Towards Privacy-Sensitive Participatory Sensing

Kuan Lun Huang, Salil S. Kanhere
School of Computer Science and Engineering
The University of New South Wales, Sydney, Australia
{klh, salilk}@cse.unsw.edu.au

Wen Hu
Autonomous Systems Lab
CSIRO ICT centre, Australia
wen.hu@csiro.au

***Abstract*—The ubiquity of mobile devices has brought forth the concept of *participatory sensing*, whereby ordinary citizens can now contribute and share information from the urban environment. However, such applications introduce a key research challenge: preserving the location privacy of the individuals contributing data. In this paper, we propose the use of *microaggregation*, a concept used for protecting privacy in databases, as a solution to this problem. We compare microaggregation with tessellation, the current state-of-the-art, and demonstrate that each technique has its advantage in certain mutually exclusive situations. We propose a hybrid scheme called, *Hybrid Variable-Size Maximum Distance to Average Vector (V-MDAV)*, which combines the positive aspects of both these techniques. Our evaluations based on real-world data traces show that hybrid V-MDAV improves the percentage of positive identifications made by the application server by up to 100% and decreases the information loss by about 40%. Furthermore, our studies show that perturbing user locations with random Gaussian noise can provide users with an extra layer of protection with very little impact on the system performance.**

I. Introduction

Over the past decade, we have witnessed an explosive growth of mobile devices that are capable of capturing, processing, and transmitting high fidelity multimedia content. Furthermore, the advances in positioning technologies and VLSI fabrication processes make geo-localization an affordable feature in mobile devices. These have motivated the research community to explore an alternative sensing paradigm referred to as *participatory sensing* [2] or *urban sensing* [1], that exploits the unique characteristics of these geo-intelligent, sensor-equipped and computationally capable mobile devices. These systems have led to the emergence of several *citizen sensing* applications, wherein, the mobile phones carried by ordinary citizens collect and share information about the urban landscape.

CarTel [3] is a system that uses mobile sensors mounted on vehicles to collect information about traffic, quality of en route Wi-Fi Access Points (APs), and potholes on the road. A similar system has been proposed in [4], which exploits sensor-rich smartphones carried by passengers for monitoring road and traffic conditions. Other applications of participatory sensing include, collecting information about urban air pollution [5], cyclist experience [6] and diet [7]. In our earlier research, we have applied the concept of participatory sensing in sharing consumer pricing information in offline markets. We have designed two systems, *PetrolWatch* [8] and *MobiShop* [9], which use mobile camera phones to collect, process and deliver pricing information from petrol stations and brick and mortar shops to potential buyers.

In a typical participatory sensing application, the sensing data uploaded by the users is invariably tagged with the location (obtained from the embedded GPS in the phone or using WiFi based localization) and time when the reading was recorded, since these provide important contextual information. This can have serious implications on user privacy, since, the sensor report uploaded by the user may reveal his/her location at a particular time. Furthermore, it may be possible to link multiple reports from the same user and determine certain private information such as the location of his/her office and residence. Simple techniques such as using pseudonyms or anonymizing the reports may not always work. For example, if an adversary has *a priori* knowledge of the user's movement patterns, it is fairly trivial to deanonymize the reports. Note that, participatory sensing relies on the altruistic participation of users for successful operation. It is thus imperative that users are assured that their privacy will not be violated to encourage sufficient participation.

In recent years, a few methods have been proposed for securing location privacy in the context of participatory sensing. Cornelius et al. have proposed AnonySense [10], a privacy-preserving architecture for realizing participatory sensing applications. Their system uses the concept of *tessellation* [11] for protection the location privacy of contributing users. In tessellation, a point coordinate is generalized to a plane in space, which is referred to as a *tile*. The sensor reports uploaded by users contain the tile id rather than the absolute location. This genearlization is guided by the principle of *k-anonymity* [12], which ensures that at least k users are located within the same tile. Hence, it is impossible for an adversary to distinguish between the k users. In this paper, we argue that the generalization approach adopted by tessellation may not be particularly suited to certain applications that require fine-grained location information. For example, consider an application that collects traffic information from the mobile phones carried by vehicular passengers [4]. If tessellation is employed, a traffic report generated by a user at one particular intersection along a road will be annotated with the tile id (which encompasses a large region), rather than the exact location of the intersection. When this report is received by the application server, the aggregated location information represented by the tile is of little use, since the server cannot ascertain which road is being referred to in the report.

We suggest a simple modification to tessellation to overcome the aforementioned problem. Next, we propose to adopt *microaggregation*, a branch of statistical disclosure control techniques [13], [14], for preserving location privacy. To protect the privacy of respondents, microaggregation creates a set of equivalence classes (ECs) such that, records are collectively represented by the mean of the respective classes. These ECs can be generated based on a wide range of criteria, for example, minimum information loss. The ECS can also be made to conform with k-anonymity, such that each EC has at least k members. Since, the mean is a numerical value, microaggregation is a natural choice for conserving the numerical properties of continuous variables [14]. Furthermore, since location is often perceived as continuous data in most popular positioning technologies, it is therefore reasonable to expect that finer fidelity and higher usability of data can be achieved with microaggregation.

This paper makes the following specific contributions:

- We demonstrate the limitations of tessellation in providing contextual support for participatory sensing applications. We then show how to eliminate these drawbacks by making modifications to tessellation.
- We propose an alternative approach, microaggregation, to ad-

dress location privacy. We compare microaggregation with our modified version of tessellation and demonstrate that each scheme has certain advantages in mutually exclusive situations. To combine the strengths of these two schemes, we propose a hybrid approach called, hybrid M-DAV.

- We use real-world user traces to evaluate the performance of these privacy-enabling methods. We show that hybrid M-DAV achieves twice the percentage of positive identifications as compared to the other schemes and a 40% improvement in information loss.
- We also propose an enhancement, which perturbs the user locations with random Gaussian noise, as an extra level of protection. We demonstrate this this extension has very little impact on the system performance.

The rest of the paper is organised as follows. In Section II we present a brief overview of the two central concepts used in this paper: (i) tessellation and (ii) microaggregation. Section III outlines the system model and assumptions. We introduce the proposed privacy-preserving techniques in Section IV. Section V presents results from our evaluations. Finally, Section VI concludes the paper.

II. Related Work

Preserving the privacy of users' locations in participatory sensing is similar to safeguarding respondents' privacy in databases, which contain continuous-valued fields. Therefore, most of the concepts and methods related to database disclosure control can be potentially applied to participatory sensing. In particular, the concept of k-anonymity [12], is widely used for preserving privacy in databases as well as in participatory sensing systems.

Kapadia et al. proposed a k-anonymity technique based on generalization in [11], referred to as tessellation. Tessellation partitions a geographic area into cells. In their implementation, these cells correspond to the Voronoi polygons constructed around Wi-Fi APs. The user distribution per cell is obtained from historical AP association records and is used to cluster cells into *tiles*. A tile is an amalgamation of cells such that the collective number of users per tile exceeds the privacy requirement k. In other words, a tile is the lowest granularity with which users represent their locations. Table I shows a sample of a 3-anonymous location database based on tessellation. Further details about the tessellation process are provided in Section V.

User ID	Location	Tile ID	Class Mean
1	(1.5, 6.0)	1	(4.33, 5.17)
2	(4.5, 4.0)	1	(4.33, 5.17)
3	(4.5, 1.0)	1	(6.33, 1.33)
4	(6.5, 2.0)	2	(6.33, 1.33)
5	(7.0, 5.5)	2	(4.33, 5.17)
6	(8.0, 1.0)	2	(6.33, 1.33)

TABLE I
Anonymized location database

Microaggregation [13] is an alternative approach, that has been used for implementing database disclosure control. Microaggregation does not generalize nor suppress the values of an attribute of a database record. Instead, it replaces the values with the mean of the EC in which the record is found. An EC is a grouping of users such that the class members are as homogeneous (i.e. similar) as possible. The member similarities are often quantified by the Information Loss (IL) metric, which effectively measures the differences between records and their representations, i.e., mean of ECs. We give the complete definition of IL and its implications in Section V. There are many algorithms proposed to generate ECs with maximum within-class homogeneity [13], [15], [16]. Maximum Distance to Average Vector (MDAV) [13] is widely recognized as one of the most efficient heuristics to date. However, MDAV cannot be readily adapted to the distribution of users in the target area, because it is a fixed class size algorithm. The variable class size variant of MDAV, called V-MDAV [16], was later proposed to ameliorate this shortcoming. The rightmost column in Table I shows the result of applying V-MDAV with the six location coordinates as inputs. The algorithmic description of V-MDAV is presented in Section IV.

Domingo-Ferrer proposed a novel protocol, which leverages microaggregation to address location privacy in Location-Based Services (LBS) [17]. Their solution assumes a peer-to-peer system. A user distorts his own location by artificially adding Gaussian variable of zero mean and standard deviation σ to the latitude and longitude. The distorted location coordinates are broadcast to nearby neighbours (i.e. peers) requesting for their Gaussian-perturbed location readings. Upon receiving the responses from its peers, the user selects $k-1$ other users such that they collectively span a region delimited by the user's privacy requirement. The mean of the group formed by the user and its $k-1$ closest neighbours is then used in all messages sent to the LBS server. However, this scheme cannot be readily adopted in participatory sensing, since these systems typically utilize a client-server architecture. In this paper, we explore the use of microaggregation for preserving location privacy in the context of participatory sensing.

III. System Model and Open Problems

In this section, we first present the system model and assumptions. Next, we present an example application, which demonstrates the limitations of using tessellation in participatory sensing.

A. System Model and Assumptions

We leverage the AnonySense architecture [10] to provide participatory sensing infrastructure support, but take a different approach to address the issues of potential leakage of private location information. In particular, we focus on the privacy protection feature of AnonySense. Recall that, AnonySense employs tessellation for implementing location privacy. Tessellation requires the existence of an additional *Map Server (MS)*, which is responsible for generating the tessellation map (i.e. dividing the entire geographical region in to tiles). Users query the MS to obtain the tessellation map, which allows them to determine the appropriate tile location that should be reported with the sensor readings.

In our implementation, a similar system entity is needed. However, in our case this entity is also able to execute various microaggregation algorithms (explained in Section IV) and is referred to as the *Anonymization Server (AS)*. The sequence of operations executed when a user contributes data is as follows: a user collects data demanded by an application with its mobile device and submits reports when it has network connectivity (via 3G/WiFi). The user consults the AS prior to submitting the report. The AS runs the appropriate microaggregation algorithm and provides the user with an anonymized location, which is used to annotate the report. The application then processes and interprets the received data using the anonymized location.

We make the following assumptions: 1) the AS is independently owned by a third-praty and is isolated from attacks, 2) the AS does not collude with applications and other system entities, and 3) users periodically upload their whereabouts to the AS (or when they submit queries) and trust the server with the confidentiality of their locations. Note that, in practice it is unrealistic to demand that users trust a single system entity with their accurate locations. Hence, we propose a scheme to relax this assumption in Section IV-D.

B. Motivating Application: PetrolWatch

We now present an illustrative example to demonstrate the drawbacks of using tessellation for location privacy in participatory sensing. In our earlier work [8], we have proposed a novel application, PetrolWatch, which allows users to automatically collect, contribute and share petrol price information using camera phones. Users mount their camera-enabled mobile phones on the car dashboard. Through the use of GPS and GIS, our system knows when the vehicle is approaching a petrol station and triggers the camera automatically. Pictures of the petrol price billboard are processed by a computer vision algorithm to extract the fuel price. The prices are annotated with the location coordinates and time and uploaded to the application server. Users can query the server to locate the cheapest petrol station in their vicinity.

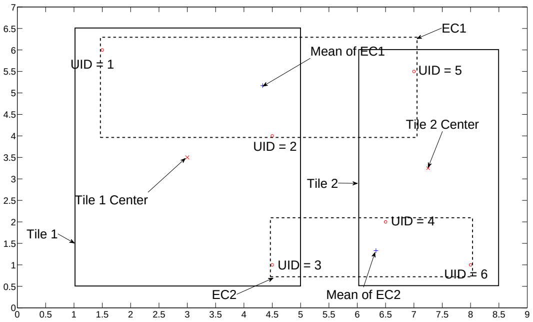

Fig. 1. Example Application: PetrolWatch

Fig. 1 illustrates a simple example of PetrolWatch assuming that tessellation is employed to provide location privacy. There are six users spread across a region of size 9km × 7km (for simplicity we assume a 2D coordinate system). This figure shows the locations of the users at a particular time instant. Assume that there is a petrol station co-located with the current location of each user (i.e. 6 petrol stations in total). Now assume that user 2, is in the process of contributing petrol price information to the application server. A query is first sent from the user to the AS requesting for the anonymized location that should be reported. Given the distribution of users, the AS constructs two tiles as shown in Fig. 1 (following the guidelines of tessellation in [11]) assuming the privacy requirement of $k = 3$ and advises the user of its anonymized location, i.e. *tile 1*. Consequently, the user annotates the report with *tile 1* instead of the actual location *(4.5, 5)*. When the report is submitted to the application, it needs to associate the received report with one of the three petrol stations located in tile 1. However, without additional information, the application is unable to determine that the petrol price included in the report corresponds to the petrol station co-located with user 2. This simple example clearly illustrates the intrinsic limitation of tessellation and serves as the primary motivation for our proposed schemes.

It should be noted that Fig. 1 shows one possible arrangement for clustering the users. It is likely that there could be other viable alternatives, which may potentially have a different impact on the performance of tessellation. A set of general instructions for tile construction are provided in [11], but there is no discussion on the impact of varying the tile configuration.

IV. Privacy Protection Approaches

In this section, we first propose a simple modification to tessellation and demonstrate how it overcomes the limitations identified in Section III. Next, we present our proposed schemes, *k*-anonymity with V-MDAV and hybrid V-MDAV. Lastly, Gaussian input perturbation is proposed to further improve location privacy.

A. Tessellation with Tile center Reporting

From the example in Section III-B, it is evident that the problem with tessellation is that it reports an entire region as the anonymized location, instead of providing the location coordinates of a point. In this regard, a natural modification to tessellation involves representing each tile by the location coordinates of the centroid of the tile. Hence, we propose a modification, wherein, user reports are annotated with the location coordinates of the center of the tile in which the user is currently located. This requires a simple modification to the AS, such that it includes the coordinates of each tile center with the tessellation map. We illustrate the operation of this modified scheme by using the same example as in Section III-B. With the above modification in place, users 1, 2, and 3 would report their positions as (3, 3.5), which is the centre of tile 1. Similarly, users 4, 5 and 6 would represent their locations as (7.25, 3.25), the centre of tile 2. This modification allows the application to better interpret the data received in the user reports. For example, comparing the Euclidean distances between the anonymized location reported by user 2 and those of the six candidate petrol stations reveals that user 2 is most likely referring to the petrol station in its vicinity.

We note here that computing Euclidean distances may not be the best strategy for the application to analyse the data received. Nonetheless, it adequately demonstrates one of the advantages of this simple modification. In the rest of this paper, we will refer to this modified version of tessellation as MT.

B. Location Anonymization with Microaggregation

Even though the above modification to tessellation works fine in most situations, it should be noted that depending on the user density, some tiles could be very large. In such cases, reporting the center of the tile may actually cause the application to incorrectly interpret the location contained in the report (this point is further elaborated in the evaluations in Section V). As an alternative, we propose the use of microaggregation for protecting location privacy in participatory sensing. In particular, we adopt the Variable-size Maximum Distance to Average Vector (V-MDAV) heuristic proposed in [16]. The V-MDAV algorithm is recursive and involves two principal successive operations: (i) Equivalence Class (EC) generation and (ii) EC extension. The former step clusters users who exhibit high geographic similarities, which is determined by their relative Euclidean distances, in groups of k. This ensures that *k*-anonymity is enforced. The latter step enables the algorithm to adapt to the user distribution by allowing geographically close users to be merged with an existing EC, despite the fact that each EC already conforms with the *k*-anonymity requirement.

We illustrate the operation of this heuristic using the same example depicted in Fig. 1. The AS generates two ECs: one encircles users 1, 2, and 5 and the other includes users 3, 4, and 6. In this approach, user 2 represents its location as the mean of the EC to which it belongs, i.e. (4.33, 5.17). This not only fulfils the k-anonymity privacy requirement (the size of each EC is 3) but also ensures that the anonymized location is in the point coordinate format.

C. Location Anonymization with Hybrid Microaggregation

We now present 2 simple examples to demonstrate that both MT and V-MDAV have their advantages in certain mutually exclusive situations. This observation motivates us to propose a novel technique that combines the best of both these methods.

Let us first consider the same example in Fig. 1. Assume that user 6 is in the process of uploading a petrol price report to the application server. We assume that the server has some background knowledge regarding the report, i.e., it knows that this report would not have referred to the petrol station in the immediate vicinity of user 4. This is a valid assumption because reports can often be filtered by other attributes, for example, the brand of the petrol station. The location data carried in the report can either be (7.25, 3.25) if MT is employed or (6.33, 1.33) in the case of V-MDAV. Assume that the application server compares the Euclidean distances of all 6 petrol stations to the location contained in the report and concludes that the report corresponds to the petrol station, which is closest to the reported location. In the case of MT, the server would incorrectly interpret that this report originated from the petrol station co-located

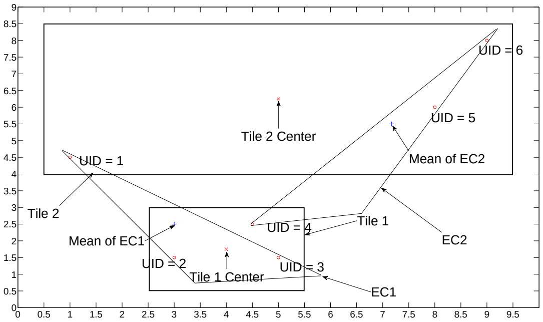

Fig. 2. An example demonstrating the benefit of Hybrid V-MDAV

with user 5. However, in the case of V-MDAV, a correct association is made with the station located in the vicinity of user 6.

Let us now consider a different example as illustrated in Fig. 2. Let us first focus on the MT algorithm. Observe that the cell in which users 2, 3, and 4 are located satisfies $k = 3$. Hence, this cell forms a tile on its own. On the other hand, the cells in which the remaining users are located need to be merged together according to the rules of tessellation. In a similar vein, with V-MDAV, users 1, 2, and 3 constitute one EC, while the remaining users are grouped into another EC. Now, assume that user 4 submits its report. The user will anonymize his location using either (4, 1.75) in the case of MT or (7.17, 5.5) in the case of V-MDAV. Using Euclidean distances for interpretation as in the previous example, the application server correctly associates the report submitted by user 4 with the petrol station located near user 4 for MT. However, V-MDAV results in an incorrect association with the petrol station near user 5.

The following observations can be made based on the above examples: 1) V-MDAV enables the application to make better decisions when the user distribution across different areas is consistent, as in Fig. 1 2) On the contrary, in areas with dense distribution of users, as in Fig. 2, MT performs better. Given that the two schemes have their advantages in contrasting situations, we propose, Hybrid V-MDAV, which attempts to combine the best of both these methods. The hybrid scheme adaptively makes a decision on whether to use MT or V-MDAV. The operation of Hybrid V-MDAV is quite simple. If the user is in a cell, which can form a tile by itself, i.e., if the number of users within the cell exceeds *k*, then MT is used. Otherwise, the algorithm switches to V-MDAV. If Hybrid V-MDAV is applied to the example in Fig. 2, then users 2, 3 and 4 would employ the MT algorithm, whereas the other users would use V-MDAV. This would overcome the incorrect association explained earlier.

D. Gaussian Input Perturbation

All of the aforementioned methods assume the existence of a trusted third-party server, which is aware of the true locations of the participating users (recall that the user queries the AS and provides its current location, each time he needs to upload a report). Clearly, this represents a single point of failure, since, if this server is compromised, the users' privacy is at risk. Further, users may not be comfortable with the idea of a server keeping track of their locations. In fact, this may be a turn off for many users and hence, they may be reluctant to participate. It is therefore, imperative to devise a strategy that does away with this requirement, without incurring substantial performance degradation.

We propose a simple perturbation scheme that artificially distorts a user's location prior to updating the AS. The artificial distortion is induced by adding a random Gaussian noise with mean μ and standard deviation σ to the X and Y coordinates of a user's location (we assume that the GPS coordinates are converted to a planar 2D coordinate system). In other words, if the current location of a user is (x, y), then the user reports its perturbed location $[x + p \times N(\mu_x, \sigma_x), y + p \times N(\mu_y, \sigma_y)]$ to the AS. The perturbation parameters, i.e., μ and σ, can be estimated from historical AP visitation records.

Assume for now that we know the number of users in each of the cells in Fig. 3 (we will explain how the details of Fig. 3 are obtained in Section V). Based on this information, we can place the users at randomly selected locations within the cell. The mean and standard deviation of these random coordinates over all cells are used as μ and σ estimates, respectively. Since the resulting σ is of the same order of magnitude as users' coordinates, a factor *p* is introduced as a scaling variable so that the perturbed value does not deviate significantly from a user's true location. *p* usually takes on a small fractional value (see evaluations in Section V).

V. Evaluations

In this Section, we present results from a simulation study that compares the performance of the three algorithms: MT, V-MDAV and hybid V-MDAV, discussed in Section IV. The simulations were conducted using real-world traces.

A. Goals, Methodology and Metrics

In our evaluations. we use the Dartmouth College campus traces that are publicly available from [18]. The traces contain log entries collected from Wi-Fi APs deployed on the Dartmouth College campus. Similar traces were used in the evaluations presented in [11]. In particular, the "*syslog/05_06*" trace under "syslog" traceset and "*aplocations*" trace under "movement" traceset are used to deduce user distributions and to plot Voronoi polygons, respectively[1]. Each record in the "*syslog/05_06*" trace represents an association, re-association or disassociation of a user device with an AP. The "*aplocations*" trace contains a list of APs deployed across the Dartmouth campus and provides information about their (x, y) coordinates as well as the floors on which they are located.

We consider a scenario wherein, a participatory sensing application similar to PetrolWatch (discussed in Section III-B) has been deployed. We assume that the application server generates tasks that require users to collect certain contextual information from some Points of Interest (PoI) in their immediate vicinity. Users who agree to participate in the application accept the tasks, collect sensor data and upload the sensor report to the server. Prior to generating the sensor report, the user polls the AS, which provides the user with the anonymized location, depending on the location privacy algorithm employed (MT, V-MDAV or Hybrid V-MDAV). The application server is aware of the true locations of all PoI. When the server receives the sensor report, it computes the Euclidean distance of each PoI from the reported location. The report is associated with the PoI that has the smallest distance.

For implementing MT, it is necessary to generate a tessellation map of the entire geographical region, which is the university campus in our scenario. In the following, we describe how this map was generated. There are 623 APs listed in the "*aplocations*" trace. In order to simplify the analysis, we perform planarization and condensation similar to [11]. Specifically, the floor numbers of APs are ignored and all APs are assumed to be located on floor 0 (planarization). Furthermore, APs located in the same building are grouped together and collectively represented by their mean (x, y) co-ordinates (condensation). Fig. 3 depicts the resulting 124 APs and their associated Voronoi polygons. We also normalize the locations of the APs so that they are confined to a region of unit square area. For generation of the user distribution per cell, we have considered records between 12pm and 6pm over a one week period from the 1st

[1] There are three separate files available for download under the "*syslog/05_06*" trace; each one of them corresponds to association records from Cisco APs, Aruba APs, and the combination of Cisco and Aruba APs. For simplicity, we only considered the traces from the Cisco AP file.

of September, 2005 to the 7th of September, 2005. The number of user associations per cell is a threshold value representing the number of users that can be statistically expected to be present in a cell for 95% of the specified time intervals. In our evaluations, this interval is assumed to be 30 minutes. There are 153 users whose distributions are marked by asterisks in Fig. 3. The coordinates of the users in a cell are randomly generated, once the threshold value for that cell is known. The cells are grouped to form tiles such that k-anonymity is attained. We use $k = 10$ in all our simulations. The tiles are shown as colored regions in Fig. 3

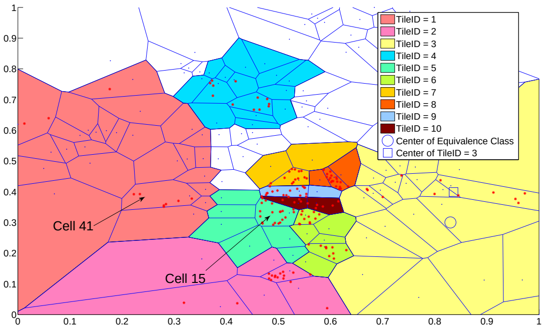

Fig. 3. Tessellation Map of the Simulation Scenario

Recall that, the users annotate the sensor reports with their anonymized locations. When the application server receives a report, it must determine the PoI that corresponds to the report. We define a metric called ***Positive Identification Percentage***, which measures the accuracy of establishing this association. This metric is defined as the ratio of the total number of positive associations to the total number of reports submitted. All the privacy-preserving algorithms being evaluated transform a user's true location to an anonymized version. Hence, it is of interest to determine how much information is lost in the transformation process. We quantify this loss of information by using the ***Information Loss*** (IL) metric, which has been commonly used to assess the performance of various microaggregation algorithms [13], [16]. IL is formally defined as follows,

$$IL = \frac{SSE}{SST}, \tag{1}$$

where

$$SSE = \sum_{j=1}^{g} \sum_{i=1}^{n} (x_i - \overline{x_j})^2, \tag{2}$$

and

$$SST = \sum_{j=1}^{g} \sum_{i=1}^{n} (x_i - \overline{x})^2, \tag{3}$$

where x_i denotes the i-th record in group j with each of the g groups containing n records. $\overline{x_j}$ and $\overline{x}$ represent the group mean and the mean of the entire dataset, respectively. SSE and SST represent the sum of squared errors with respect to the group mean and the mean of the entire dataset, respectively. Note that, SSE measures the distances between the actual locations of the users and their anonymized locations.

B. Simulation Results

We conduct a set of simulations to evaluate the positive identification percentage and information loss achieved by MT, V-MDAV and Hybrid V-MDAV. We assume that the PoI are co-located with the users. We assume that a subset of the entire user population submits reports to the application server. We vary the percentage of users reporting data from 20% to 100% in increments of 20%. The server associates each report with a PoI using the shortest Euclidean distance. Figure. 4 represents the average value of both metrics (IL and positive identification percentage).

One can readily observe that the performance of all three algorithms do no vary significantly with an increase in the number of users contributing data. Hybrid V-MDAV achieves a 40% reduction in IL as compared to MT. The performance of the hybrid scheme is marginally better than that of V-MDAV. We explain the inferior performance of MT by using tile 3 in Fig. 3 as an illustrative example. Observe that the center of tile 3 denoted by a circle is quite distant from the actual locations of the users. Recall that, in MT, users report the center of the tile as their anonymized location. On the contrary, with V-MDAV and hybrid V-MDAV the same set of users would report a much closer coordinate, which is represented as a square, as their anonymized location. As a result, the SSE is larger with MT as compared to the other two algorithms. Consequently, MT achieves higher IL. One might argue that the performance gap could be improved by shrinking the size of tile 3 such that it only includes those cells in which users are found. This is a valid argument. However, one must remember the following: 1) the tiles in Fig. 3 were constructed to fit all user distributions, which also account for the Gaussian perturbation extension and 2) to the best of our knowledge, there do not yet exist any real-time algorithms that produce optimal tessellation maps, which can adapt to the user distributions.

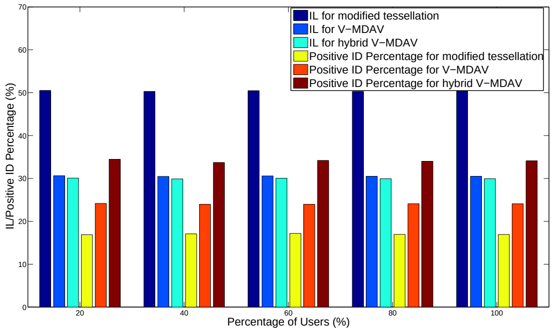

Fig. 4. Percentage of Positive Identifications and IL as a function of the percentage of users uploading reports.

Fig. 4 also suggests that there exists an inverse relationship between the two metrics, IL and percentage of positive identifications. For example, MT, which has the highest IL results in the lowest positive identification percentage. Similarly, hybrid V-MDAV, which achieves the highest positive identification percentage has the lowest IL. Observe that, hybrid V-MDAV improves the positive identifications made by the server by more than 100%, in comparison with MT. The significant improvement achieved by hybrid V-MDAV in comparison with V-MDAV (about 50%) can be explained by an illustrative example. Consider cell 15, which accommodates 20 users. According to the rules of hybrid V-MDAV, these 20 users replace their locations with the center of the cell. Since, the users are densely concentrated near the centre of the cell, the application server can interpret the true locations with high accuracy. On the other hand, V-MDAV separates these users by grouping some of them with those in cell 41 in an attempt to lower IL while keeping the size of EC in check, i.e., between 10 and 19. As a result, the reported location is somewhere in between the cells, which is not close to actual users. Hence, the application server tends to make wrong associations, which is reflected in the lower positive identification rate. It should be noted that, even the best performing hybrid V-MDAV scheme only achieves a moderate level of positive identifications. This is because the application server employs the simplistic Euclidean estimation technique for making the PoI associations. We intend to investigate alternate techniques in our future work.

C. Impact of Gaussian Input Perturbation

Next, we study the impact of Gaussian input perturbation on the performance of the three algorithms. Recall, that in this enhancement, the users do not report their true locations to the AS. Instead, a random Gaussian noise is added to the location reported to the AS. We repeat the previous set of simulations for different values of p, which range from 0.02 to 0.2 in increments of 0.02. Recall, that p is the scaling factor used for perturbing the true locations of the users (see Section IV-D). The larger the value of p, the greater is the deviation from the true location.

Figs. 5 and 6 illustrate the impact of Gaussian input perturbation on the three algorithms when 40% and 80% of users report data. Since, the results exhibit some fluctuations, we fit them with polynomials of degree 1 to reveal the general trends. As in the previous simulations, the percentage of users submitting reports has negligible impact on the performance. Furthermore, the additional input pertubation degrades the performance of all three algorithms. The level of performance degradation is more substantial for larger values of p. These results are expected since, the users are increasingly distorting their locations that are reported to the AS. Fig. 6 reveals that the performance gain of hybrid V-MDAV gradually diminishes as p increases. Increasing the value of p implies that the user distribution is more sparse, i.e., fewer cells are sufficient to provide the required level of anonymity on their own. Therefore, the V-MDAV component of the hybrid algorithm tends to dominate. As a result, the performance of these two schemes converge. The results depicted in Fig. 6 also indicate that it is possible to guarantee satisfactory performance, without requiring the users to reveal their true locations to the third-party AS. As long as the perturbation parameters are adequately chosen, the performance degradation can be limited. For example, we only observe a 5% loss when $p = 0.06$ with hybrid V-MDAV. This achieves a good balance between user privacy and system performance.

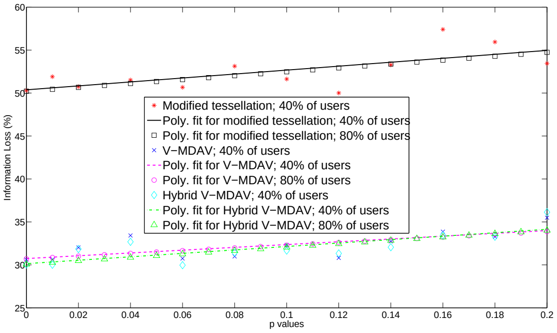

Fig. 5. Impact of Gaussian Perturbation on IL.

VI. Conclusion

In this paper, we have proposed hybrid V-MDAV for preserving location privacy in participatory sensing. Hybrid V-MDAV combines the positive aspects of tessellation and microaggregation, two popular privacy-preserving concepts. Our evaluations based on real-world data traces show that hybrid V-MDAV improves the percentage of positive identifications made by the application server by up to 100% and decreases the information loss by about 40%. Furthermore, our studies show that perturbing user locations with random Gaussian noise can provide users with an extra layer of protection with a minimal impact on the performance.

References

[1] A. Campbell, S. Eisenman, N. Lane, E. Miluzzo and R. Peterson, People-centric Urban Sensing, in *Proceedings of Second Annual International Wireless Internet Conference (WICON)*, pp. 2-5, August 2006.

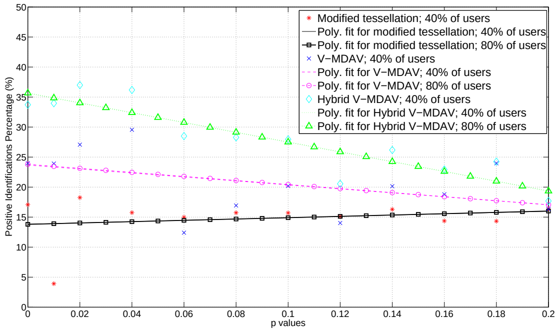

Fig. 6. Impact of Gaussian Perturbation on Positive Identification Percentage.

[2] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, M. B. Srivastava, Participatory Sensing, in *Proceedings of the World Sensor Web Workshop, in conjunction with ACM Sensys 2006*, November 2006.

[3] B. Hull, V. Bychkovsky, Y. Zhang, et. al., CarTel: A Distributed Mobile Sensor Computing System, in *Proceedings of ACM SenSys 2006*, pp. 125-138, November 2006.

[4] P. Mohan, V. Padmanabhan, R. Ramjee, Nericell: Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones, in *Proceedings of ACM SenSys 2008*, November 2008.

[5] E. Paulos, R. Honicky, E. Goodman, Sensing atmosphere, in emphProceedings of the Workshop on Sensing on Everyday Mobile Phones in Support of Participatory Research in conjunction with ACM SenSys 2007, November 2007.

[6] S. Eisenman, E. Miluzzo, N. Lane, R. Peterson, G. Ahn and A. Campbell, The Bikenet Mobile Sensing System for Cyclist Experience Mapping, in *Proceedings of ACM SenSys 2007*, November 2007.

[7] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estin and M. Hansen, Image Browsing, Processing and Clustering for Participatory Sensing: Lessons from a DietSense Prototype, in *Proceedings of the Workshop on Embedded Networked Sensors (EmNetS)*, June 2007.

[8] Y. Dong, S. S. Kanhere, C. T. Chou and N. Bulusu, Automatic Collection of Fuel Prices from a Network of Mobile Cameras, in *Proceedings of IEEE DCOSS 2008*, June 2008.

[9] S. Sehgal, S. S. Kanhere and C. T. Chou, Mobishop: Using Mobile Phones for Sharing Consumer Pricing Information, Demo Paper in *Proceedings of IEEE DCOSS 2008*, June 2008.

[10] C. Cornelius, A. Kapadia and N. Triandopoulos, AnonySense: Privacy-Aware People-Centric Sensing, in *Proceedings of ACM MobiSys 2008*, June 2008.

[11] A. Kapadia, N. Triandopoulos, C. Cornelius, D. Peebles and D. Kotz, AnonySense: Opportunistic and Privacy-Preserving Context Collection, in *Proceedings of Sixth International Conference on Pervasive Computing (Pervasive)*, pp. 162-179, May 2007.

[12] L. Sweeney, K-anonymity: A Model for Protecting Privacy, in *International Journal of Uncertainty, Fuzziness, and Knowledge-Basrd Systems*, 2002.

[13] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical Data-oriented Microaggregation for Statistical Disclosure Control, in *IEEE Transactions on Knowledge and Data Engineering*, Vol. 14, No. 1, pp. 189-201, 2002.

[14] J. Domingo-Ferrer and V. Torra. Ordinal, Continuous and Heterogeneous k-anonymity through Microaggregation. *Data Mining and Knowledge Discovery*, Vol. 11, pages 195-212, 2005.

[15] M. Laszlo and S. Mukherjee. Minimum Spanning Tree Partitioning Algorithm for Microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 7, pages 902-911, 2005.

[16] A. Solanas, A Martinez-Balleste. V-MDAV: A Multivariate Microaggregation With Variable Group Size, in *17th COMPSTAT Symposium of the IASC*, Rome, 2006.

[17] J. Domingo-Ferrer. Microaggregation for Database and Location Privacy. In *Next Generation Information Technologies and Systems-NGITS'2006*, Vol. 4032 of *Lecture Notes in Computer Science*, pp. 106-116, 2006.

[18] D. Kotz, T. Henderson, and I. Abyzov. CRAWDAD trace, url-http://crawdad.cs.dartmouth.edu/meta.php?name=dartmouth/campus.