Secure Cloaking Area Based on User Profile Similarity

Priti Jagwani and Saroj Kaushik

***Abstract*—Location-based services (LBS) are the applications of mobile technology, GIS and Internet that utilize the location of the user in order to provide the service. However, access to a person's location might allow the service provider to breach the privacy of the person. Location privacy based on k-anonymity addresses this threat by cloaking the person's location such that there are at least k – 1 other people within the cloaked area. Then instead of the exact location, cloaking area will be revealed to service provider. In this paper it has been shown that only k anonymity is not sufficient to protect location privacy. Problem of location privacy preservation is addressed via technique of anonymizing similar users together. The proposed work aims at generating a cloaking area which is containing users of similar profile. The cloaking area generated as a result is more secure. Further, for the validity of this claim statistical techniques have been applied and it has been proved that cloaking area having more similar users is safer. Experimental evaluations have been performed to calculate the area of cloaking region and time required to generate it. It has been found that for enhanced privacy levels our system puts up a little overhead.**

***Index Terms*—Cloaking area, k anonymity, profile, location privacy, anonymization.**

I. Introduction

In the last decade, mobile communication, internet and field of GIS has enjoyed unprecedented growth all over the world. The recent advances in these technologies propelled the increase of many context aware services. Location-based service (LBS) is a particular example of context aware services. LBS is the request of usable, personalized information delivered at the point of need, which includes information about new or interesting products and services, promotions automatic updates of travel reservations, etc. Location of the user is a key input for getting location based service. On the very other hand knowledge of a person's location can be used by an adversary to physically locate the person. So, LBS users have genuine concerns about their personal privacy and security.

To address this trepidation about privacy many solutions are available in literature. K anonymity is one of the prevalent solutions which is very popular. In this, instead of revealing the exact location, a bounding box is reported containing at least k people. This bounding box is commonly known as cloaking area/region. This cloaking area, is an extended area from the exact position of a mobile user, and is computed by the anonymization server/ middleware which is considered as a trusted party. As a result the exact location is abstracted within the region. Middleware in turn delegates this cloaking area along with query request to location service provider(LSP). This provides for more security, since one cannot be individually identified within the bounding box.

Computation of cloaking area/region (CR) is well studied in literature. K anonymity based CR extends a cloaking area until there are 'K-1' other users included [1]. While l-diversity based cloaking area schemes extends the area until there are 'l-1' semantically different locations included [2]. Thus there can be two different CRs for a single user based on any two different schemes (like l diversity and k anonymity).

Authors in [3] presents a novel location privacy protection technique, which protects the location semantics from an adversary. In this scheme, location semantics are first learned from location data. Then, the trusted anonymization server performs the anonymization using the location semantic information by cloaking with semantically heterogeneous locations. Xue *et al.* in [4] presented the concept of Location Diversity. Location Diversity improves spatial k-anonymity by ensuring that each query can be associated with at least l different semantic locations.

Authors in [5] shown that satisfying k-anonymity is not enough in an environment where a group of colluded service providers collaborate with each other, a user's privacy can be compromised. They presented a detailed analysis of such attack on privacy and proposed a new privacy definition called s-proximity. But in this work value of K as well as value of 'S' is being asked by user as a part of their profile. Shin, Heechang *et al* in [6] addressed the problem of privacy preservation via anonymization. They extended the notion of k-anonymity by proposing a profile based k-anonymization model. The proposed approaches generalize both location and profiles to the extent specified by the user. Authors in [7] proposed a semantic aware scheme for cloaking. Cloaking can be done based on various metrics like l diversity, k anonymity etc. there can be more than one cloaking regions for a user based on different schemes [8].

In this paper it has also been argued that k anonymity alone is not sufficient to protect privacy of a user especially when the profiles of user present in a cloaking area are diverse. A detailed analysis of attack model and adversary knowledge is presented in Section II. In this work notion of k anonymity is refined by including k-1 similar profiles in the cloaking area. Intuitively, one can say that cloaking area having similar profiles will be more secure as compared to the area generated by taking random profiles. For this similarity indexes between the user profiles presented in that particular area have been taken. It has been found that a cumulative value of similarity of a Cloaking region (CR) containing similar profiles is significantly greater than similarity of another CR containing random users. We

Manuscript submitted August 25, 2015; revised November 10, 2015.

Priti Jagwani is with School of IT, Indian Institute of Technology, Hauz Khas, New Delhi 110016, India (e-mail: jagwani.priti@gmail.com).

Saroj Kaushik is with the Department of Computer Science and Engineering, Hauz Khas, New Delhi 110016, India (e-mail: Saroj@cse.iitd.ac.in).

DOI: 10.7763/IJET.2016.V8.933

proved this claim statistically using technique of dummy variable regression.

The technical contributions of proposed work are as follows:

- This work proposed strengthens the fact that k anonymity alone is not sufficient for protecting privacy of users. Privacy can be compromised if there is sufficient diversity in the profiles of users present in a cloaking area.
- We generated a secure cloaking area containing similar profiles and proved statistically that the generated cloaking area is secure. To our knowledge this is the first work to provide the statistical proof of the fact that CR having similar profiles is more secure.
- Experiments have been performed to calculate overhead occurred in terms of more time required to generate cloaking area. Also area of the cloaking region will be more for our approach but both are within the acceptable limits.

Remaining paper is organized as follows: Next section contains the details about attack model and adversary knowledge, Section II contains proposed method along with the description of dataset and statistical proof. Experiments and evaluations are presented in Section IV. Finally Section V concludes the paper and contains future work also.

II. Attack Model and Adversary Knowledge

In our setup a cloaking area has been generated by using a trusted middleware model. In the model, a mobile user requests a service to the middleware/anonymization server which is a trusted entity. In this request, the mobile user specifies its exact current position. Then the middleware computes the cloaking area using either a number of users (location k-anonymity) or locations (location ℓ-diversity) nearby the user's position. The cloaking area is passed to Location service provider (LSP, an untrusted entity). LSP returns all results related to the cloaking area. The middleware filters out unnecessary results and gives back the result corresponding to the mobile user's current location. As a result, the exact position is not exposed to LBS applications, because a cloaking area is used instead of the position. Though the delegating middleware knows the exact position of a mobile device, LSP only see an abstracted range of an area.

We assume that the LSP or any other adversary is well aware of anonymity level i.e. value of K and static profile data of all the users that have made query at some time is also known to the adversary. Although the generated cloaking area provides some degree of privacy, it is vulnerable to privacy attacks irrespective of the schemes used to generate it. If cloaking area is generated using l diversity it can suffer from location similarity attack [3] while if the cloaking area is generated using k anonymity approach it may suffer from heterogeneity attack and conformity attack [5].

In location similarity attack [3], if the cloaking area includes only semantically similar locations, the adversary would be able to infer semantic meanings from the extended area. For example, if the cloaking area only includes an elementary school, high school, and university, then the adversary could infer that a mobile user is doing work related to 'teaching' or 'studying'. In heterogeneity attack given by [5] if the members in the anonymity set are too much diversified and the query requester possess some exclusive identifiable property, then he/she may no longer be remain indistinguishable. For example, a query for women hospital has been issued. In response to this, the CR and corresponding anonymity set containing only one female with all male users, is vulnerable to identification.

TABLE I: Sample Dataset

Age	Work Class	education	marital status	Occupation	relation ship	Race	Gender
39	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male
50	Self-emp-not-inc	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male
38	Private	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male
53	Private	11^{th}	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male
28	Private	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female
37	Private	Masters	Married-civ-spouse	Exec-managerial	Wife	White	Female
49	Private	9^{th}	Married-spouse-absent	Other-service	Not-in-family	Black	Female
52	Self-emp-not-inc	HS-grad	Married-civ-spouse	Exec-managerial	Husband	White	Male

III. Proposed Method

The main focus of the solution is on minimizing the probability of re-identifying the actual query requester by generating a secure cloaking area. We aim to include the users similar to that of query issuer in cloaking area. Profile similarity between users can be calculated using any similarity index available in literature. We have used Jaccard index for this purpose. This similarity index is in the form of a number varying between 0 and 1. Now user profiles present in a cloaking area are having a similarity value with respect to the query issuer.

A. Dataset Used

For the proposed solution we have taken 'Adult Census' dataset from UCI ML repository. This dataset consists of profiles of 32560 users [9]. Sample dataset is shown in Fig.

1. Each profile is having at least 10 attributes describing about personal and demographic features of the users. Some attributes of them are discrete, and some are continuous. Further continuous attributes are discretized by using disjoint ranges and each profile is converted into a vector form. Table I is presenting a sample of the dataset being used.

The above described dataset is having only profile data but for our experiments location of users on the map is also needed. So apart from this Minnesota Web-based Traffic Generator (MNTG) to generate traffic [10] objects on the spatial map has been used. Further, the profile vectors from 'Adult Census' dataset are associated with the objects of traffic model generated.

B. *Statistical Proof of Security*

Intuitively, it is very clear that cloaking area having more similar profiles will be more secure in terms of privacy. More the users having similar profiles in an area less will be the identification probability of the query issuer. The example described below establishes the above claim.

Two cloaking regions CR1 and CR2 are taken. The similarity index of profiles is calculated for CR 1 and CR2 using Jaccard's index. Table II is showing similarity indexes of users of CR1 and CR2 with the query issuer. These similarity indexes of CR1 and CR2 are regressed using a dummy variable regression technique. Now to prove that the difference between the averages of CR1 and CR2 is significant, a test hypothesis has been coined. This test hypothesis is :

H0: $\beta = 0$; (null hypothesis)

Ha: $\beta >< 0$; (alternative hypothesis)

where β is the difference of averages.

Test Statistics will be:

$t = \beta / Se(\beta)$.

where Se is the standard error in β.

TABLE II: SIMILARITY VALUES OF USERS IN CR1 AND CR2 (W.R.T QUERY ISSUER)

Users	CR1	CR2
U1	.6250	.3750
U2	.6667	.41
U3	.5	.4
U4	.444	.4
U5	.6667	.41
U6	.7143	.25
U7	.5556	.4286
U8	.5714	.6250
U9	.3000	.5714
U10	.3750	.3750
Average	***.54187***	***.4245***

Value of $Se(\beta)$ has been determined by Linest function. Using t table it has been observed that on permitting 1%, 5% and 10% error, value of t always exceeds critical value so null hypothesis has been rejected. This indicates that β is significantly larger which means that difference between the two averages is significant. So we can say that CR1 is more secure than CR2.

C. *Generating Secure Cloaking Area*

A location privacy system with a trusted middleware which creates anonymization group considering profiles of the users present in that area, has been proposed. Main steps to generate a secure cloaking area are:

1) A cloaking area around the user/ query issuer is generated using Nearest Neighbor Cloak (NNC). In NNC method K-1 nearest neighbors around the query issuer are located and a minimum bounding box enclosing those users is regarded as the cloaking region. Although any standard method available in literature for generating cloaking area can be used. Cloaking area generated will be abounding box around the query issuer's location.
2) Profile similarity index between the query issuer and all users is calculated using Jaccard's index.
3) Now check whether the similarity index of all users is greater than a specified value (let's say α). Increased value of α signifies more strict privacy requirements. For our experiments value of α as 0.5 has been taken.
4) If the area generated is not having K-1 profiles satisfying the above requirement, cloaking region should be expanded using an incremental strategy to cover at least K-1 profiles whose similarity index is greater than or equal to α.

IV. EXPERIMENTAL EVALUATION AND RESULTS

A prototype version of the application (middleware) which generates cloaking area has been implemented. Performance of our system has been evaluated and the findings are summarized in the figures given below. Performance of the system was measured in terms of the metrics: CR Construction Time and CR Size (w.r.t to total data space).

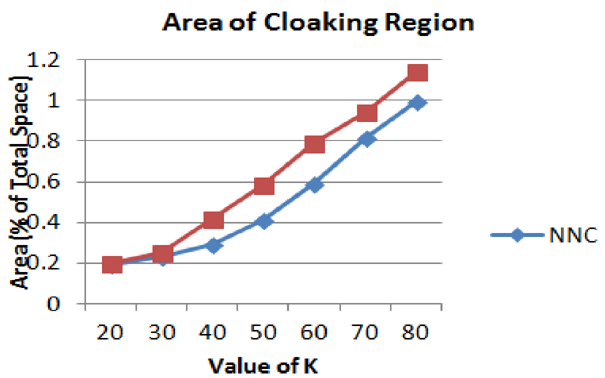

Fig. 1. Area of cloaking region.

These metrics measured the time required for constructing a cloaked region and the size of the constructed CR. The graph in Fig. 1 demonstrate that our approach yields cloaked region with a nominally large size which is acceptable considering the enhanced level of privacy it offers

CR construction time is shown in Figure 2. It has been observed that CR construction time has also increased when similar profiles of users are considered but this increase also is not huge.

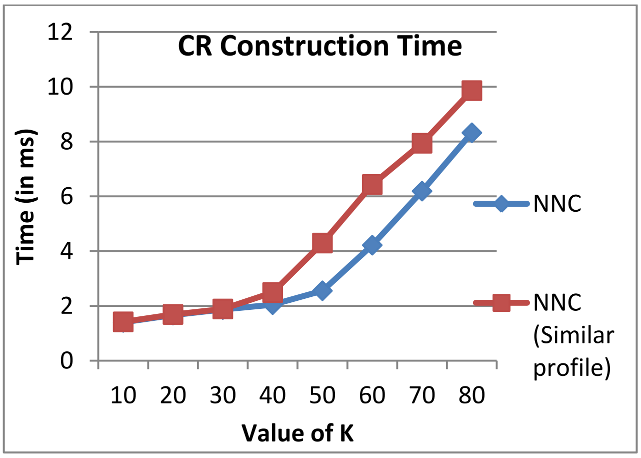

Fig. 2. CR construction time.

From the above figures it is evident that both the parameters (CR construction time and CR area) show higher values for increased anonymity levels but for low anonymity levels (i.e. value of K till 30) CR construction time and CR area, are almost similar to existing NNC method of cloaking area generation. This low level of anonymity suffices for various practical applications. Moreover, the overhead occurred is for achieving considerable privacy amplification.

V. Conclusion

This paper proposes novel location privacy protecting techniques for generating a secure cloaking area. In this work notion of K anonymity has been extended by proposing a profile based K-anonymization model that provides privacy even when profiles of mobile users are known to an adversary. Location privacy is protected by generating cloaking area containing profiles similar to that of query issuer. Statistical results validate our proposed claim. Experiments have been done to determine the overhead incurred. This overhead is calculated in terms of increased area of cloaking region and increased time required in generating that cloaking area. For low values of anonymity, overhead is almost negligible while for higher anonymity levels overhead is marginal with significant increase in privacy protection.

References

[1] J. W. Johnston, "Similarity indices I: What do they measure?" 1976

[2] H. I. Kim, Y. K. Kim, and J. W. Chang, "A grid-based cloaking area creation scheme for continuous LBS queries in distributed systems," *Journal of Convergence*, vol. 4, no. 1, pp. 23-30, 2013.

[3] B. Lee, J. Oh, H. Yu, and J. Kim, "Protecting location privacy using location semantics," in *Proc. the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1289-1297.

[4] M. Xue, P. Kalnis, and H. K. Pung, "Location diversity: Enhanced privacy protection in location based services," *Location and Context Awareness*, 2009, pp. 70-87

[5] C. S. Hasan, S. Ahamed, and M. Tanviruzzaman, "A privacy enhancing approach for identity inference protection in location-based services," in *Proc. 33rd Annual IEEE International Computer Software and Applications Conference*, 2009, vol. 1, pp. 1-10.

[6] H. Shin, V. Atluri, and J. Vaidya, "A profile anonymization model for privacy in a personalized location based service environment," in *Proc. 9th International Conference on Mobile Data Management*, 2008, pp. 73-80.

[7] M. Li, Z. Qin, and C. Wang, "Sensitive semantics-aware personality cloaking on road-network environment," *International Journal of Security & Its Applications*, vol. 8, no. 1, 2014.

[8] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "I-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, 2007.

[9] M. Lichman, "UCI machine learning repository irvine," CA: University of California, School of Information and Computer Science. (2013)

[10] MNTG. [Online]. Available: http://mntg.cs.umn.edu/tg/index.php

Priti Jagwani is a PhD student at Indian Institute of Technology, Delhi, India. She is working with Aryabhatta College, University of Delhi, India. Her research efforts focus on privacy in location based services.

Saroj Kaushik is a professor at Computer Science and Engineering Department of Indian Institute of Technology Delhi, India. Her primary research interests are Artificial Intelligence, Natural Language Processing. In addition, she directs research in the areas of artificial intelligence, location based services, recommender systems and natural language processing. She received her PhD degree from IIT Delhi.

.