

Anonymization Techniques for Preserving Data Quality in Participatory Sensing

Tishna Sabrina¹, Manzur Murshed², and Anindya Iqbal³

¹Department of EEE, United International University, Dhaka, Bangladesh. tishna.sabrina@eee.uiu.ac.bd

²Faculty of Science and Technology, Federation University, Churchill Vic 3842, Australia. manzur.murshed@federation.edu.au

³Department of CSE, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh. anindya@cse.buet.ac.bd

Abstract—Participatory sensing is a revolutionary new paradigm where citizens voluntarily sense their surroundings using readily available sensing devices such as mobile phones and share this information for mutual benefit of community members. To encourage ample participation of users, ensuring their privacy is inevitable. Existing techniques that attempt to protect location privacy with spatial cloaking suffer from irrecoverable data quality degradation. To the best of our knowledge, very few works provided a solution preserving high data quality/utility at the destination server, however, suffered from unacceptable computational overhead. This paper presents an improved deterministic alternative and also a faster variant by exploiting several optimization issues. Theoretical formulations and extensive simulation results are presented to establish the applicability of our proposed techniques.

Keyword – Anonymization; privacy; participatory sensing.

I. INTRODUCTION

Participatory sensing system (PSS) facilitates community people share data for mutual benefit. Unlike web applications, data is sensed from point of interests (POIs) at the proximity of places participating users visit in course of their daily life using ad-hoc sensing devices mounted on cell phones, vehicles, etc. The sensed data is sent to servers via some lightweight inexpensive wireless communication networks. Servers combine data from users and reply to queries accordingly. Consumer price information sharing is one of the most potential applications of PSS. Community people may report price of goods by scanning dockets or capturing advertisement photo to the server. Later on, when someone queries about the cheapest provider of specific goods within a certain area, server may reply matching the requestor's location with previously received shop-price information. Dong *et al.* [1] presented such an application, *PetrolWatch*, where a participatory sensing network of mobile cameras automatically collects fuel prices and serves the users by reporting the cheapest fuel station of the user's locality.

Since an observer has to provide her identification for claiming incentive, she can easily be associated with her location privacy. With some prior information, knowledge of location may also lead to the compromise of inferable privacy e.g., reporting nearby a specialized medical treatment facility may assist speculating the reporter's medical condition if someone knows only the time of his doctor's appointment.

The research challenge of protecting privacy of participants in PSS has naturally gained attention of researchers in recent times [2]–[11]. While ensuring privacy, quality of data is also needed to be maintained to provide reliable response to queries. Here, data quality implies the ability to map an object to its attribute at the desired end. For example, if the fuel price at the suggested fuel station is not the cheapest, due to loss of data quality,

dissatisfied users will not participate in future. Finding an acceptable privacy-quality trade-off is indispensable to ensure voluntary participation of critical mass. However, very few works in literature could maintain quality/utility of retrieved data at the destination. For example, in *PetrolWatch* application, spatial cloaking techniques such as the one used in [2] can only offer a geographic region where more than one candidate objects may exist relevant to the reply of a query and the user will become confused. The philosophy of these techniques is to introduce uncertainty by reducing precision, which do not serve the purpose of many PSS applications.

Therefore, the ultimate requirement of privacy preserving techniques from PSS is that each observation from a participant can be transmitted with sufficient anonymity and, on the other side, the destination server can de-anonymize individual data with acceptable accuracy. The aim of this paper is to provide a solution to this significant research challenge.

We propose a deterministic greedy k -anonymization technique. The concept of k -anonymity states that an observation is k -anonymous if the observed object is indistinguishable from $k - 1$ other objects. Our *Probabilistic Greedy Anonymization Scheme* (PGAS) technique [3] that achieves data quality suffering from high computational overhead, i.e., $O(N!)$ time where N implies the number of objects in PSS. Consequently, for many application scenarios with $N > 7$, the PGAS technique is not practical. This is a major limitation as many PSS applications have to inevitably deal with more than N objects in the region of interest. To overcome this limitation, unlike PGAS, we design a deterministic greedy heuristic. We refer to the proposed technique as the *Deterministic Greedy Anonymization Scheme* (DGAS). The fast variation of DGAS is named *Fast Deterministic Greedy Anonymization Scheme* (FDGAS) which works with highest degree of anonymity, i.e., $k = N - 1$.

II. SYSTEM MODEL

The basic entities of a participatory sensing system are mobile nodes (MNs) or users that collectively sense attributes of N POIs and report them to the application server (ApS). The ApS replies to user query about the attribute of any specific POI. However, the reports sent to ApS are anonymized with the help of a third party anonymization server (AS). The AS performs a directional anonymization such that efficient data quality can be reached with minimal received collective data by ApS.

Let us term the k -anonymized form of an observation report (OR) by an MN as *Anonymized Rule* (AR). For example, when 3-anonymity is desired, \$10 price observed for POI 1 may use AR {1, 2, 3}: \$10 to anonymize POI 1 with POI 2 and POI 3. The detail of the

greedy anonymization approach is discussed in the subsequent section.

Similarly, we may term the reports as Report towards ApS (RApS). For example, the RApS for an observation of POI 1 by MN having user id $u1$ observed at time t may assume the form as $[u1, \{1, 2, 3\}: \$10, t]$. Upon receiving successive RApS the ApS continues performing joint decoding with the help of the decoder algorithm until it reaches desired decodability.

The other component of the system is the adversary model. The adversary assumes position near the ApS and tries to capture as many reports as possible and decode them jointly using the same decoding algorithm used by ApS. They can be tackled by dividing POIs in a region into several groups, each served by its dedicated AS, and allowing a user to either sense POIs from only one group or use separate user id for each group. We refer the readers to our prior work [3] for more detail on this.

III. DETERMINISTIC GREEDY ANONYMIZATION SCHEME

First, we are going to discuss the process by which AS generates ARs from received observations. The approach is a greedy one that tries to augment POI-attribute association with generation of new AR considering the already generated ARs. We define a couple of relevant notions before discussing the concept.

Definition 1 (Decodability): *A particular outcome of an anonymization scheme satisfies D-decodability iff D or more POIs can be associated to their correct attributes. N-decodability is also referred to as full decodability.*

Definition 2 (Possible Anonymization Subsets): *Possible Anonymization Subsets (PAS) are the anonymized subset parts of all possible ARs corresponding to a particular POI for a fixed k and N . The total number of PAS corresponding to a POI is $\binom{N-1}{k-1}$.*

Definition 3 (Conforming Attribute Assignment Set): *A Conforming Attribute Assignment Set (CAAS) is a subset of all possible POI-Attribute assignment set where conforming N price tuples are listed eliminating the non-conforming ones, each time an AR is generated.*

Each permutation of POI-attribute assignment is checked against a generated AR and only the conforming ones constitute the CAAS.

Definition 4 (Anonymized Rules Set): *An Anonymized Rules Set (ARS) is a set of generated ARs selected so far by the greedy algorithm of AS. Each time a new AR is generated and being included in ARS, the CAAS is updated.*

As the ORs arrive at AS, our greedy anonymization scheme aims to choose an AR that will maximally reduce the size of CAAS. The ultimate target of our proposed scheme is to attain a single cardinality CAAS which may be formalized as follows.

Objective 5 *Achieving CAAS cardinality 1 with minimal observation.*

At the fresh start, when the AS has no information, i.e., no OR has yet been reported to AS, the cardinality of all

possible POI-attribute assignment permutations equals to $N!$. As the ORs start arriving, the AS is to give suggestion in a way to achieve CAAS-cardinality 1. Keeping that objective in mind we develop our greedy algorithm that focus on finding AR that reduces CAAS-cardinality maximally. Whenever AS gets an OR, it checks the all possible ARs for it and selects the one which will reduce CAAS-cardinality maximally.

Consider an example of *PetrolWatch* with four POIs having prices \$10, \$20, \$30, and \$40, respectively. In this $N = 4$ system, the initial size of all possible POI-attribute assignment permutation set is $N! = 24$. Now let us observe a simple example how one-by-one seven ARs are generated by AS for seven incoming ORs $[1, \$10], [3, \$30], [2, \$20], [4, \$40], [2, \$20], [1, \$10]$ and $[2, \$20]$. When the first OR $[1, \$10]$ is received by the AS, for desired anonymity $k = 3$, there are three possible ARs that can be ranked based on their ability of CAAS-cardinality reduction. As any of the three possible ARs, $\{1,2,3\}: \$10$, $\{1,2,4\}: \$10$, and $\{1,3,4\}: \$10$, can reduce CAAS-cardinality by 6, we can pick any. Let us assume that we pick $\{1,2,3\}: \$10$ for the 1st observation, after which cardinality of CAAS reduces to $N! - 6 = 18$. When the second OR $[3, \$30]$ arrives, among the three possible ARs, $\{1,2,3\}: \$30$ can reduce cardinality by 6, whereas both the other two $\{1,3,4\}: \$30$ and $\{2,3,4\}: \$30$ can reduce that by 2. Hence, the AS selects the AR $\{1,2,3\}: \$30$ in response to this observed report and after applying that the cardinality of CAAS reduces to $18 - 6 = 12$. Thus, by adding a new AR to the ARS, some of the existing conforming tuples no longer remain conforming to the new set of ARS. Effectively, only the tuple representing correct POI-attribute association for all POIs survives as the single cardinality CAAS.

At the ApS end, initially as it has no RApS yet, the decoding starts with all possible POI-attribute assignment permutation set with cardinality $N!$. Upon receiving each RApS, it cancels out the non-conforming tuples from the CAAS.

IV. OPTIMIZATION ISSUES

In this section, we investigate the scopes of optimization to improve decodability.

A. Recurrent Observations

In [3], all possible $\binom{N-1}{k-1}$ ARs are considered for each new observation; whereas DGAS only considers ARs that are not included in the already generated ARS. This is significant in reducing the PAS size from which a suitable AR is to be selected. Selecting repeating ARs by the AS cannot improve decodability as proved by the following theorem.

Theorem 1. *Performance of the decoder is independent of the order of ARs and duplicate ARs do not improve the performance.*

Proof. Irrespective of the order of ARs, the conforming attribute N -tuples are fixed for the set of ARs and duplicate ARs neither make any non-conforming tuple conforming nor does it cause any conforming tuple to become non-conforming. $\square$

B. Recurrent Anonymizations

At the service provider is end, the sooner we reach at single cardinality CAAS the better the purpose is served. From that realization, we focus on selecting ARs that reduce CAAS-cardinality maximally at every stage. While doing this, we make an observation which is established by the following theorem.

Theorem 2: *The anonymized subset part of an already selected AR for a particular observation can perform best cardinality reduction as compared to other possible subsets when selected for a different observation in a $k = N - 1$ scenario.*

Proof. For a $k = 3$ and $N = 4$ scenario, let us assume that the very first observation is $[i_1, a_{i_1}]$. At initial stage, any AR can be selected randomly as all possible ARs result in same cardinality reduction for they are all equally likely at that very first instance. If the AR $\{i_1, i_2, i_3\}$; a_{i_1} is selected for the first OR it implies that a_{i_1} cannot be the attribute of the POI i_4 . Then, for the next observations of the remaining 2 POIs, the same anonymization subset part $\{i_1, i_2, i_3\}$ can give best cardinality reduction if selected. For the other possible ARs, some of the non-conforming attribute N -tuples have already got eliminated once the previous AR has been selected. However, upon selecting the previous AR at the first place, all of the non-conforming attribute N -tuples for selecting next AR with the same anonymized subset part representing a different observation are all kept conforming. That means, after m number of observations, $C_m \geq C'_m$, where C_m and C'_m denote counts of cardinality reduction using this theorem and otherwise respectively. $\square$

Note that this theorem is valid only for $k = N - 1$, which is the best possible privacy preserving case. Instead of DGAS if we follow this theorem for encoding, the encoder benefits from less memory complexity and less temporal complexity which is shown later.

C. PAS-based Decoding

Inspired by the Theorem 2, we find that for the special cases where, $k = N - 1$, decoding can also be made less computationally complex by using an approach termed as PAS-based decoding. According to this approach, the decoder checks in the ARS whether for any single attribute, a complete PAS for any of the POIs has appeared or not. Once it obtains a complete PAS for a particular POI, then that POI is actually decoded and can now be correctly associated with that single attribute. Next the decoder checks in the ARS whether PAS for any other POI having the already decoded POI/s as the member of the subset, has appeared or not. For example, if we already have obtained PAS_{i_1} , then the decoder will search for $PAS_{i_1 i_2}$.

When all but a single POI get decoded, no need to look for PAS and the decoding becomes pretty straight forward. Any sole appearance of that POI in the observation sequence is enough to decode it.

In this special case of $k = N - 1$, both the encoding and decoding now has no need to count or update CAAS. It has been empirically found that it performs the same as

compared with DGAS from the viewpoint of achieving decodability while being far less complex in computation as well as in memory requirement. Hence, it may perform satisfactorily in bigger scenarios with larger values of N .

D. Fast Deterministic Greedy Anonymization

When $k = N - 1$, to anonymize an observation $[i, a]$, any subset that has already been used for anonymizing observations from any other POI but i will eliminate maximum possible non-conforming tuples. Hence, on the AS side, upon receiving an OR among all possible ARs the one that has appeared maximally in the already generated ARS for any other OR gets selected. On the decoding side, the decoder checks in the ARS whether for any single attribute, a complete PAS for any of the POIs has appeared or not. Once it obtains a complete PAS for a particular POI, then that POI is actually decoded and can be correctly associated with that single attribute. We then use the same approach to decode another POI in the reduced problem space $(N - 1, k - 1)$ and so on.

V. SIMULATION RESULT

A. Simulation Setup

In our simulated PSS, the MNs report observations to AS in a random fashion. Number of POIs (N) were varied from 4 to 6. Since the system is designed in such a way that every locality will have multiple ASs dealing with exclusive POIs, for the same types of POIs 4 to 6 is quite realistic. All the results presented here are obtained by averaging 1000 simulation runs using a custom simulator.

B. Decodability Performance

Fig. 1 depicts the D -decodability performance of DGAS which varies for different values of D for a certain N and k . Performance curve for each set up is identified as $N, k(D)$ where $D \in \{1, 2, \dots, N\}$. With the increase in desired level of decodability, the required number of user observations (M) also needs to be increased. For example, to achieve 1-decodability i.e., to become confirm about one POI-attribute association among the total of four POIs where $k = 3$, the number of required $M = 7$. However, to arrive at full decodability the number of necessary reports, $M = 18$. In some application scenarios, it may be sufficient to have less than full decodability with requirement of fewer M .

C. Comparative Performance

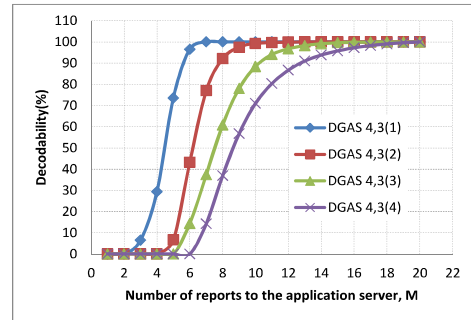


Fig. 1. Data quality trend when $N = 4$ and $k = 3$, for all possible D -decodabilities, where $D \in \{1, 2, \dots, N\}$ generated from M observation reports.

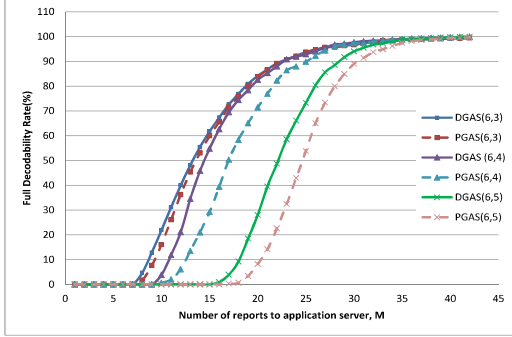


Fig. 2. Data quality trend when $N = 6$ and $k \in \{3,4,5\}$, for DGAS and PGAS as compared from M number of observation reports.

Fig. 2 compares decodability performance of DGAS with a Random subset selection technique for different values of N and $k = N - 1$. Here, the alternative technique called “Random” refers to a simple, distributed random selection mechanism of k -anonymous subsets by individual MNs without the help of any coordinated anonymization algorithm. Performance curve for each set up is identified as (N, k) where $N \in \{4,5,6\}$. The figure shows that in all cases, a certain proportion of the simulation runs achieved 100% full decodability with significantly fewer observations when we used our proposed DGAS approach instead of the random one. As the number of POI increases, the gap between the decodability performance curves also increases.

Fig. 3 presents the comparative performance of our proposed DGAS with PGAS [3] for the same values of N and k . For all cases, a certain proportion of the simulation runs achieved 100% full decodability with almost same ORs for both DGAS and PGAS. However, prior to reaching 100% full decodability, DGAS is running over PGAS. For example, our proposed DGAS can reach 28% full decodability with $M = 20$, whereas PGAS can reach only 8.2% full decodability with same number of ORs, when $k = 5$. This gap increases when operating with higher degree of anonymity, i.e., higher value of k .

VI. RELATED WORK

To protect PSS privacy, several techniques have been used such as Mix-network, pseudonym, encryption-based etc. Mix-network based schemes such as LAP [4] depend on the trustworthiness of intermediate nodes which is a hard assumption. Pseudonym based location-aware scheme [5] continuously change identification and hence cannot

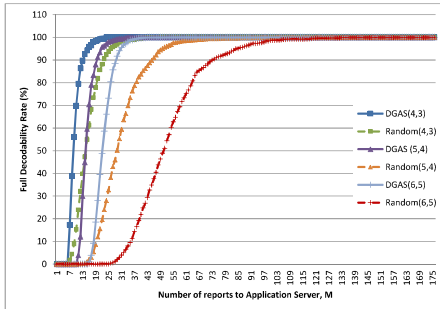


Fig. 3. Data quality when $N \in \{4,5,6\}$ and $k = N - 1$, for DGAS and Random scheme as compared from M number of observation reports.

support incentive design or preserve reputation of a user for a long period. k -anonymity based location-privacy schemes such as [6] may suffer from privacy attack of adversary participants or third-party. To address this challenge, a distributed k -anonymity based scheme has been proposed in [7] that incorporate delay in real time operation and not suitable where fine-grained information is required. Reference [8], [9] use cryptographic techniques and combines voting approaches to support users’ reputation. However, encryption on small sensed data in energy-constrained devices is not feasible in real world. A hybrid privacy-preserving mechanism has been proposed in [11] that combine anonymization, data obfuscation, and encryption techniques to increase the privacy of the users while improving the quality of information and the energy consumption.

VII. CONCLUSION

We have addressed the highly significant problem of location anonymization while maintaining data quality in the context of participatory sensing. The proposed deterministic strategy DGAS and its computationally efficient variation FDGAS overcome the limitations of our previous probabilistic strategy PGAS [3]. These are likely to contribute in making participatory sensing more popular technology in future to the community ensuring privacy of participants without compromising accuracy of data.

REFERENCES

- [1] Y. Domg, S. Kanhare, C. T. Chou, and N. Bulusu, “Automatic collection of fuel prices from a network of mobile cameras,” In Proc. DCSS, 2008.
- [2] K. Huang, S. Kanhare, and W. Lu, “Preserving privacy in participatory sensing systems,” Computer Comm., 2009.
- [3] M. Murshed, A. Iqbal, T. Sabrina, and K. M. Alam, “A subset coding based k-anonymization technique to trade-off location privacy and data quality in participatory sensing systems,” in Proc. IEEE NCA, 2011.
- [4] H. Hsiao et. al., “Lap: Lightweight anonymity and privacy,” in Proc. IEEE S&P, 2012.
- [5] S. Gao, J. Ma, W. Shi, G. Zhan, and C. Sun, “Trpf: A trajectory privacy preserving framework for participatory sensing,” IEEE Trans. On Information Forensics and Security, 2013.
- [6] K. Vu, R. Zheng, and L. Gao, “Efficient algorithms for k-anonymous location privacy in participatory sensing,” in Proc. INFOCOM, 2012.
- [7] D. Christin, D. Bub, A. Moerov, and S. Kasem-Madani, “A distributed privacy-preserving mechanism for mobile urban sensing applications,” in Proc. ISSNIP, April 2015.
- [8] E. Cristofaro and C. Soriente, “Participatory privacy: Enabling privacy in participatory sensing,” IEEE Network, 2013.
- [9] A. Michalas and N. Komninos, “The lord of the sense: A privacy preserving reputation system for participatory sensing applications,” in Proc. IEEE ISCC, 2014.
- [10] F. Qiu, F. Wu, and G. Chen, “Slicer: A slicing-based k-anonymous privacy preserving scheme for participatory sensing,” in Proc. IEEE MASS, 2013.
- [11] I. Laurens, D. Mendez, and M. Labrador, “Privacy, quality of information, and energy consumption in participatory sensing systems,” in Proc. IEEE PerCom, 2014.