Computer Communications 33 (2010) 1266–1280

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Preserving privacy in participatory sensing systems ☆

Kuan Lun Huang [a,*], Salil S. Kanhere [a], Wen Hu [b]

[a] *School of Computer Science and Engineering, The University of New South Wales, Sydney, Australia*
[b] *Autonomous Systems Lab, CSIRO ICT Centre, Australia*

ARTICLE INFO

Article history:
Available online 1 September 2009

Keywords:
k-Anonymity
l-Diversity
Anonymity
Privacy
Participatory sensing

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 privacy of the individuals contributing data. In this paper, we study two different privacy concepts, *k*-anonymity and *l*-diversity, and demonstrate how their privacy models can be applied to protect users' spatial and temporal privacy in the context of participatory sensing.

The first part of the paper focuses on schemes implementing *k*-anonymity. We propose the use of microaggregation, a technique used for facilitating disclosure control in databases, as an alternate to tessellation, which is the current state-of-the-art for location privacy in participatory sensing applications. We conduct a comparative study of the two techniques and demonstrate that each has its advantage in certain mutually exclusive situations. We then propose the Hybrid Variable size Maximum Distance to Average Vector (Hybrid-VMDAV) algorithm, which combines the positive aspects of microaggregation and tessellation. The second part of the paper addresses the limitations of the *k*-anonymity privacy model. We employ the principle of *l*-diversity and propose an *l*-diverse version of VMDAV (LD-VMDAV) as an improvement. In particular, LD-VMDAV is robust in situations where an adversary may have gained partial knowledge about certain attributes of the victim.

We evaluate the performances of our proposed techniques using real-world traces. Our results show that Hybrid-VMDAV improves the percentage of positive identifications made by an application server by up to 100% and decreases the amount of information loss by about 40%. We empirically show that LD-VMDAV always outperforms its *k*-anonymity counterpart. In particular, it improves the ability of the applications to accurately interpret the anonymized location and time included in user reports. Our studies also confirm that perturbing the true locations of the users with random Gaussian noise can provide an extra layer of protection, while causing little impact on the application performance.

1. 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 [3], 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, mobile phones carried by ordinary citizens collect and share information about the urban landscape.

Cartel [4] is a system that uses mobile sensors mounted on vehicles to collect information about traffic, quality of en route Wi–Fi access points, and potholes on the road. A similar system has been proposed in [5], which exploits sensor-rich smartphones carried by passengers for monitoring road and traffic conditions. Micro-Blog [6], on the other hand, is an architecture which facilitates real-time recording and sharing of multimedia contents. Other applications of participatory sensing include, collecting information about urban air pollution [7], cyclist experience [8], and diet [9]. Moreover, riding the recent wave of social networks such as Facebook and MySpace, [10] presents CenceMe, which is a novel application that exploits the capabilities of mobile phones to automatically infer people's sensing presence. In our earlier

☆ Part of this work has been published at the 5th IEEE PerCom International Workshop on Sensor Networks and Systems for Pervasive Computing (PerSeNS 09) [1]. The extension to the PerSeNS paper includes the investigation of the applicability of the concept of *l*-diversity in the context of participatory sensing; the algorithm implementation and evaluations.

* Corresponding author. Tel.: +61 406120793.
E-mail addresses: klh@cse.unsw.edu.au (K.L. Huang), salilk@cse.unsw.edu.au (S.S. Kanhere), Wen.Hu@csiro.au (W. Hu).

0140-3664/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2009.08.012

research, we have applied the concept of participatory sensing in sharing consumer pricing information in offline markets. We have designed two systems, PetrolWatch [11] and MobiShop [12], which use mobile camera phones to collect, process and deliver pricing information from service stations and brick and mortar shops to potential drivers and buyers.

In a typical participatory sensing application, the sensing data uploaded by users are invariably tagged with the location (obtained from the embedded GPS in the phone or using Wi–Fi based localization) and time when the readings are recorded, since these provide important contextual information. This can have serious implications on user privacy, since the sensor reports uploaded by users may reveal their locations at particular times. 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 [13] or suppressing user identity [14] may not always work. For instance, if an adversary has *a priori* knowledge of a user's movement patterns, it is fairly trivial to deanonymize his/her reports. Note that, participatory sensing relies on the altruistic participation of users for widespread penetration and successful operation. It is thus imperative that users are assured that their temporal and spatial privacy will not be violated to encourage sufficient participation.

In recent years, a few methods have been proposed for securing location privacy in participatory sensing systems. Kapadia et al. in [16] implemented a novel technique called tessellation, which addressed the public concern for users' location privacy in events of data contribution. In tessellation, a point coordinate is enlarged to a region, which is referred to as *tile*, containing at least k users. Sensor reports uploaded by users contain tile identifiers (tile IDs) rather than their exact locations.[1] The act of transforming a value from a finer granularity (point in a plane) to a coarser equivalent (region in a plane) is often called *generalization*. Generalization is an important class of implementation techniques for the well-known k-anonymity concept [17]. k-anonymity is a desirable property for reports collected by applications. The collection of reports received by an application is k-anonymous if it represents groups of users with the size of each group being at least k. Further, members of a group share similar values for some attributes. Tessellation, in line with the above description, is therefore regarded as an instantiation of the concept of k-anonymity. In this paper, we argue that the underlying generalization may make tessellation particularly unsuitable for certain applications which require fine-grained location information. For example, consider an application that collects traffic information from mobile phones carried by vehicular passengers [5]. 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 tile ID is of little use, since the server cannot ascertain which road is being referred to in the report.

We suggest a minor modification to tessellation in an attempt to address the above-mentioned issue. Next, we adopt microaggregation, a commonly used technique to implement k-anonymity for statistical disclosure control [21,22], as an alternative of tessellation for location privacy in participatory sensing. One of the useful properties of microaggregation is its ability to operate on continuous-valued numerical attributes. This makes it a good candidate approach to not only ensure spatial privacy (cf. tessellation) but also temporal privacy. Similar to tessellation, microaggregation protects users of participatory sensing applications by creating groups[2] of users from reports stored at application servers. Akin to tessellation, the resulting groups all have at least k members that share the same values for some selected attributes, e.g. location and time. More importantly and distinguishably, the common values can also assume numerical values, e.g. mean location of a group of users. Referring to the prior traffic monitoring example, the numerical format of spatial information may offer more contextual insight for the problem at hand. There are many implementations under the umbrella of microaggregation (see [21,22] and the references therein). In this paper, we apply a particular instance called Variable size Maximum Distance to Average Vector (VMDAV) algorithm because of its demonstrated algorithmic efficiency.

This paper focuses on spatial and temporal privacy of users, which are two universal attributes that are expected to be included in user reports for all participatory sensing applications. We presume the existence of an adversary who does not know the true values of time and location of user reports. However, the adversary has means to find out the temporal and spatial properties of his victims. For example, the adversary may overhear the conversation between Bob and his friend and find out that, Bob is scheduled for a medical treatment sometime in the afternoon on Wednesday. The goal of the adversary is to use this prior temporal information to find out Bob's medical conditions. More specifically, consider a case in which the aforementioned adversary is the administrator of a participatory sensing application, to which Bob has registered as a user. Bob employs tessellation for his location privacy when interacting with the application. The adversary is able to exploit his prior knowledge about Bob to narrow down the search among reports uploaded on Wednesday afternoon and conclude that Bob was somewhere in region A, which corresponds to the cancer treatment facility of a hospital. This allows the adversary to infer the fact that Bob is most likely to have cancer. The above attack is often called *background knowledge attack* [30]. The privacy model characterized by k-anonymity is vulnerable to compromise under these attacks. As the above example shows, even though region A is shared among k users, the lack of variation exposes Bob to what is referred to as *attribute disclosure*. One approach to circumvent this problem is to define a set of *exclusion zones* [28]. Exclusion zones refer to those sensitive areas that should not be used to replace users' true locations. In the above scenario, the cancer treatment facility constitutes an exclusion zone. The drawback of this idea is that it may reduce the penetration of the application if the envisioned deployment area involves a substantial amount of exclusion zones.

In light of the type of adversary described above, we view privacy in participatory sensing from a different perspective. We illustrate this using the same example above. Our goal is not to stop the adversary from knowing which group Bob is placed (via prior temporal knowledge). Instead, we seek to prevent him from acquiring absolute semantic information about the location of Bob. One way to achieve this is to ensure that, each group has multiple values for the location attribute. For example, instead of having (tile ID = 1) as the only location value, the group to which Bob belongs may have (tile ID = 1), (tile ID = 3), (tile ID = 7).[3] This makes it harder for the adversary to become aware of the Bob's medical conditions. This privacy model is formally characterized by the well-known l-diversity concept [30]. In this paper, we propose an l-diverse version of VMDAV to address the limitations of k-anonymous privacy methods.

In summary, this paper makes the following specific contributions:

- We demonstrate the limitations of tessellation in providing contextual support for participatory sensing applications. We then

[1] Note that users also do not reveal the precise time of a sensing event. Instead, they report the time interval over which the event takes place. For example, if a sensing took place at 12:23, the report documenting this event would log the time as [12:00 – 12:30], an interval of 30 min.

[2] A group is often referred to as an equivalence class in microaggregation literature.

[3] Assuming (tile ID = 3) and (tile ID = 7) refer to areas other than the cancer treatment facility.

show how our modified version of tessellation, TwTCR, eliminates these drawbacks.

- We propose the use of an alternative implementation, VMDAV, to address location privacy. We compare VMDAV with TwTCR 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-VMDAV.
- We demonstrate that k-anonymous techniques such as TwTCR, VMDAV, and Hybrid-VMDAV, are insufficient to defend against attribute disclosure. We therefore propose an l-diversity based extension of VMDAV. We show how this algorithm prevents attribute disclosure while providing both temporal and spatial privacy.
- We use real-world user traces to evaluate the performances of proposed privacy-preserving schemes. We first compare the performance among k-anonymous techniques, and show that Hybrid-VMDAV achieves twice the percentage of positive identifications as compared to TwTCR and VMDAV, while reducing 40% of the amount of information loss. Next we demonstrate that LD-VMDAV can consistently outperform VMDAV and show that it can prevent attribute disclosure.
- We also propose an enhancement, which perturbs user locations with random Gaussian noise, as an extra layer of protection. We demonstrate that this extension has very little impact on the performance of our proposed schemes.

The rest of the paper is organized as follows. In Section 2, we present a brief overview of the two central concepts used in this paper: (1) k-anonymity and (2) l-diversity. We also include in this Section some prior implementation developments relevant to the two core concepts. In particular, the techniques of tessellation and microaggregation are described in more details. Section 3 outlines the system model and describes a motivating application. We introduce our k-anonymous privacy-preserving techniques in Section 4. Section 4 also explores the viability of introducing Gaussian input perturbation as an extra layer of privacy protection. Section 5 is reserved for the study of l-diversity for temporal and spatial privacy in participatory sensing. A detailed explanation of the l-diversity algorithm, LD-VMDAV, is also included in Section 5. Section 6 provides results from our evaluations. Finally, Section 7 concludes the paper.

2. Related work

Preserving users' privacy 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, k-anonymity [17], has been widely used for privacy preservation in databases as well as in participatory sensing systems.

2.1. The concept of k-anonymity

The concept of k-anonymity is easy to understand. A report collected by an application is k-anonymous if it is indistinguishable, with respect to some chosen attributes, among $k-1$ other reports received by the same application. The indistinguishability is achieved by replacing the true values of selected attributes with common ones. In participatory sensing applications, especially those involving location data, value substitutions are often performed over the sensitive attributes.[4] Location is an example of sensitive attribute, since it is commonly perceived by users as confidential information. There are a multitude of algorithms implementing k-anonymity but they can be classified according to the mechanisms by which the common values are generated [21]. *Generalization* refers to the techniques where data granularity is reduced, e.g. replacing a street-level location value with a city-level equivalent. *Perturbation*, on the other hand, does not reduce data granularity but artificially changes the attribute values according to some pre-determined functions, e.g. adding random Gaussian noise to location coordinates.

2.2. Tessellation: k-anonymity by generalization

Kapadia et al. proposed tessellation, which is a k-anonymous technique primarily aimed at addressing location privacy in participatory sensing applications, as part of the AnonySense architecture in [16]. Tessellation belongs to the generalization category. It involves partitioning a geographic area into a collection of cells and amalgamating neighboring cells to form tiles, which users can use to mask their true positions. In other words, a tile is the lowest granularity with which users represent their locations. In their implementation, these cells corresponded to the Voronoi cells constructed from the locations of Wi–Fi access points (APs) on the Dartmouth College campus. The user distribution per cell was obtained from historical AP activity records and was used to cluster cells into tiles. Columns 1–5 of Table 1[5] show a sample of a 3-anonymous reports based on tessellation. The true time and location (i.e. columns 2 and 3) are included in the table for references only. In reality, these are absent from the reports submitted to applications. In other words, a user report at the application consists of the following fields: ⟨User ID, Anonymized Time, Anonymized Location⟩. Further details about how the tile IDs are decided are provided in Section 3.2. Note that, in Table 1, time values are also generalized. More specifically, time is reported at the granularity of one hour. For example, 12:31 is represented by the interval 12:00 – 13:00.

2.3. VMDAV: k-anonymity by Perturbation

Microaggregation [21] is an alternative approach to implement k-anonymity. Its operation involves creating a set of equivalence classes, within which members share common values for sensitive (in the context of participatory sensing) attributes. These common values are typically the averages of attributes. An equivalence class refers to the grouping of records[6] such that class members are as similar as possible. Member similarity is often measured by the relative distances between attribute values, e.g. Euclidean distances between location coordinates. Microaggregation is an example of perturbation techniques since, it does not generalize values of the sensitive attributes but changes them according to the average function. Many algorithms have been proposed to generate equivalence classes with maximum within-class similarity [21,23,24]. Maximum Distance to Average Vector (MDAV) [21] has been widely recognized as one of the most efficient heuristics to date. However, it has also been found to perform poorly if the distribution of records exhibited prominent features. Taking location as an example, such feature may manifest itself as regions with exceptionally populated users. The poor performance in these circumstances is due to its inability to vary the size of the resulting equivalence classes. The variable size variant of MDAV, which has been termed VMDAV, was later proposed in [24] to ameliorate this shortcoming. The rightmost column of Table 1 shows the result of applying VMDAV with the location coordinates of the six users as input. The full algorithmic description

[4] On the contrary, the value replacement in conventional databases occur on quasi-identifiers. These refer to attributes whose values can be obtained elsewhere and used to identify individuals. Postal code, gender, date of birth are all examples of quasi-identifier.

[5] A table in this paper refers to a collection of reports submitted by users and stored by a participatory sensing application.

[6] A record refers to an entry of the table stored at an application. It represents a report from a participating individual.

Table 1
Example of 3-anonymous reports maintained at the application.

User ID	Time	Location	Anonymized time (generalization)	Tile ID	Class ID: class mean
1	12:31	(1.5, 6.0)	(12:00–13:00)	1	Class 1: (4.33, 5.17)
2	12:48	(4.5, 4.0)	(12:00–13:00)	1	Class 1: (4.33, 5.17)
3	12:01	(4.5, 1.0)	(12:00–13:00)	1	Class 2: (6.33, 1.33)
4	17:05	(6.5, 2.0)	(17:00–18:00)	2	Class 2: (6.33, 1.33)
5	17:35	(7.0, 5.5)	(17:00–18:00)	2	Class 1: (4.33, 5.17)
6	17:48	(8.0, 1.0)	(17:00–18:00)	2	Class 2: (6.33, 1.33)

of VMDAV is provided in Section 4.2.

Domingo-Ferrer proposed a novel protocol, which applied microaggregation to address location privacy in Location-Based Services (LBS) [25]. Their solution assumes a peer-to-peer system. In their scheme, a user distorts his own location by artificially adding a Gaussian variable of zero mean and standard deviation σ to his latitude and longitude. The distorted location coordinates are broadcast to nearby neighbors (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 his $k-1$ closest neighbors is then used in all messages sent to the LBS server. There are still many open problems in distributed (peer to peer) participatory sensing [16]. Therefore, this scheme cannot be readily adopted in our context. In this paper, we leverage the client–server architecture, albeit the distributed counterparts are gradually gaining momentum and popularity [27].

2.4. Problems with k-anonymity

In general, k-anonymity protects user privacy by replacing attribute values with those which are common to k records. Even though this protection model is sufficient to defend against identity disclosure, it has been discovered by several authors [29,30] that k-anonymity alone cannot prevent attribute disclosure. *Identity disclosure* refers to the case where an individual is linked to a specific record in the table. *Attribute disclosure*, on the other hand, occurs when confidential properties about an individual are acquired from the semantic meaning of an attribute. To elaborate on these two privacy compromises, we use the cancer treatment facility example from Section 1 in conjunction with Table 1. We assume that the adversary knows (through mutual conversations) that his victim's medical appointment is scheduled at 12:30 p.m. This prior temporal knowledge does not permit the adversary to precisely identify which of the first three records is uploaded by the victim. In other words, identity disclosure is prevented. However, this knowledge does allow the adversary to unambiguously conclude that the victim is in (tile ID = 1), which in this scenario corresponds to the cancer treatment facility. Thus the victim is exposed to location attribute disclosure.

Two types of attack have been identified to cause attribute disclosure: (1) background knowledge attack and (2) homogeneity attack [30]. Background knowledge attack refers to the situation, wherein an adversary eliminates unlikely candidates and learns information about his victim using some prior knowledge about the individual. Homogeneity attack, on the other hand, occurs when an adversary exploits the monotony in attribute values to acquire properties of victims. Both types of attack are used in the above example to reach attribute disclosure: background knowledge (temporal information) enables the adversary to exclude the last three records, while homogeneity attack confirms his belief that the victim has been to the cancer treatment facility and is thus likely to suffer from cancer.

2.5. The concept of l-diversity

In light of the aforementioned disclosure risks, Machanavajjhala et al. [30] proposed an ingenious approach, which is now well-known as l-diversity, to further enhance the privacy of individuals. Formally a group of reports is l-diverse if these reports contain at least l well-represented values for sensitive attributes and that, a table satisfies l-diversity if all constituting groups are l-diverse. In [30], the authors propose different ways to interpret the term *well-represented*, but for simplicity, we explain the most intuitive, *distinctive* l-diversity here and use it in our proposed algorithm in Section 5. In distinctive l-diversity, the user reports are grouped such that each group has l distinct values for sensitive attributes. To illustrate this, we refer the reader to Table 1. Now assume that users are arranged in groups of 3, e.g. users: 1, 2, 3 and users 4, 5, 6, and ignore the tile ID column. The resulting representation is an example of a distinctive 2-diverse table. Specifically, users in the first group (users 1, 2, and 3) have the same time value, e.g. 12:00–13:00, but two different values for location, e.g. (4.33, 5.17) and (6.33, 1.33). This eliminates the monotony in location and thus protects users from location attribute disclosure as described in Section 2.4. The implementation of l-diversity does not require the design of new algorithms since, it has been proven in [30] that any k-anonymity algorithms are l-diversity compatible with minor changes to test conditions. An example of such an implementation can be found in Han's work [31], wherein an l-diversity version of VMDAV was proposed for statistical disclosure control.

3. System model and motivating application

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 a location sensitive participatory sensing application.

3.1. System model and assumptions

3.1.1. System model

We leverage the AnonySense architecture proposed in [15] to provide participatory sensing infrastructure support, but take a different approach to address the issue of potential disclosure of private location information. In particular, we focus on the privacy protection aspects of the architecture. Fig. 1 gives a pictorial description of the system architecture.

To assist a participatory sensing application, the architecture depends on four core services: (1) a collection of mobile nodes (MNs), (2) a registration authority (RA), (3) a task server (TS), and (4) a report server (RS). Further, it assumes the existence of a Mix Network (MIX), which provides a medium for anonymous communications. MNs are devices with sensing and communication capabilities and are mostly carried by humans (in some cases, they are attached to objects such as vehicles). It should be noted that the participation of MNs in sensing is voluntary. The RA is the central hub for trust establishment. It verifies the integrity of other service components and issues certificates and keys so that they can anonymously authenticate each other. The TS is responsible for the downward communication between the application and MNs. It ensures the tasks from the application are genuine and do

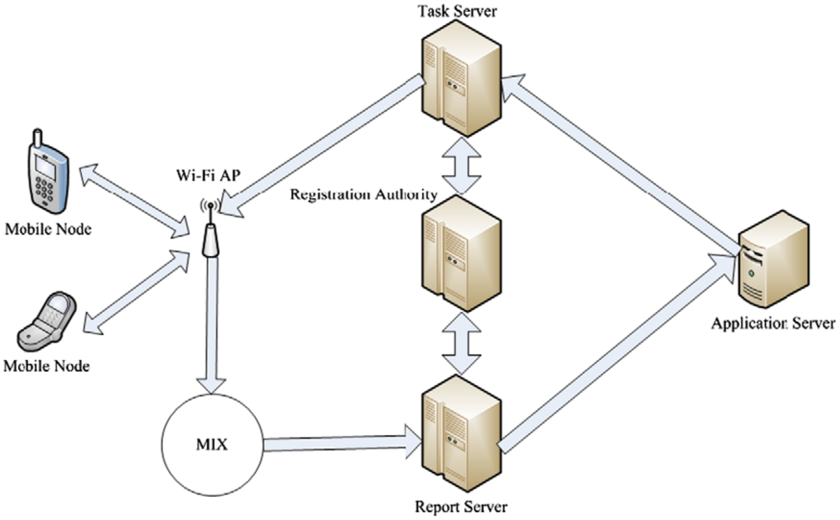

Fig. 1. AnonySense architecture.

not violate MN carriers' privacy requirements. Its counterpart, RS, aggregates reports from MNs to provide additional level of protection and channels the reports to the trusted application. MIX is used by MNs to de-link their reports before they reach the RS. The anonymous property of messages traversing inside MIX prevents the message recipient from linking multiple reports to the same origin.

In their related work [16], the authors proposed tessellation for providing privacy in the architecture. The operation of tessellation requires the existence of an additional map server (MS), which is responsible for the generation of tessellation map (i.e. dividing a geographic region into tiles). MNs query the MS for the tessellation map, which allows them to determine the appropriate tiles that should be reported with 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 4) 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/Wi–Fi). The user consults the AS prior to submitting the reports. The AS runs the appropriate microaggregation algorithm and provides the user with anonymized locations, which are used to annotate the reports. The application then processes and interprets the received data using the anonymized locations.

3.1.2. Trust assumption

We make the following assumptions regarding the trust laid upon system components: (1) the AS is independently owned by a third-party operator 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 users to trust a single system entity with their accurate information. Hence, we propose a scheme to relax this assumption in Section 4.4.

3.1.3. Threat model

We focus on the temporal and spatial information included in user reports. The threat model presumes the existence of a hostile adversary, who does not know the true values of time and location corresponding to the user reports. However, the adversary is assumed to have means to find out the temporal and spatial properties of his victim, e.g. time of day or the suburb in a city. In this paper, we assume such an adversary possessing some degree of temporal knowledge about individuals. For example, he may know the time period over which certain individuals are more likely to use PetrolWatch (an example participatory sensing application used in the rest of the paper and detailed in the next sub-section), e.g. on their way back home from work. The goal of the adversary, with this prior temporal knowledge at his disposal, is to either identify his victim precisely (identity disclosure) or to deduce the nature of the places that his victim has visited (attribute disclosure). We also assume that the adversary is able to observe submitted reports, which consist of the ⟨User ID, Anonymized Time, Anonymized Location⟨ columns of Table 1. This is possible via eavesdropping or being a malicious application administrator.

3.2. 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 [11], we have designed a novel application, *PetrolWatch*, which allows users to automatically collect, contribute and share fuel pricing information using camera phones. Users mount their camera-enabled mobile phones on the car dashboard. Through the use of GPS and GIS, PetrolWatch knows when the vehicle is approaching a service station and triggers the camera automatically. Pictures of fuel pricing billboard are processed by computer vision algorithms to extract fuel prices. Fuel prices are annotated with location coordinates of the service station and time at which the capture takes place, and uploaded to the application server. Users can query the server to locate the cheapest petrol station in their vicinity.

Fig. 2 is the pictorial representation of Table 1 and illustrates a simple distribution of users for PetrolWatch, assuming that tessellation is employed to provide location privacy. There are six users spread across a region of size 9 km × 7 km (for simplicity we assume a 2D coordinate system). Fig. 2 captures the locations of users at a particular time instant. Assume that there is a service station co-located with the current location of each user (i.e. six service stations in total) and that, a user only records pricing information of the co-located service station. Now suppose that user 2 is in the process of uploading fuel pricing information to the application server. A query is first sent from the user to the AS requesting for an anonymized location that should be reported. Given the distribution of users, the AS constructs two tiles as shown in Fig. 2 (fol-

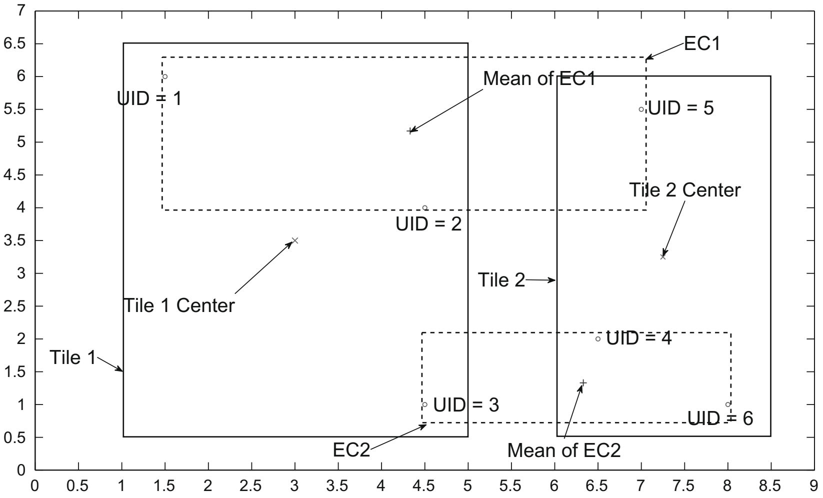

Fig. 2. User distribution in the example application: PetrolWatch.

lowing the guidelines of tessellation in [16]) assuming the privacy requirement is $k = 3$, and advises the user of his anonymized location, i.e. tile 1. Consequently, user 2 annotates his report with tile 1 instead of his actual location (4.5, 4). When the report is submitted to the application, it needs to associate the received report with one of the three service stations located in tile 1. However, without additional information, it is unable to confidently determine that the fuel prices included in the report correspond to the service station co-located with user 2. This simple example clearly illustrates the intrinsic limitation of using tessellation as a means for location privacy, and serves as the primary motivation for our proposed schemes in the next Section.

It should be noted that Fig. 2 shows only one possible arrangement for user clustering. It is likely that other viable alternatives exist, which may potentially have a different impact on the performance of the application. A set of general instructions for tile construction is given in [16], but it provides no discussions on the impact of varying tile configurations.

4. *k*-anonymous privacy-preserving schemes

In light of the aforementioned limitation, we propose a simple modification to tessellation and demonstrate how our scheme solves the problem posed by PetrolWatch. This result is presented in Section 4.1. In Section 4.2, we investigate how perturbation-based *k*-anonymous techniques can be applied as an alternative approach. In particular, we propose the use of VMDAV for addressing privacy issues in participatory sensing. Section 4.3 describes some important observations from the previous two sub-sections and introduces Hybrid-VMDAV, which is another alternative scheme attempting to deliver the benefits of both generalization (tessellation) and perturbation (VMDAV). Section 4.4 incorporates a simple input perturbation mechanism with our proposed privacy-preserving schemes. It represents our attempt to free users from having their precise location information known to the AS.

Note that in theory, VMDAV can be applied to any numerical attributes, e.g. location and time. However, since tessellation was designed primarily to address location privacy in participatory sensing, we limit the use of VMDAV on the location attribute as well to facilitate a fair comparison. In other words, the timing information associated with reports is processed by generalization as described in Section 2.2. We defer the usage of VMDAV for both temporal and spatial privacy to Section 5.

4.1. Tessellation with tile center reporting

The example in Section 3.2 shows that the problem with tessellation in providing location privacy is that it uses a region, rather than a point coordinates, for location anonymization. In this regard, a natural modification to it is to represent each tile by the coordinates of its center. Hence, we propose a modification, wherein, a user's reports are annotated with location coordinates of the center of the tile in which he is currently observed. This requires a simple update to the AS, such that it includes the coordinates of tile centers in the tessellation map. We illustrate the operation of this scheme by using the same example as depicted in Fig. 2. With the above change in place, users 1, 2, and 3 anonymize their positions using (3, 3.5), which is the center of tile 1. Similarly, users 4, 5, and 6 mask their locations with the center of tile 2, (7.25, 3.25). This alteration provides the application with more options to analyze the data contained in user reports. For example, searching for the shortest Euclidean distances between the anonymized location reported by user 2 and the positions of the six candidate service stations reveals that, user 2 was most likely referring to the one in his vicinity.

We acknowledge here that the method of shortest Euclidean distance may not be the best strategy for an application to analyze received positional data. Nonetheless, it adequately demonstrates one of the advantages of this numerical value driven approach. In the rest of this paper, we refer to the above alternate tessellation scheme as TwTCR.

4.2. Location anonymization with microaggregation

Even though TwTCR overcomes the obstacles encountered in Section 3.2, it should be noted that depending on user density, some tiles may have considerably large areas. In such cases, reporting the center of tiles may lead to data infidelity and cause the application to erroneously interpret the locations contained in reports (this point is further elaborated in the evaluations in Section 6). We propose the use of microaggregation as an alternative to achieve location privacy in these situations. In particular, we adopt the VMDA heuristic proposed in [24]. The pseudo code of VMDAV is reproduced in Fig. 3 for reference.

We illustrate the outcome of this heuristic using the example from Table 1 which is depicted in Fig. 2. The AS generates two equivalence classes: one encompasses users 1, 2, and 5 and the other includes users 3, 4, and 6. In this approach, user 2 substitutes his position with the mean location coordinates of the equivalence class to which he belongs, i.e. (4.33, 5.17). This anonymization not only meets 3-anonymity (the size of each equivalence class is 3) but also ensures that a numeric location value is provided to the application.

4.3. Location anonymization with hybrid microaggregation

So far, we have introduced two k-anonymous privacy-preserving schemes, TwTCR and VMDAV. An immediate question at this point would be if there was any reason to favor one over the other? To this end, we present two simple examples to demonstrate that both TwTCR and VMDAV have their advantages in certain mutually exclusive situations. These observations motivate us to propose a novel technique that combines the best of both methods.

Let us first consider the example in Fig. 2. Assume that user 6 is in the process of uploading his fuel pricing report to the application server. We assume that the server has some background knowledge regarding this report, e.g. it knows that this report would not have referred to the service 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 service station. The location data carried by the report can be either (7.25, 3.25) if TwTCR was employed or (6.33, 1.33) if VMDAV was used. Assume that the application server compares the Euclidean distances of all six service stations to the location contained in the report, and concludes that the report corresponds to the service station closest to the reported location. In the case of TwTCR, the server mistakenly makes the decision that this report referred to the service station co-located with user 5. On the other hand, with V-MDAV, a correct association can be made with the chosen service station being located in the vicinity of user 6.

Let us now consider a different example with a different user distribution as depicted in Fig. 4. Let us first focus on TwTCR. Observe that the cell in which users 2, 3, and 4 are located satisfies the privacy requirement $k = 3$ on its own and hence, this cell forms a tile. On the other hand, the cells in which the remaining users are found need to be merged together according to the rules of tessellation. VMDAV, on the other hand, creates two equivalence classes. Users 1, 2, and 3 constitute one equivalence class, while the remaining users are grouped into the other one. Now, assume that user 4 is to submit his report. He will anonymize his location using either TwTCR-generated (4, 1.75) or VMDAV-produced (7.17, 5.5). Note that, these anonymized values are generated by the AS. Using Euclidean distances for interpretation as in the previous example, the application server can correctly associate the report submitted by user 4 with his co-located service station if TwTCR was chosen. On the contrary, VMDAV would have led to an incorrect association with the deduced service station being the one near user 5.

The following observations can be made based on the above examples: (1) VMDAV enables an application to make better decisions when the user distributions across different areas are relatively consistent, as in Fig. 2. (2) On the contrary, in areas with dense distribution of users, as in Fig. 4, TwTCR performs better. Given that the two schemes have their advantages in contrasting sit-

```
VMDAV Algorithm
(01) Function input: user data D, privacy requirement $k$, output: set of equivalence classes M
(02) Compute matrix of distance(D)
(03) C = Center of the dataset(D)
(04) while (There are more than $k - 1$ unassigned users), do
(05) e = pick the with the largest distance from the center of the dataset (D,C)
(06) ec($i$) = build an equivalence class around e(e, D, $k$)
(07) ec($i$) = extend the above equivalence class if necessary (ec($i$,D,$k$)
(08) end while
(09) ec(1)...ec(s) = Assign remaining unassigned users(D, ec(1)...ec(s))
s is the number of equivalence classes resulted from steps 4 and 8
(10) M = Aggregate attribute values, i.e. computing mean attribute values.
(11) return M
(12) end Function
```

Fig. 3. Pseudocode for VMDAV.

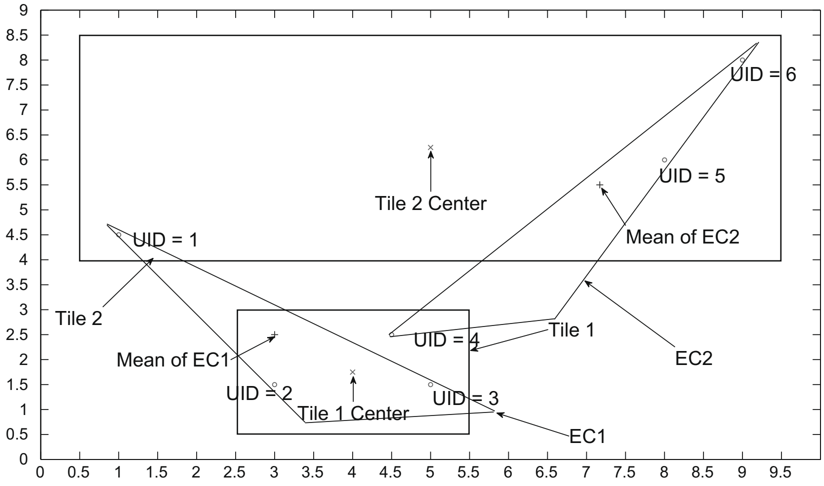

Fig. 4. An example demonstrating the benefit of Hybrid-VMDAV.

uations, we propose Hybrid-VMDAV, which attempts to combine the best of both methods. The hybrid scheme adaptively makes a decision on whether to operate in TwTCR or VMDAV mode. The operation of Hybrid-VMDAV is quite simple. If a user is in a tile-forming cell, i.e. the number of users within the cell exceeds k, then TwTCR is used. Otherwise, the algorithm switches to V-MDAV mode. If Hybrid-VMDAV was applied to the example in Fig. 4, users 2, 3, and 4 would anonymize their locations using the value suggested by TwTCR, whereas the other users would use VMDAV-generated value. This overcomes the erroneous association explained earlier.

4.4. Gaussian input perturbation

All schemes discussed thus far assumed the existence of a trusted third-party server, which is aware of the true locations of participating users (recall that a user queries the AS and provides his current details, each time he needs to upload a report). Clearly, this architecture is not resilient against a single point of failure since, if this server was compromised, then user privacy is at risk. Further, users may not be comfortable with the idea of a system entity keeping track of their locations. In fact, this may be a turn off factor 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 cell of the Voronoi diagram for the area of interest (we will explain the construction and property of Voronoi diagram in Section 6). Based on this information, we can place 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 a user's coordinates, a factor p is introduced as a scaling factor so that the perturbed value does not deviate significantly from his true location. p usually takes on a small fractional value (see evaluations in Section 6).

Note that, the proposed perturbation scheme is the simplest of its kind. It is introduced for the purpose of investigating the viability of user-side pre-processing in the face of a distrustful AS. It has been shown by several authors [18,19] that, merely adding random noise to data does not protect privacy. They argue that correlations among different pieces of data or between data contributors can be exploited to reconstruct the data, unless the noise is too large to the extent where data utility is completely removed. For example, if a user perturbs his location with different Gaussian random numbers every time he updates the AS, it is possible to track his locations progressively more accurately by averaging past updates to cancel out noise. In [20], the authors develops the mathematical foundations and architectural components to perturb user data such that, the reconstruction of data from noisy versions is avoided while still allowing the computation of aggregate information. It is possible to incorporate the techniques proposed in [20] with our privacy-preserving techniques for better privacy against the AS. We do not pursue this topic in this paper but leave it as a potential future work.

5. From *k*-anonymity to *l*-diversity

The privacy protection schemes proposed in Section 4 are based on k-anonymity. In short, users location privacy are preserved by ensuring that they anonymize their positions with different coordinates, and those coordinates are shared among a group of users. It was mentioned in Section 2 that, the level of privacy enabled by k-anonymity is insufficient to defend against attribute disclosure. In what follows, we examine if the k-anonymous schemes proposed in Section 4 lead to attribute disclosure.

Recall that in Section 4, we have assumed that the temporal information in user reports is generalized by simple techniques such as increasing the time granularity. Consequently, the discussions in Section 4 omitted temporal privacy and focused exclusively on spatial privacy. In this Section, we consider temporal and spatial attributes simultaneously since, we expect most practical participatory sensing applications would involve both of these in user reports. Correspondingly, users would be concerned about preserving the privacy of both these attributes. Further, in line with the threat model as described in Section 3.1.3, we assume one of the attributes is designated as primary, and the other as secondary depending on users perceived importance. We leverage the capability of VMDAV in operating on any numerical values and anonymize user times (along with locations) using value perturbations.

This Section is organized as follows: Section 5.1 describes an example showing that the k-anonymous schemes discussed in Section 4 are also prone to attribute disclosure. We investigate if l-diversity can be applied to solve this shortcoming in Section 5.2. Section 5.3 details our VMDAV-inspired l-diversity algorithm, LD-VMDAV, and points out a few of its attractive features.

5.1. Scenario and attribute disclosure

Consider a 3-anonymous set of PetrolWatch reports shown in Table 2. Columns 2 and 3 are the actual times and locations of users, respectively, and are not included in user reports. Columns 4 and 5 contain the values of location and time, respectively, that are generated by VMDAV at the AS and returned to users for protecting their temporal and spatial privacy. Note that, since location and time attributes are considered simultaneously in this example, a minor change must be made to the VMDAV algorithm presented in Section 4.2. In particular, the distance metric used in step (02) and (05) of Fig. 3 must now refer to the combined spatial and temporal distances (cf. spatial distance used in Section 4). In other words, members of equivalence classes in Table 2

Table 2
Another example of 3-anonymous reports maintained at the application.

User ID	Time	Location	Anonymized time	Anonymized location
11	12:01:31	(0.77, 1.00)	14:07:22	(0.72, 0.78)
7	16:14:21	(1.00, 0.73)	14:07:22	(0.72, 0.78)
6	14:06:15	(0.37, 0.63)	14:07:22	(0.72, 0.78)
8	17:18:32	(0.65, 0.00)	17:18:03	(0.52, 0.27)
1	17:07:30	(0.82, 0.42)	17:18:03	(0.52, 0.27)
12	17:28:07	(0.10, 0.40)	17:18:03	(0.52, 0.27)
2	16:27:57	(0.00, 0.28)	16:55:49	(0.15, 0.55)
5	17:07:55	(0.23, 0.57)	16:55:49	(0.15, 0.55)
10	17:11:34	(0.21, 0.81)	16:55:49	(0.15, 0.55)
9	18:08:45	(0.32, 0.91)	15:33:03	(0.33, 0.73)
3	15:04:08	(0.22, 0.78)	15:33:03	(0.33, 0.73)
4	13:26:15	(0.44, 0.49)	15:33:03	(0.33, 0.73)

are similar in terms of location and time (cf. location only in Table 1).

We now illustrate how elements of homogeneity and background knowledge attacks described in Section 2.4 can be used to cause attribute disclosure. Suppose that user 8 in Table 2 has a fuel pricing report ready for the application server. He uses the suggested attribute values, i.e. (17:18:03), (0.52, 0.27), from the AS to maintain his privacy in the report. Based on the threat model postulated in Section 3.1.3, an adversary is able to deduce that, user 8 must be the owner of one of the reports in the second equivalence class. Although this observation is insufficient for an exact identity match (since there are two other users sharing the same position), it nonetheless reveals the victim's whereabouts on a coarser scale. Depending on the context or user preferences, this coarse locational representation may be undesirable. For example, [0.52, 0.27] can be mapped to a region of the child care center in which the victim's child is placed. The inadvertent release of this information is an example of attribute disclosure, and can at times be considered too intrusive. Note that, the above disclosure is eventuated by (1) the adversary's background temporal knowledge and (2) the homogeneity of the anonymized locations in the second equivalence class.

5.2. Employing l-diversity in participatory sensing

It was mentioned in Section 2.5 that attribute disclosure could be effectively avoided if diversity was introduced. In what follows, we give a conceptual overview of the improvements that can be achieved by *l*-diversity. We use the previous example in Table 2 to illustrate this. For simplicity, distinctive 2-diversity is used.

Given the example in Table 2, the most intuitive approach to produce a set of reports satisfying distinctive 2-diversity is to combine two equivalence classes into a group, with group members sharing a common time. For example, the odd-numbered equivalence classes can be merged to form the first group, wherein members replace their actual time with their mean time value, e.g. (15:31:36). Group 2 can be formed by merging the even numbered equivalence classes in a similar manner. Table 3 shows the result of the above operation. Notice that for reports in Table 3, users in a group receive the same temporal anonymization but are given two different anonymized locations. In other words, the set of reports represented by Table 3 is 2-diverse in terms of the location attribute. For the threat model considered (see Section 3.1.3), the information carried by the reports in Table 3 reduces the probability of attribute disclosure. For example, assuming that using prior temporal information the adversary can deduce that the victim's report belongs to group 1. The adversary is now presented with two possible locations for the victim. If these locations are sufficiently apart, then the adversary has difficulty in narrowing down on the victim's precise location. The probability of attribute disclosure is reduced by 50% in this example (cf. 100% for reports in Table 1).

Table 3
An example of 2-diverse reports maintained at the application.

Group ID	Class ID	User ID	Time	Location	Anonymized time	Anonymized location
1	1	11	12:01:31	(0.77, 1.00)	15:31:36	(0.72, 0.78)
1	1	7	16:14:21	(1.00, 0.73)	15:31:36	(0.72, 0.78)
1	1	6	14:06:15	(0.37, 0.63)	15:31:36	(0.72, 0.78)
1	3	2	16:27:57	(0.00, 0.28)	15:31:36	(0.15, 0.55)
1	3	5	17:07:55	(0.23, 0.57)	15:31:36	(0.15, 0.55)
1	3	10	17:11:34	(0.21, 0.81)	15:31:36	(0.15, 0.55)
2	2	8	17:18:32	(0.65, 0.00)	16:25:33	(0.52, 0.27)
2	2	1	17:07:30	(0.82, 0.42)	16:25:33	(0.52, 0.27)
2	2	12	17:28:07	(0.10, 0.40)	16:25:33	(0.52, 0.27)
2	4	9	18:08:45	(0.32, 0.91)	16:25:33	(0.33, 0.73)
2	4	3	15:04:08	(0.22, 0.78)	16:25:33	(0.33, 0.73)
2	4	4	13:26:15	(0.44, 0.49)	16:25:33	(0.33, 0.73)

Although the above approach works, its application needs some careful thought. Consider an alternative grouping of the records in Table 2, wherein group 1 consists of the first two equivalence classes while the second group is made of the remaining ones. The resulting groups of reports are different from those in Table 3 but they are still 2-diverse. However, the anonymized locations of these two equivalence classes are very close. For example, the anonymized location coordinates for the reports in this group, i.e. (0.72, 0.78) and (0.52, 0.27), may be mapped to two different areas of the same complex, e.g. different facilities in a hospital. This can neutralize the application of *l*-diversity if the user considers disclosure of his presence in the hospital to be a violation of privacy. There is another problem which may not be immediately obvious from the example in Table 3, but can have a negative implication on the performance of applications. Consider the first group of reports in Table 3. Observe that a single temporal value (15:31:36) is used. This may be too coarse-grained for the information to be meaningful to some applications.

5.3. Implementation of l-diversity for Participatory Sensing

The discussions in Section 5.2 identified two issues which need to be addressed in any *l*-diversity implementations: (1) the semantic relationship between locations and (2) timing accuracy. In the following, we propose an *l*-diverse extension of VMDAV (detailed in Section 4.2) called LD-VMDAV. As in VMDAV, this algorithm is also executed at the AS (see Section 3.1.1 for detailed overview of the system). We show that the LD-VMDAV creates a set of *l*-diverse reports with significantly reduced spatial correlations and timing errors. The implementation of LD-VMDAV is based on successive applications of the VMDAV algorithm. The first pass of VMDAV anonymizes the primary attribute while the second pass produces anonymized values for secondary attribute. Application designers can designate the attribute that is the most important (from the perspective of user privacy) as primary and less important one as secondary. In the rest of this discussion, we assume that location is the primary attribute and time is the secondary attribute. Specifically, LD-VMDAV involves the following two steps:

1. VMDAV is first executed over the entire dataset but only with respect to the primary attribute. The parameter k is set to the required k-anonymity level, i.e. $k = 3$ to be consistent with example in Table 3.
2. VMDAV is executed again over the entire dataset but this time only with respect to the secondary attribute. The parameter k' is set to the product of the required k-anonymity level and the required l-diversity level, i.e. $k' = k \times l = 3 \times 2 = 6$.

Table 4 shows an example of the output of the first step, where $*$ denotes values yet to be determined. It is clear that each anonymized location is shared among at least $k = 3$ users. The next step

Table 4
Results of first step of LD-VMDAV.

Class ID	User ID	Time	Location	Anonymized time	Anonymized location
1	8	17:18:32	(0.65, 0.00)	** : ** : **	(0.64, 0.30)
1	1	17:07:30	(0.82, 0.42)	** : ** : **	(0.64, 0.30)
1	4	13:26:15	(0.44, 0.49)	** : ** : **	(0.64, 0.30)
2	7	16:14:21	(1.00, 0.73)	** : ** : **	(0.72, 0.78)
2	11	12:01:31	(0.77, 1.00)	** : ** : **	(0.72, 0.78)
2	6	14:06:15	(0.37, 0.63)	** : ** : **	(0.72, 0.78)
3	2	16:27:57	(0.00, 0.28)	** : ** : **	(0.11, 0.41)
3	12	17:28:07	(0.10, 0.40)	** : ** : **	(0.11, 0.41)
3	5	17:07:55	(0.23, 0.57)	** : ** : **	(0.11, 0.41)
4	9	18:08:45	(0.32, 0.91)	** : ** : **	(0.25, 0.82)
4	10	17:11:34	(0.21, 0.81)	** : ** : **	(0.25, 0.82)
4	3	15:04:08	(0.22, 0.78)	** : ** : **	(0.25, 0.82)

of the algorithm determines the unknown anonymized times and produces the final output[7] as illustrated in Table 5. Observe that the reports in Table 5 are at least 2-diverse, with group 1 exhibiting 4-diversity and group 2 exhibiting 3-diversity (both with respect to location).

The independent executions of VMDAV with respect to each attribute in LD-VMDAV are important because it creates a number of remarkable features as observed from reports in Table 5. First, contrary to the intuitive approach in Section 5.2, more than two equivalence classes constitute a group. Second, even though the algorithm was designed to meet the user-specified diversity level (2-diverse in this example), the resulting groups of reports always exhibit more than the required diversity (diversity level of 4 and 3 in the example in Table 5). Our experiments with the real-world trace data (in Section 6) suggests that this is a generic property of the algorithm. This is a highly desirable property from the perspective of location privacy since, the more diverse the location values, the harder it is for an adversary to deduce the true location. Third, notice that the anonymized locations in a group demonstrate a reasonable amount of separation, which means the problem of multiple references to a common sensitive place as described in Section 5.2 can be avoided. Lastly, the anonymized times used by reports in Table 5 cause smaller inaccuracy. For example, the second group in Table 5 uses the anonymized time of (17:23:44) to represent the actual times, which are all in the interval from 17:00 to 18:00. The improvement in data accuracy will be formally quantified in Section 6, where the *information loss* and *positive identification percentage* metrics are defined.

The example in Table 5 shows that LD-VMDAV is able to simultaneously account for spatial and temporal privacy, albeit different protection mechanisms are involved. Spatial privacy for owners of reports in Table 5 is provided via *l*-diversity but, *k*-anonymity is enforced to guard their temporal privacy. Note that, by interchanging the primary and secondary attributes in the algorithm, it is easy for LD-VMDAV to swap the protection mechanisms, i.e. *l*-diversity for temporal privacy and *k*-anonymity for location. In this case, the resulting anonymized reports are resilient against temporal attribute disclosure. More specifically, an adversary with *spatial* prior knowledge is unable to deduce the time at which his victim has visited a particular location.

6. Evaluations

We present results from a simulation study that compares the performance of the proposed privacy-preserving schemes, TwTCR, VMDAV, Hybrid-VMDAV, and LD-VMDAV. Our evaluation focuses on their costs, in particular, the errors induced by anonymization and the accuracy of application decisions are of interests. Section 6.1 describes the simulation setup and the evaluation methodology. Section 6.2 introduces the two metrics used to assess the algorithm performance. Simulation results are provided in Section 6.3.

6.1. Overview of simulation setup

6.1.1. Simulation scenario

In the following evaluations, we consider a scenario wherein, a participatory sensing application similar to PetrolWatch (described in Section 3.2) has been deployed. We assume that the application server generates tasks that require users to collect certain contextual information from some points of interest in their immediate vicinities. Users who agree to participate in the application accept the tasks, collect sensor data, annotate sensor reports with time and location, and upload the reports to the server via the architecture described in Section 3.1.1. Prior to generating sensor reports, a user contacts the AS with his desired privacy parameters (k, l).[8] The AS in response provides the user with his anonymized time and location whose values depend on the privacy algorithm executed, e.g. TwTCR, VMDAV, Hybrid-VMDAV, or LD-VMDAV.[9] The application is aware of the location coordinates of all points of interest. When the server receives sensor reports, it applies the method of shortest Euclidean distance (discussed in Section 4.1) to determine the ⟨point of interest, report⟩ associations.

6.1.2. Data

Our evaluations are based on real-world trace data. In particular, the Dartmouth College campus traces, which are made publicly available on CRAWDAD [26], are used. These traces contain log entries collected from Wi–Fi APs deployed around the Dartmouth College campus. We choose the "*syslog/05_06*" trace[10] under "syslog" traceset and "*aplocations*" trace under "movement" traceset to deduce user distributions and to overlay a Voronoi diagram over the campus map. Each record in the *syslog/05_06*" trace logged the association, re-association or disassociation of a user's Wi–Fi enabled device with an AP. The "*aplocations*" trace contains a list of APs deployed across the college campus and provides information about their (x, y) coordinates as well as the floors on which they are located.

6.1.3. Methodology

The use of TwTCR requires the region of interest, i.e. the college campus, to be tessellated. In what follows, we describe in detail

[7] To maintain a consistent use of terminology, we use *Class ID* to label the equivalence classes generated by step 1 and *Group ID* for those produced by step 2 of the LD-VMDAV.

[8] If $l = 0$, it means the user does not opt for *l*-diversity level of protection.

[9] For TwTCR and Hybrid-VMDAV, only value generalization is available to produce anonymized time values. On the other hand, value generalization as well as value perturbation can be used for temporal anonymizations if VMDAV or LD-VMDAV is used. The default mode of operation for VMDAV and LD-VMDAV is value perturbation.

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

Table 5
An example of reports generated by LD-VMDAV.

Group ID	Class ID	User ID	Time	Location	Anonymized time	Anonymized location
1	3	2	16:27:57	(0.00, 0.28)	14:33:24	(0.11, 0.41)
1	4	3	15:04:08	(0.22, 0.78)	14:33:24	(0.25, 0.83)
1	1	4	13:26:15	(0.44, 0.49)	14:33:24	(0.64, 0.30)
1	2	11	12:01:31	(0.77, 1.00)	14:33:24	(0.72, 0.78)
1	2	6	14:06:15	(0.37, 0.63)	14:33:24	(0.72, 0.78)
1	2	7	16:14:21	(1.00, 0.73)	14:33:24	(0.72, 0.78)
2	3	12	17:28:07	(0.10, 0.40)	17:23:44	(0.11, 0.41)
2	3	5	17:07:55	(0.23, 0.57)	17:23:44	(0.11, 0.41)
2	4	9	18:08:45	(0.32, 0.91)	17:23:44	(0.25, 0.83)
2	4	10	17:11:34	(0.21, 0.81)	17:23:44	(0.25, 0.83)
2	1	8	17:18:32	(0.65, 0.00)	17:23:44	(0.64, 0.30)
2	1	1	17:07:30	(0.82, 0.42)	17:23:44	(0.64, 0.30)

how this process is accomplished. There are 623 APs listed in the "*aplocations*" trace. In order to simplify the analysis, we perform planarization and condensation similar to [16]. In the planarization step, the floor numbers of APs are ignored and all APs are assumed to be located on floor 0. Furthermore, APs located in the same building are grouped together and collectively represented by their mean (x, y) coordinates, this completes the condensation step. The result of the above simplification is shown pictorially in Fig. 5. Fig. 5 contains 124 APs and has a set of Voronoi cells overlaid. A Voronoi cell has the following property: all points within its interior are closer to its generating point than to any others, e.g. in our context, the generating points are the positions of APs. This property allows us to define the boundary of a region in which users of an AP can be observed. We also normalize the locations of APs so that they are confined to a region of unit square area.

To estimate user distribution per cell, we consider traces between 12 p.m. and 6 p.m. over a week period from the 1st of September, 2005 to the 7th of September, 2005. The number of user associations per cell is a threshold value, which represents the number of users that can be statistically expected in a cell (connected to the cell AP) for 95% of the specified time intervals. In our evaluations, this interval is 30 min. There are 153 users whose distributions are marked by asterisks in Fig. 5. The coordinates of users in a cell are randomly generated, once the threshold value for that cell is known. Neighboring cells are grouped to form tiles such that k-anonymity is attained. We use $k = 10$ in all our simulations, unless otherwise stated. The tiles are shown as colored regions in Fig. 5.

6.2. Metrics

6.2.1. Application accuracy

Collecting sensor readings from participating users is only part of the objectives of participatory sensing. Eventually, an application receiving sensor reports must be able to use the embedded information as much as possible. In other words, an application should ideally make ⟨point of interest, report⟩ association decisions with high accuracy. To this end, we define a metric called *Positive Identification Percentage* (PIP) to measure the precision of applications. Specifically,

$$PIP = \frac{\text{total number of correct (positive) associations by the application}}{\text{total number of reports received by the application}}, \tag{1}$$

Note that, depending on the attribute of interest, there are different ways to define positive association in the numerator of (1). In terms of location, a positive association refers to the case in which an application correctly identifies the intended points of interest from users anonymized locations. If we assume that an application performs no further processing, e.g. shortest Euclidean distance calculations, on time but logs its values as specified in sensor reports. Then, a positive association of time means that the difference between the time tuple ⟨actual time, anonymized time⟩ does not exceed the tolerance level specified by the application. For example, if an application can operate with a tolerance level of 30 min, then the time tuple ⟨17:30, 17:10⟩ flags a positive association in the temporal domain.

6.2.2. Errors induced by anonymization

All the evaluated schemes anonymize users true attribute values. It is therefore of interest to see how much information is lost as a result of this process. We adopt the commonly used Informa-

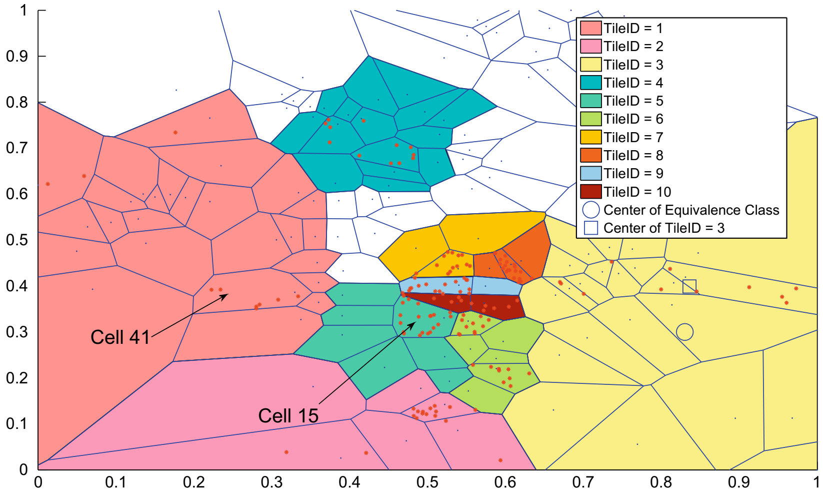

Fig. 5. Tessellation map of the simulation scenario.

tion Loss (IL) metric [21] for this purpose. Information Loss is defined as

$$IL = \frac{SSE}{SST}, \tag{2}$$

where

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

and

$$SST = \sum_{j=1}^{g} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{4}$$

where x_i denotes the ith record in group j, and each of the g groups containing n records. $\bar{x}_j$ and $\bar{x}$ represent the group mean and the mean of the entire dataset, respectively. SSE is the sum of squared errors with respect to class (group) mean while SST is the same quantity but with respect to the mean of dataset. Note that, SSE measures the distances between the actual attribute values and their anonymized versions.

6.3. Simulation results

We conduct a set of simulations to evaluate the PIP and IL achieved by TwTCR, VMDAV, Hybrid-VMDAV, and LD-VMDAV. Similar to the PetrolWatch application described in Section 3.2, we assume that the points of interest are co-located with users. In order to realistically reflect the real-world usage of applications, we investigate the impact of varying the proportion of users submitting reports. In particular, we alter the percentage of users reporting data (active users) from 20% to 100% in 20% increments. All simulations are repeated for 1000 times and average values are recorded.

The presentation of simulation results is organized as follows: the first subset compares the performance of privacy-preserving schemes based on k-anonymity. In particular, TwTCR, VMDAV, and Hybrid-VMDAV are assessed. The second subset is devoted to the simulation results for our l-diversity implementation. Specifically, the performance of LD-VMDAV is thoroughly evaluated against its k-anonymity counterpart, VMDAV. In the last part we present results that evaluate the impact of additional input perturbation (as described in Section 4.4) on the performance of VMDAV.

6.3.1. Comparison of TwTC, VMDAV, and Hybrid-VMDAV

Recall that, the main motivation behind the design of TwTCR, VMDAV, and Hybrid-MDAV is to provide alternatives to tessellation for location privacy (see Section 4). Thus this set of simulations are mainly instrumented to compare the relative strengths of these techniques in relation to location privacy. In other words, the calculations of PIP and IL are only with respect to location. We also assume that, there is no difference in the anonymized times produced by the three techniques, i.e. they all apply value generalization on the temporal attribute.

Fig. 6 shows the PIP and IL for k-anonymous schemes over a range of active users. One can readily observe that the performance of all three algorithms do not vary significantly with an increase in the number of users contributing data. In other words, the accuracy enabled by the proposed schemes is not affected by an increasing system load. Hybrid-VMDAV achieves a 40% reduction in IL as compared to TwTCR. The performance of the hybrid scheme is marginally better than that of VMDAV. We explain the inferior performance of TwTCR by using tile 3 in Fig. 5 as an illustrative example. Observe that the center of tile 3 denoted by a circle is quite distant from the actual locations of users. Recall that, in TwTCR, users report the center of tile as their locations. On the contrary, with VMDAV and Hybrid-VMDAV the same set of users would report a much closer square-denoted coordinates in place of their actual locations. As a result, the SSE is larger with TwTCR for users within tile 3 (recall that SSE measures inter-class distances) as compared to the other two alternates. Consequently, TwTCR produces higher IL. One might argue that the performance gap can 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. 5 are constructed to fit all user distributions, which also account for the subsequent Gaussian perturbation extension and (2) to the best of our knowledge, there is no real-time algorithm that can produce optimal tessellation maps, which can adapt to constantly fluctuating user distributions.

Fig. 6 also suggests that there exists an inverse relationship between PIP and IL. For instance, TwTCR, which has the highest IL results in the lowest PIP. Similarly, Hybrid-VMDAV, which achieves the highest PIP, has the lowest IL. Observe that, Hybrid-VMDAV improves the positive identifications made by the server by more than 100%, in comparison with TwTCR. The significant improvement achieved by Hybrid-VMDAV over VMDAV can be explained by considering cell 15 in Fig. 5, which accommodates 20 users. According to the rules of Hybrid-VMDAV, these 20 users replace their locations with the center of cell 15. Since these users are all located near the cell center, the application server can interpret the true locations with high accuracy. On the other hand, VMDAV separates these users by grouping some of them with those in cell 41 in an attempt to lower IL while keeping the size of equivalence class in check, i.e. between 10 and 19.[11] The result is a reported location somewhere in between cells, which is not close to the users and the point of interest to which their reports refer. Hence, the application server tends to make wrong associations resulting in a lower positive identification rate. It should be noted that, even the best performing Hybrid-VMDAV only allows an application to achieve a moderate level of accuracy. This is because the simplistic Euclidean estimation technique is employed for making the $\langle$point of interest, report$\rangle$ associations. We intend to investigate alternate techniques in our future work.

6.3.2. VMDAV and LD-VMDAV

This part of the simulation compares the performance of k-anonymous and l-diverse versions of VMDAV. Contrary to the location-only analysis in the previous sub-section, both temporal and spatial privacy are considered here. To this end, random times are generated for the 153 users in Fig. 5 in addition to their existing random locations. Further, to establish a consistency between time and location attributes, VMDAV uses value perturbation (see Section 5.1) to generate anonymized times (cf. value generalization in previous sub-section) as well as locations.

Fig. 7 shows the IL and PIP produced by VMDAV and LD-VMDAV for an anonymity level of 10 and a diversity level of 2, i.e. $(k, l) = (10, 2)$, and over a range of active users. Note that, with $k = 10$ and $l = 2$, LD-VMDAV creates groups of equivalence classes with 20 users (see Section 5.3), therefore, an anonymity level $k' = 20$ is required for VMDAV to facilitate a fair comparison. Note that, we use the prime notation to differentiate the anonymity levels input to VMDAV and LD-VMDAV. Recall that, two parameters are required for LD-VMDAV, e.g. k and l, while a single parameter is sufficient for VMDAV. We refer to k' as the equivalent anonymity for VMDAV.

Similar to earlier results shown in Fig. 6, varying the number of active users does not affect the IL and PIP achieved by LD-VMDAV. Also note that, IL with respect to location and time are separately presented in Fig. 7a and b, respectively. As one

[11] The optimal class size for VMDAV is between k and $2k - 1$ [24].

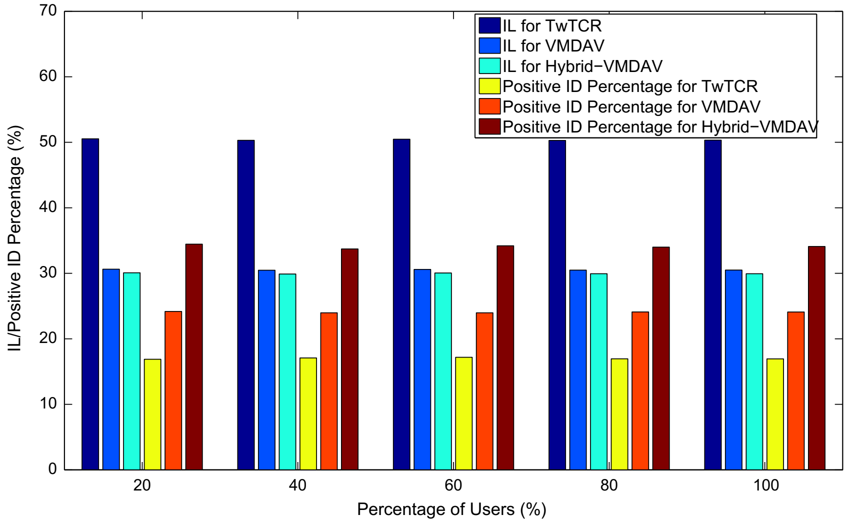

Fig. 6. PIP and IL as a function of the percentage of users uploading reports.

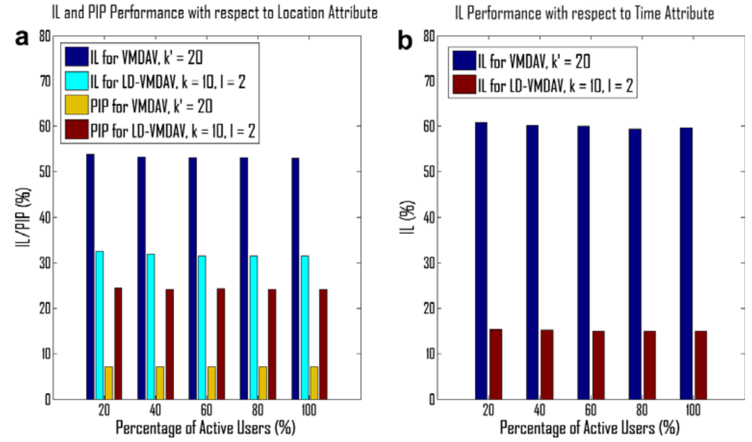

Fig. 7. IL and PIP results for LD-VMDAV and VMDAV.

can readily observe, LD-VMDAV outperforms VMDAV by reducing the amount of information loss. In particular, the IL for the location and time attribute is reduced by 40% and 75%, respectively. The improvement in location errors is not surprising since, in VMDAV a larger number of users, e.g. 20, are grouped to form equivalence classes. The resulting means of the location coordinates are thus expected to deviate more from the true user values. The huge reduction in timing errors is largely due to the de-coupling of location and time anonymizations built in LD-VMDAV. In VMDAV, no such separation exists and member similarity is measured by the combined location and time distances. Note that, users who are similar in terms of location may not necessarily have close time values.

In terms of PIP, it can be seen from Fig. 7a that LD-VMDAV enables an application to make much more accurate spatial decisions in comparison with VMDAV for an equivalent anonymity, i.e. k'. Observe also that the PIP values for LD-VMDAV in Fig. 7a are identical to those for VMDAV in Fig. 6. This is not an unexpected result since, the PIP metric used in this set of simulation measures the ability of the application to establish correct location associations. Now recall that, VMDAV and LD-VMDAV use the same mechanism to anonymize user locations. Hence, with the same parameter value ($k = 10$ for VMDAV in Fig. 6 and $k = 10$ for LD-VMDAV in Fig. 7), both algorithms should produce the same outcome.

Fig. 8 shows the impact of varying the diversity level on IL for LD-VMDAV. Since the anonymity level remains unchanged at $k = 10$, there is no difference in the location IL (Fig. 8a) for LD-VMDAV. However, increasing the level of diversity causes a proportional increase in the equivalent anonymity for VMDAV. As a result, the location IL for VMDAV increases. In addition, since VMDAV factors in location and time distances simultaneously, it is also subjected to increased timing errors as shown in Fig. 8b. Increasing the diversity level also has a negative impact on the temporal errors introduced by LD-VMDAV as seen in Fig. 8b. A larger diversity value creates a larger group of equivalence classes. As a result, reports with greater temporal disparity are merged in the same group, thus increasing the temporal IL. However, LD-VMDAV consistently outperforms VMDAV. The improvement achieved reduces slightly from 75% for $l = 2\%$ to 55% for $l = 4$.

6.3.3. Impact of Gaussian input perturbation

The last part of our simulation focuses on investigating the impact of Gaussian input perturbation on the performance of TwTCR, VMDAV, and Hybrid-VMDAV. Recall that in Section 4.4, users do not report their true locations to the AS in this enhancement. Instead, a random Gaussian noise is added to the true location prior to updating the AS. The simulations are run for different values of p, which range from 0.02 to 0.2 in increments of 0.02. Recall also that, p is the scaling factor used for controlling the amount of perturbation on user locations (see Section 4.4). The larger the value of p, the greater is the deviation from the true value.

Figs. 9 and 10 illustrate the impact of Gaussian input perturbation on TwTCR, VMDAV, and Hybrid-VMDAV when 40% and 80% of users contribute reports. Since the results exhibit some fluctua-

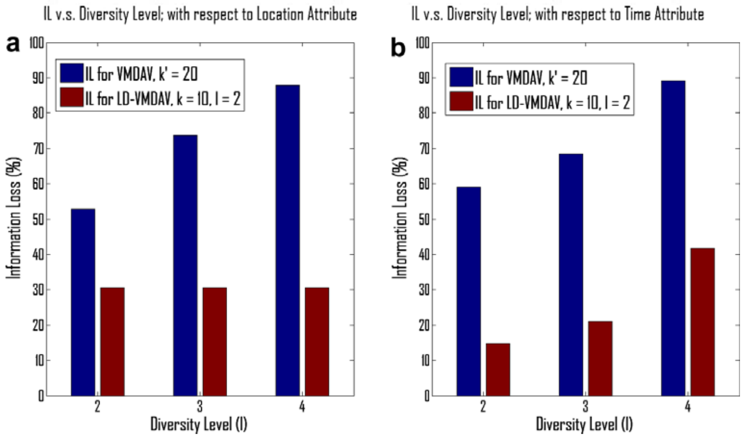

Fig. 8. Impact of varying the diversity level on IL.

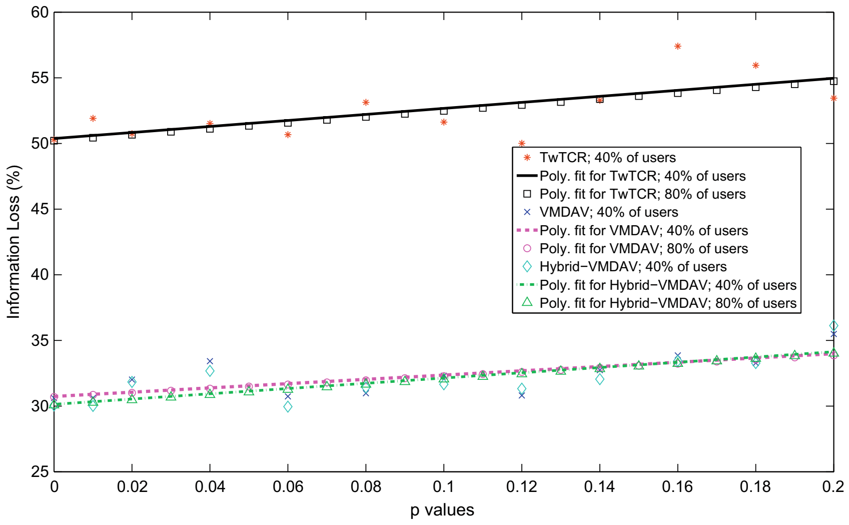

Fig. 9. Impact of Gaussian input perturbation on IL.

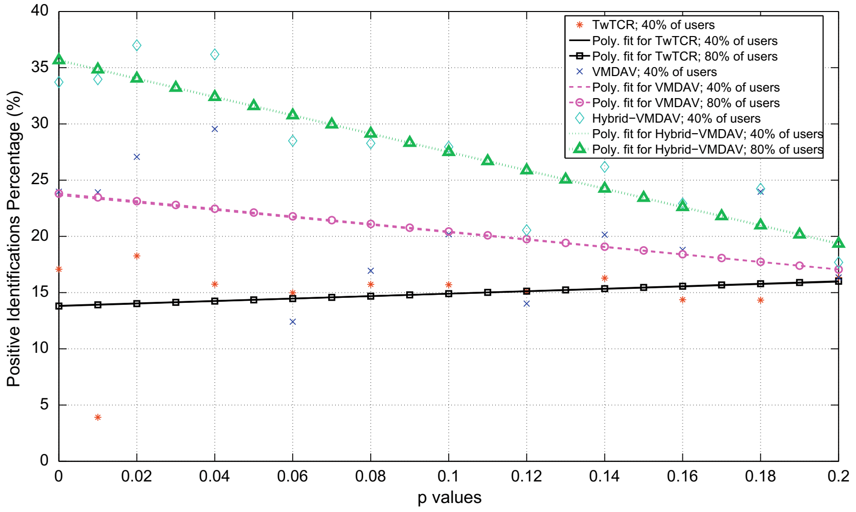

Fig. 10. Impact of Gaussian input perturbation on PIP.

tions, we fit them with polynomials of degree 1 to reveal the general trends. As in the previous simulations, the percentage of active users has negligible impact on the performance. Furthermore, the additional input perturbation degrades the performance of all three *k*-anonymous schemes. The level of performance degradation is more substantial for larger values of *p*. These results are expected since users are increasingly distorting their locations registered with the AS. Fig. 10 reveals that the performance gain of Hybrid-VMDAV gradually diminishes as *p* increases. Increasing the value of *p* implies that the resulting user distribution is more sparse, i.e. fewer cells are sufficient to provide the required level of anonymity on their own. Therefore, the VMDAV component of the hybrid algorithm tends to dominate. As a result, the performance of these two schemes converge. The results depicted in Fig. 10 also indicate that it is possible to guarantee satisfactory performance, without requiring 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-VMDAV. This achieves a good balance between user privacy and system performance.

7. Conclusions

This paper addresses user privacy in participatory sensing systems. The *k*-anonymity and *l*-diversity privacy models were thoroughly investigated. In the first part of this paper, we proposed TwTCR and VMDAV to overcome the shortcomings of the current state-of-the-art tessellation in securing location privacy of users in participatory sensing. We showed that these algorithms achieved better results in two contrasting situations and proposed Hybrid-VMDAV to take advantage of both schemes. The second part of this paper focused on demonstrating the inability of *k*-anonymous schemes in preventing attribute disclosure. Based on our threat model, we then proposed LD-VMDAV, a two-stage applications of VMDAV, to enhance user privacy. LD-VMDAV is based on the concept of *l*-diversity. We showed that, LD-VMDAV strengthens users location privacy by diversifying values for anonymized location while ensuring *k*-anonymity for anonymized times.

Our evaluations based on real-world data traces showed that Hybrid-VMDAV improved the percentage of positive identifications made by an application server by up to 100% and decreased the amount of information loss by about 40%, in comparisons with TwTCR. Our simulation results also indicated that LD-VMDAV outperformed its *k*-anonymous counterpart in terms of IL and PIP, while providing better privacy for users. Lastly, our studies suggested that perturbing user locations with random Gaussian noises can provide users with an extra layer of protection with a minimal impact on system performance.

References

[1] K.L. Huang, S.S. Kanhere, W. Hu, Towards privacy-sensitive participatory sensing, in: Proceedings of the the 5th International Workshop on Sensor Networks and Systems for Pervasive Computing (PerSeNS 2009), TX, March 2009.

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

[4] 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.

[5] 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.

[6] S. Gaonkar, J. Li, R.R. Choudhury, Micro-Blog: sharing and querying content through mobile phones and social participation, in: Proceedings of MobiSys 08, Breckenridge, CO, USA, June 17–20, 2008.

[7] E. Paulos, R. Honicky, E. Goodman, Sensing atmosphere, in: Proceedings of the Workshop on Sensing on Everyday Mobile Phones in Support of Participatory Research in Conjunction with ACM SenSys 2007, November 2007.

[8] S. Eisenman, E. Miluzzo, N. Lane, R. Peterson, G. Ahn, A. Campbell, The Bikenet mobile sensing system for cyclist experience mapping, in: Proceedings of ACM SenSys 2007, November 2007.

[9] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estin, 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.

[10] E. Miluzzo, N. Lane, K. Fodor, R. Peterson, S. Eisenman, H. Lu, M. Musolesi, X. Zheng, A. Campbell, Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application, in: Proceedings of the International Conference on Embedded Networked Sensor Systems (SenSys). ACM Press, New York, 2008, pp. 337–350, doi:10.1145/1460412.1460445.

[11] Y. Dong, S.S. Kanhere, C.T. Chou, N. Bulusu, Automatic collection of fuel prices from a network of mobile cameras, in: Proceedings of IEEE DCOSS 2008, June 2008.

[12] S. Sehgal, S.S. Kanhere, C.T. Chou, Mobishop: using mobile phones for sharing consumer pricing information, Demo paper, in: Proceedings of IEEE DCOSS 2008, June 2008.

[13] G. Calandriello, P. Papadimitratos, J.-P. Hubaux, A. Lioy, Efficient and robust pseudonymous authentication in VANET, in: VANET 07: Proceedings of the Fourth ACM International Workshop on Vehicular Ad Hoc Networks, ACM Press, New York, pp. 1928, 2007.

[14] K.P. Tang, J. Fogarty, P. Keyani, J.I. Hong, Putting people in their place: an anonymous and privacy-sensitive approach to collecting sensed data in location-based applications, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 93102, 2006.

[15] C. Cornelius, A. Kapadia, N. Triandopoulos, AnonySense: privacy-aware people-centric sensing, in: Proceedings of the Sixth International Conference on Mobile Systems, Applications, and Services, MobiSys'08, CO, June, 2008.

[16] A. Kapadia, N. Triandopoulos, C. Cornelius, D. Peebles, D. Kotz, AnonySense: opportunistic and privacy-preserving context collection, in: Proceedings of Sixth International Conference on Pervasive Computing (Pervasive), pp. 162–179, May 2007.

[17] L. Sweeney, K-anonymity: a model for protecting privacy, International Journal of Uncertainty Fuzziness and Knowledge-basrd Systems (2002).

[18] H. Kargutpa, S. Datta, Q. Wang, K. Sivakumar, On the privacy preserving properties of random data perturbation techniques, in: Proceedings of the IEEE ICDM, pp. 99–106, 2003.

[19] Z. Huang, W. Du, B. Chen, Deriving private information from randomized data, in: Proceedings of ACM SIGMOD Conference, pp. 37–48, June 2005.

[20] R.K. Ganti, N. Pham, Y.-E. Tsai, T.F. Abdelzaher, PoolView: stream privacy for grassroots participatory sensing, in: Proceedings of the 6th ACM conference on Embedded Network Sensor Systems, SenSys'08, pp. 281–293, November 2008.

[21] J. Domingo-Ferrer, J.M. Mateo-Sanz, Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering 14 (1) (2002) 189–201.

[22] J. Domingo-Ferrer, V. Torra, Ordinal, continuous and heterogeneous *k*-anonymity through microaggregation, Data Mining and Knowledge Discovery 11 (2005) 195–212.

[23] M. Laszlo, S. Mukherjee, Minimum spanning tree partitioning algorithm for microaggregation, IEEE Transactions on Knowledge and Data Engineering 17 (7) (2005) 902–911.

[24] A. Solanas, A Martinez-Balleste. V-MDAV: a multivariate microaggregation with variable group size, in: 17th COMPSTAT Symposium of the IASC, Rome, 2006.

[25] 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.

[26] D. Kotz, T. Henderson, I. Abyzov, CRAWDAD trace. Available from: <http://crawdad.cs.dartmouth.edu/meta.php?name=dartmouth/campus>.

[27] G. Zhong, U. Hengartner. A distributed *k*-anonymity protocol for location privacy, in: Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom), TX, pp. 253–262, March 2009.

[28] B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A.M. Bayen, M.Annavaram, Q. Jacobson, Virtual trip lines for distributed privacy-preserving traffic monitoring, in: Proceedings of the Sixth International Conference on Mobile Systems, Applications, and Services, MobiSys'08, CO, June 2008.

[29] T.M. Truta, B. Vinay, Privacy protection: *p*-sensitive *k*-anonymity property, in: Proceedings of the 22nd International Conference on Data Engineering Workshops, The Second International Workshop on Privacy Data Management (PDM'06), p. 94, 2006.

[30] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, *l*-diversity: privacy beyond *k*-anonymity, in: Proceedings of the 22nd International Conference on Data Engineering (ICDE), p. 24, 2006.

[31] H Jian-min, C Ting-ting, Y Hui-qun, An improved V-MDAV algorithm for *l*-diversity, in: International Symposium on Information Processing, pp. 733–739, May 2008.