b_1(1+f)}\prod_{k=2}^{j-1}\left(1-\frac{\prod_{l=k+1}^{j-1}b_l}{m^{j-k}}\right)^{b_k}. \quad (1)$$

The same work proves the approximation for $t\approx 1$

$$p_{CCPB} = (1-t)\left(1-\frac{1}{m}\right)^{m(j+f-1)}+o(t-1). \quad (2)$$

3.2 Network cost

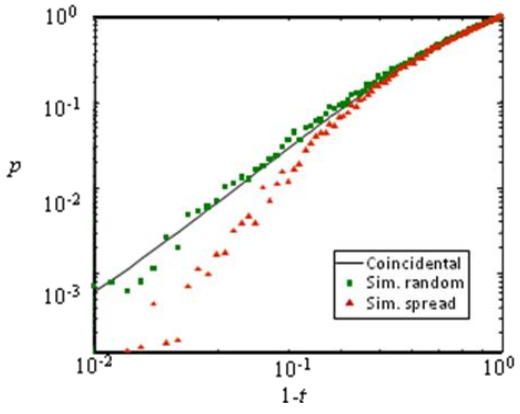
Regarding network costs, [9] shows that the number of connections and query messages which are required by our protocol in the *m*-by-*n* trellis is $O(m^2n)$, and the total number of queries transmitted through the trellis is $O(m\ n^2)$. Relative to the minimum of m n attained in an ideal scenario with benign participants, the corresponding overhead is linear, precisely, $O(m)$ and $O(n)$ respectively. Provided that users in the first column generate a fixed number of queries, the total number of queries processed by the LBS is $O(m\ n)$, and the relative overhead is asymptotically 1. In [9] we describe simple variations of the protocol to remove the need for forged query processing altogether. In addition, the privacy protocol is completely transparent to the implementation of the query-response function in the LBS. This is an advantage with respect to cryptographic PIR mechanisms operating on the assumption of a reduced response space, for instance, a lookup table implementation of the query-response function.

4. EXPERIMENTAL ANALYSIS OF PRIVACY

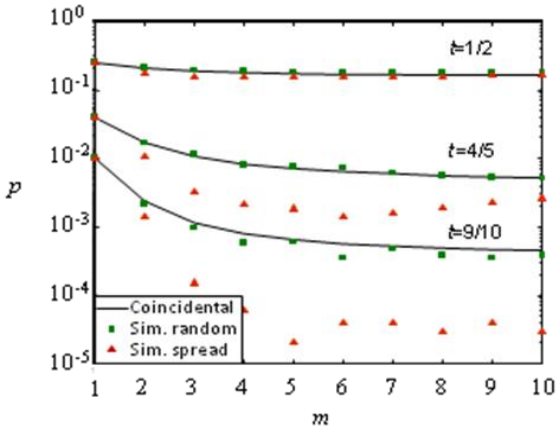
We simulated two query forwarding policies for the trellis structure described in Sec. 2.3, using the Monte Carlo technique. In the first policy users randomly forward the queries that arrive to them from the previous column. In the second policy users spread queries as much as possible. It can be seen in Fig. 6. that lower probabilities of privacy risk are obtained by spreading. After simulation of the mentioned policies, it can be seen that as t increases, the spreading policy further reduces the probability regardless of the variation of *m*, *j* and *f*. The probability is reduced drastically as *m* increases but not as significantly with *f* when the trust probability of users is high. Furthermore, it can be considered that probabilities obtained with random forwarding policy are close to the theoretical probability of CCPB of equation (1), unlike maximum spreading policy, which is considerably lower.

The following plots show averages of 10 000 simulated outcomes. When the experimental result is far from the theoretical probability, it means that the number of simulated outcomes yields insufficiently accurate averages. As a consequence, if the simulation with random forwarding is not accurate, neither will be the simulation with spread forwarding.

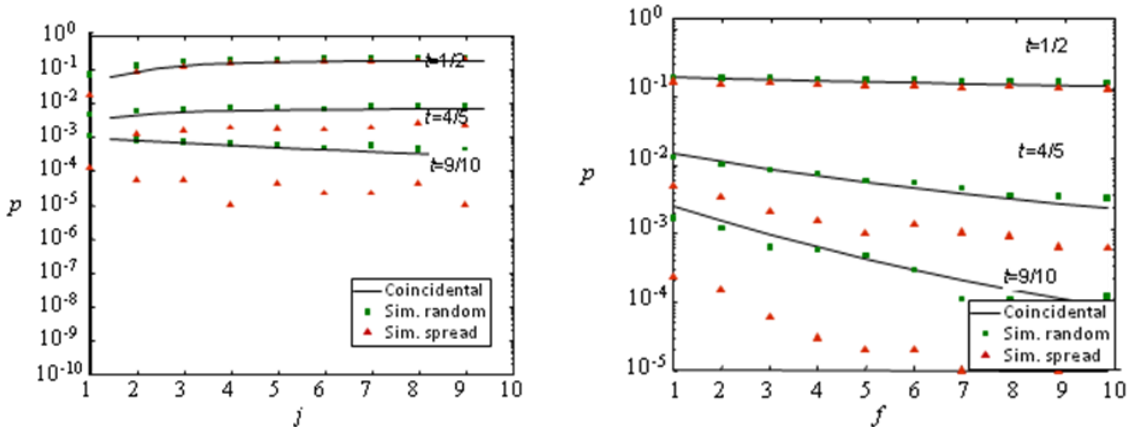
As for the graphic representation of the theoretical probability, the plots of Fig. 6 also suggest that larger *m* and *f* help reduce p_{CCPB}, in keeping with the exact formula (1). For trustworthy users, p_{CCPB} decreases exponentially with j and f, but approaches a saturation level for large m. This phenomenon is supported by the curves depicted for t=9/10 in Fig. 6. The intraquery splitting alternative commented on in Sec. 2.2 may be an additional degree of freedom in the protocol to alter the probability of CCPB, but the probability of GCPB will remain equal to $(1-t)^{2m}$. While both probabilities can always be reduced by generating additional forged queries at intermediate nodes, this comes at the cost of network traffic, and LBS processing time.



(a) m=5, j=5, f=4



(b) m=1,…,10, j=5, f=4



(c) m=5, j=2,…,10, f=4 (d) m=5, j=5, f=1,…,10

Fig. 6. Probability of CCPB(1), and simulation of the protocol with random forwarding and maximum spread forwarding.

5. CONCLUSION

Location-based services are undoubtedly essential representatives of the ICTs. Due to their inherent capability to infer private information from LBSs users, techniques to protect the user privacy are of paramount importance. In this work, we have proposed a collaborative privacy protocol for LBSs that despite not requiring TTPs, is highly scalable in terms of privacy risk. Precisely, one of the main strengths of the protocol is that it benefits from an exponentially decreasing probability of guaranteed privacy breach, at the expense of only linearly increasing relative communication costs, with respect to the size parameters of the trellis.

More specifically, users group themselves into a trellis of m rows and n columns, where queries are exchanged and permuted in such a way that privacy is preserved throughout to a scalable degree. In fact, complete privacy breach is only guaranteed under the collusion of 2m users together with the LBS, increasingly unlikely with large m. There exists a tradeoff between privacy and latency, due to the fact that users must wait for others to cooperate before sending their queries, and that a latency constraint in turn imposes an upper bound on the average number of participants in the trellis. Creating and maintaining the ad hoc network structure needed by our protocol has been shown to be feasible with a small number of nodes and sufficient for practical applications.

Furthermore, the probability was simulated following two policies. First, forwarding queries to the next column randomly, and secondly, spreading them as much as possible. The Monte Carlo simulation concluded that spreading leads to lower probabilities of privacy risk. It can be seen that as t (trust probability) increases, the spreading policy further reduces the probability regardless of the variation of m (number of users in a row), j (user column index) and f (number of queries sent by a user of the first column). The probability is reduced drastically as m increases. However, the dependence of the probability is not as significant on f, especially for high values of t. Furthermore, it can be considered that probabilities obtained with random forwarding policy are close to the theoretical probability of CCPB, unlike maximum spreading policy, which is considerably lower.

ACKNOWLEDGEMENT

This work was partly supported by the Spanish Government through projects CONSOLIDER INGENIO 2010 CSD2007-00004 “ARES”, TSI2007-65393-C02-02 “ITACA” and TSI2007-65406-C03-01 “E-AEGIS”, and by the Government of Catalonia under grants 2005 SGR 00446 and 2005 SGR 01015.

REFERENCES

1. Agostino Ardagna, C. et al.,2008. A Multi-Path Approach for k-Anonymity in Mobile Hybrid Networks. In *Proceedings of the 1st International Workshop on Privacy in Location-Based Applications*. Malaga, Spain. No. 6.
2. Benford S, Magerkurth C, Ljungstrand P. Bridging the physical and digital in pervasive gaming. *Communications of the ACM* March 2005; 48(3):54 – 57.
3. Chow C, Mokbel MF, Liu X. A peer-to-peer spatial cloaking algorithm for anonymous location-based services. GIS '06: *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems, ACM*, 2006; 171–178.
4. Domingo-Ferrer J. Microaggregation for database and location privacy. NGITS, vol. 4032, Etzion O, Kuflik T, Motro A (eds.), Springer, 2006; 106–116. *Finantial Times. 3G iPhone sales hit 1m during opening weekend. Webpage July 2008*. http://www.ft.com/cms/s/0/9d4ea864-51cc-11dd-a97c-000077b07658.html?nclick_check=1.
5. Domingo-Ferrer J., Forne J., Domingo-Ferrer J., From t-closeness to PRAM and noise addition via information theory. Privacy Stat. Databases (PSD), *Lecture Notes Comput.Sci. (LNCS),Springer-Verlag*: Istambul, Turkey, 2008.
6. Duri S, Gruteser M, Liu X, Moskowitz P, Perez R, Singh M, Tang JM. Framework for security and privacy in automotive telematics. *Proceedings of the 2nd international workshop on Mobile commerce, ACM* Press New York, NY, USA, 2002; 25– 32.
7. Finantial Times. 3G *iPhone sales hit 1m during opening weekend*. Webpage July 2008. http://www.ft.com/cms/s/0/9d4ea864-51cc-11dd-a97c-000077b07658.html?nclick_check=1.
8. L Ramaswamy BG, Liu L. A distributed approach to node clustering in decentralized peer-to-peer networks. *IEEE Transactions on Parallel and Distributed Systems* 2005; 16(9):814–829.
9. Rebollo-Monedero, D. et al., 2008. Private Location-Based Information Retrieval through User Collaboration. *Research report, UPC, Aug. 2008*. http://davidrebollomonedero.googlepages.com/Rebollo-CollaborativePrivacyforLBS08.pdf.
10. Samarati P. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering* 2001; 13(6):1010–1027.
11. Solanas, A. and Martínez-Ballesté, A., 2008. A TTP-free protocol for location privacy in location-based services. In *Computer Communications*, Vol 31, No. 6. pp 1181—1191.
12. Solanas, A. and Martínez-Ballesté, A., Fourth European PKI Workshop: theory and practice, 2008. Springer Berlin / Heidelberg, *Lecture Notes in Computer Science*, Privacy Protection in Location-Based Services through a Public-Key Privacy Homomorphism. Palma de Mallorca, Spain, pp. 362 – 368.
13. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* 2002; 10(5):571–588.
14. Sweeney L, k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems2002; 10(5):557–570.