Geoinformatica (2014) 18:165–191
DOI 10.1007/s10707-013-0193-z

User-side adaptive protection of location privacy in participatory sensing

Berker Agir · Thanasis G. Papaioannou · Rammohan Narendula · Karl Aberer · Jean-Pierre Hubaux

Received: 30 May 2013 / Revised: 31 August 2013 /
Accepted: 24 September 2013 / Published online: 22 October 2013

Abstract The participatory sensing paradigm, through the growing availability of cheap sensors in mobile devices, enables applications of great social and business interest, e.g., electrosmog exposure measurement and early earthquake detection. However, users' privacy concerns regarding their activity traces need to be adequately addressed as well. The existing static privacy-enabling approaches, which hide or obfuscate data, offer some protection at the expense of data value. These approaches do not offer privacy guarantees and heterogeneous user privacy requirements cannot be met by them. In this paper, we propose a user-side privacy-protection scheme; it adaptively adjusts its parameters, in order to meet personalized location-privacy protection requirements against adversaries in a measurable manner. As proved by simulation experiments with artificial- and real-data traces, when feasible, our approach not only always satisfies personal location-privacy concerns, but also maximizes data utility (in terms of error, data availability, area coverage), as compared to static privacy-protection schemes.

Keywords Location privacy · Privacy requirements · Utility · Participatory sensing

1 Introduction

The recent advances in sensor technology have led to a wide availability of privately-held low-cost sensors in mobile phones, vehicles, home appliances, etc. In turn, this has led to the development of the participatory sensing paradigm, that enables vast sensor data–from privately-held sensory devices–to be collected. In this paradigm, mobile devices send, either continuously or on demand, sensor data along with

This work has been partially supported by the EU project OpenIoT (ICT 287305).

B. Agir (✉) · T. G. Papaioannou · R. Narendula · K. Aberer · J.-P. Hubaux
School of Computer & Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, 1015, Switzerland
e-mail: berker.agir@epfl.ch

their locations and timestamps to an aggregation entity. The participatory sensing paradigm paves the way for innovative applications of great social and business interest, such as air pollution monitoring [19], early earthquake detection [16], and electrosmog exposure. This concept has attracted much attention from the research community, e.g., [4, 7], as it is an alternative to the costly and difficult-to-manage deployment of dedicated sensor-network infrastructures. However, participatory sensing faces serious challenges: data accuracy (i.e., due to the low-cost sensors), user privacy protection, finding incentives for users to contribute to the system, etc. Most importantly, users who are sensitive about their private information, such as their location (and inferred activities), are not expected to be willing to contribute to the system. Therefore, it is necessary in such systems to implement privacy-protection mechanisms such as anonymization of the source, and obfuscation of the location and/or time information attached to data. The usefulness of the sensed data for the corresponding application, however, depends on the accuracy, the availability and the spatio-temporal correctness of the data, all of which are negatively affected by privacy-protection mechanisms. For example, data accuracy decreases when location or time information are obfuscated, hence a trade-off emerges between data utility and user privacy. If a certain scheme that rewards mobile users according to the utility of their sensed data is in place, then, assuming that users are utility maximizers, they can more willingly provide data and enjoy a satisfactory level of privacy-protection.

An adversary who has access to users' spatio-temporal traces can find users' activity schedules [31]. To this end, the main objective of a protection mechanism is to provide untraceability. There exist location-privacy protection mechanisms [15] employing techniques such as data hiding or location obfuscation, with limited effectiveness against powerful adversaries that can exploit spatio-temporal associability of users' observable events, in order to partially or fully discover user trajectories. This happens because these techniques are employed with static parameters that cannot satisfy the user privacy requirements when the correlation between sequential user actions (i.e., user mobility and data emissions) and user's context lead to excessive leakage of personal information. Therefore, a user can never be sure that a static privacy protection will be successful against adversaries at all times. Another core problem of most of the existing approaches is the assumption that all users require the same level of privacy. This results in an unnecessarily high level of protection for some users and in insufficient protection for others. This second problem was addressed by Xiao and Tao [30] in the context of anonymization of datasets, which is not directly comparable to our context (where we do not consider anonymization).

In this paper, we propose an innovative approach for adaptive location-privacy protection in the participatory sensing context. Our objective is to provide the user with a statistical privacy guarantee at the lowest possible utility loss for the application. In order to achieve this objective, we define a personal privacy threshold θ, which is a lower bound on user location privacy. Before taking any privacy-protection action, in order to meet θ at the minimal utility cost, the privacy level of the user is dynamically measured on the user's device and compared with θ. Our adaptive scheme for location-privacy protection is lightweight, realistic and thus easily deployable at mobile devices. We consider two threat models: (a) A semi-honest aggregation server that attempts to extract and exploit private location information based on the emitted sensor data. (b) An active-tracking aggregation server,

which employs both the (partial) location history of the user and the emitted data for extracting and exploiting private location information. Using artificial- and real-data traces, we experimentally show that our approach, when feasible, satisfies the personal location-privacy protection requirements, based on the privacy techniques employed. By comparing our results with both real and artificial trajectories, we establish that the effectiveness of our approach is independent of mobility patterns. Moreover, it is shown that our approach increases the utility of the participatory sensing application, as compared to static privacy-protection policies, especially when user mobility history is partially available at the adversary. Finally, our work experimentally analyzes in a thorough manner the trade-off between utility and privacy in the context of participatory sensing. Note that our approach is compatible with most continuous or sporadic location-based applications (including location-based services).

The remainder of this paper is organized as follows: In Section 2, we present an overview of the related work. In Section 3, we describe our context, define personalized privacy, and set our privacy and utility-efficiency metrics. We present our adaptive privacy-protection scheme in Section 4 and, in Section 5, we experimentally assess its effectiveness. Finally, we conclude our paper in Section 6.

2 Related work

Privacy-preserving participatory sensing has been widely addressed by the research community in the past [2] including privacy of data itself, of data source identity and of user location. k-anonymity based solutions address data privacy and obfuscate a sensitive data item with a collection of $k-1$ items (referred to as *anonymity set*), so that the original item cannot be distinguished. Spatial cloaking techniques are proposed to preserve location-privacy, they often involve location *generalization* or *perturbation*. The idea behind generalization is to report a larger area instead of the exact user location (location obfuscation), whereas location perturbation applies a certain function on the real location of the user, e.g., the average location of multiple users. In addition, noise can be added to the data in an approach referred to as *randomization*, dummy data can be reported, or the sensed data can be *hidden* (i.e., not submitted at all to the application server). In our adaptive scheme, we employ location obfuscation and data hiding for privacy-protection.

The downside of any privacy-preserving mechanism in data-driven applications such as participatory sensing is the potential loss of accuracy or precision in the reported data and/or loss of samples. Krause et al. [14] address location privacy and experimentally analyze the trade-off between accuracy and privacy. They employ two methods of location-privacy protection: location obfuscation and sparse querying. The combination of these two methods diversify the users chosen for querying in order to minimize the privacy breach of a single individual user. In our paper, we significantly enhance the study of the trade-off between utility and privacy by studying the effect of privacy on additional utility aspects apart from accuracy, namely data completeness and area coverage.

According to [30], any privacy-preserving mechanism should consider the personalized privacy requirements of the participating users, because individuals typically have varying privacy requirements. In addition, we argue and experimentally prove in Section 5 that data utility can also be improved by personalized privacy protection

that avoids excessive privacy preservation. In [30], the authors formalize personal privacy specifications and apply a data generalization technique for satisfying individual privacy requirements. The authors in [9] propose a location-privacy protection mechanism based on personalized k-anonymity for location-based services (LBS). However, they employ a trusted third-party that implements the privacy-protection scheme, which is contrary to our approach; similarly, Vu et al. [27] also propose a trusted third-party based k-anonymity approach. In [4, 18], the users might decide to selectively activate sensing (and hide in other times) depending on a variety of factors, such as presence in sensitive locations (home or office), or their current social surroundings (presence of friends or family members). However, hiding is applied not based on a rigorous privacy assessment, but based on a fixed probability value. The authors in [17] analytically prove that trajectory inference is still possible in a LBS if data hiding is the only mechanism used for location-privacy protection and they suggest designing new policies that consider users' past events, as we do in our work in the context of participatory sensing.

In addition to being a client-based location-privacy preserving mechanism, our approach supports continuous location dissemination. Several client-based solutions exist in the literature [11, 12, 21, 23]. SybilQuery [23] generates, for each user's query, $k-1$ other queries so that the LBS server cannot distinguish the real query from the Sybil ones. However, this work requires the user to determine a priori the source and destination of the real query, thus it does not support real-time continuous dissemination. In addition, it does not apply any transformation/obfuscation on the trajectories, which allows an adversary to obtain the full real trajectory, once it is identified partially. A distributed k-anonymity cloaking mechanism is proposed in [11], which identifies neighbors using on-board wireless modules and exploits secure multi-party computation in a collaborative manner in order to compute a cloaking region. However, this work does not support continuous querying. Finally, Jadliwala et al. [12] present a concept called *privacy-triggered communications*, which is a generic framework that fits our work; our work differs in detailed utility and privacy analysis.

Last but not least, [3, 5] propose cryptographic approaches for protecting the identity of the participants in participatory sensing. Groat et al. [10] consider multidimensional data to evaluate the user privacy, i.e., they consider spatio-temporal dimensions, the sensed data and more. But, they do not take into account the continuous data disclosure, which would be disastrous for the users in case of an attack on a multidimensional scale. More on the privacy issues in participatory sensing applications can be found in the survey paper by Christin et al. [2].

In summary, to the best of our knowledge, none of the existing work proposes a location-privacy protection scheme combining the following properties: (i) dynamic estimation of user privacy based on the history of mobility and data submissions, (ii) adaptive satisfaction of personalized privacy requirements, (iii) user-side residence, and (iv) independence of any trusted third parties.

3 System model and performance metrics

People are concerned about the potential (though unconfirmed) health risks due to base stations [29]. Therefore, in this paper, we consider the application of

electrosmog monitoring by means of participatory sensing, as this case study fits the continuous data dissemination scenario.

We assume that a mobile user can always submit its sensor data using its own data plan through the cellular network. In this context, mobile users (or just "users") sense their environment and send their sensor data to a certain data-collection entity called an aggregation server (AS). Such data is valuable only if it is accompanied by the location and time information, hence the reported data packets are triplets in the form of $\langle value, location, time \rangle$. Our objective is to provide the user with a statistical privacy guarantee at the lowest possible utility loss for the application in this setting.

In our approach, we avoid relying on a trusted third party, because in reality it is difficult to establish such an entity that is trusted by all participants. Furthermore, we assume that users do not collaborate with each other in order to protect their location privacy, because this approach is energy-costly and enables users to collude in order to breach others' privacy. In such a setting, hiding the identities of mobile users is rather unrealistic. Consequently, we focus on user untraceability and do not consider hiding user identity as a protection mechanism (i.e., the AS knows the source of each sensed data).

For presentation clarity and computational limitations, throughout the remainder of the paper, we assume the monitored area to be partitioned into cells and the time to be slotted. Henceforth, we use the terms 'location' and 'grid cell' interchangeably. In the remainder of this section, we specify the adversary models, define personalized privacy, and describe metrics for the evaluation of privacy and utility.

3.1 Threat models

We consider two threat models; the adversary in both is assumed to be the AS, who records and exploits the private information that it obtains. The communication between the AS and the users is assumed to be encrypted, and the AS knows the identities of users.

In the first model, the AS is assumed to be semi-honest [1], meaning that it follows the protocols, it does not collude with other entities and it does not tamper with the system to obtain private information about the users. Furthermore, it does not deploy devices to monitor the whereabouts of users (no global or local eavesdropping). As a result, it can only try to infer private information based on the data it collects from the users. In this model, the AS has no background information on the users' mobility.

In the second model, the AS is assumed to be an active adversary and to deploy a limited number of tracking devices constrained by cost and resources. In this regard, we assume that the AS is able to detect user presence in a fraction of locations and it uses the information collected to reconstruct the original traces. For example, the AS can do this by sniffing the control channels of the cellular communication where the handshakes between the users and the base stations are exchanged in clear text. At this point, we argue the AS cannot optimally choose the monitored locations, because it cannot know the location sensitivities of the individuals. One approach would be to monitor the hotspot or generally-sensitive areas (such as hospitals) in a city, but then any other user movement would not be captured. Moreover, some users might not be very privacy-sensitive to their presence in hotspot areas. Therefore, we assume that the AS chooses randomly the locations to monitor. The AS uses the tracking data to build a *spatio-temporal probability distribution* for

each user. For example, on Mondays 9am with probability 0.8 a particular user is at work, and the probabilities assigned to user's other possible locations sum up to 0.2. To reveal the user trajectories, this background knowledge is combined with the location information contained in the emitted data from the mobile users. Note that the notion of the spatio-temporal probability distribution is very generic and can model other kinds of background knowledge as well, i.e., user habits, user location sensitivity, location semantics, etc.

We also assume that, in both threat models, the only other background information that the AS has about the users is their maximum possible speed (also known to the users themselves). Nevertheless, mobile users are assumed to be honest, which means that they do not attempt to tamper with their sensor measurements or collude with the adversary, but they might reduce the data accuracy (in terms of location/time), in order to protect their privacy. Last, we assume no interaction among users; consequently, there is no risk of potentially malicious users aiming to track other ones.

3.2 Personalized privacy

In most of the existing location-privacy protection approaches [15], fixed parameters are statically employed in the proposed mechanisms for all the users. This approach has a negative effect on both the privacy levels of the users and the utility of the system, as will be shown in Section 5.

First of all, such a static approach does not take into account the trajectory history of users. It implicitly assumes that a uniform parameter for a particular location-privacy protection mechanism will always provide the same level of protection, which is not the case because spatio-temporal correlation between disclosed events might reveal partial or full trajectories of users.

Another problem resulting from this approach is the negative effect on the utility of the system due to the fact that in some cases the provided location privacy can be much higher than what a user actually wants. For example, a user might still achieve satisfactory privacy-protection by providing 4 grid cells in an obfuscated area instead of 6, and therefore increase the system utility.

According to Westin [28], "each individual is continually engaged in a personal adjustment process in which he balances the desire for privacy with the desire for disclosure and communication of himself to others, in light of the environmental conditions and social norms set by the society in which he lives". In this spirit, considering the aforementioned issues about the static/uniform parameter selection, we define personalized privacy as follows:

Definition 1 Given a set P of protection mechanisms and $P^{a_i} \subseteq P$ being the subset of mechanisms that can be implemented by the user i with ability a_i, the *personalized privacy* for user i with privacy threshold θ_i is defined by the formula:

$$\exists p \in P^{a_i} : p(z, \omega, H) \geq \theta_i, \tag{1}$$

where $z \in Z$ is an instantiation of adversary capabilities Z, $\omega \in \Omega$ is a particular user action from the set of available actions Ω, H is the history of user actions, and $p(.)$ is the privacy level resulting from the mechanism p as estimated by the user.

According to this definition, personalized privacy is the individual's ability to employ all the necessary privacy protection mechanisms so as to adapt to privacy leakage resulting from his/her activities and/or the changing privacy-breaching capabilities of the adversary, as observed by the individual.

3.3 Privacy metrics

In the literature, several metrics have been proposed for measuring the level of location privacy. Some approaches, such as [9], utilize as metrics the parameters of the privacy-protection mechanism, e.g., the size of the anonymity set, the size of the obfuscation area. However, these simple metrics do not assess the actual location-privacy protection level of users, because they do not take into account the adversary's capabilities, users' history of events and the correlations between users' events.

Dwork proposes the concept of differential privacy [8], which is a privacy measurement approach for statistical databases. In this approach, the privacy is measured by the predictability of the existence of a single record based on a statistical result obtained from a database. In our work, single data values are provided to the server; thus, differential privacy is not a suitable metric in our case.

A more useful metric is *entropy*, which is an information theoretical approach to privacy measurement. In [6, 22], entropy H is proposed as an anonymity metric to measure the privacy offered by a system. It is defined as $H = -\sum_i p_i \log_2 p_i$, where p_i is the attacker's estimate of the probability that a participant i is responsible for some observed action. In our context of participatory sensing, the observed actions consist of reported locations at a specific time, thus entropy can be used to measure how well the actual location is hidden among this location set, i.e., the uncertainty of the adversary about the actual location of a user.

Although entropy adequately assesses the uncertainty of the adversary, it does not measure the correctness of adversary's estimation. To this end, Shokri et al. [24] propose a distortion-based metric, where the uncertainty and the correctness of the adversary are measured by assigning probabilities to all possible trajectories of a user and by calculating the distances of the fake trajectories from the real ones. The distances are multiplied with their respective probabilities in order to obtain the *expected distortion* (ED) of location inference for a corresponding user. ED is given by the following formula:

$$ED(u,t) = \sum_{\Psi} D(\text{actual}(u,t), \Psi(t)) \cdot \Pr(\Psi, t) \tag{2}$$

where $ED(u,t)$ is the expected distortion of user u at time t and Ψ represents all the observed trajectories of user u. actual(u,t) gives the actual location of user u at time t and $\Psi(t)$ is the location on trajectory Ψ at time t. $\Pr(\Psi, t)$ is the probability assigned to trajectory Ψ at time t by the adversary. In our work, we define $D(loc_1, loc_2)$ as a normalized distance function that gives distance between locations loc_1 and loc_2 in [0, 1] and therefore, the privacy level computed is in the interval [0, 1], where 0 means no privacy protection and 1 means full privacy protection. This is done by normalizing the actual distance by an upper bound distance per time step (e.g., the

maximum driving speed in our case).[1] We employ the Euclidean distance function, but other choices of distance functions are possible as well.

Shokri et al. [25, 26] further extend the idea behind the distortion-based metric and present a comprehensive location-privacy quantification framework. This framework formalizes the attack of the adversary, takes into account its background information on users' mobility patterns and calculates users' location-privacy protection levels based on the adversary's *accuracy*, *correctness* and *certainty* about users' actual trajectories. The authors also propose a software tool, called *Location Privacy Meter* (LPM), which implements this framework. The LPM consists of several attack strategies based on complex algorithms. In our context, since it suits our scenario, we consider the *localization attack*, which is explained in [26]. This attack is an attempt to find the most likely location, at each time instant, for a particular user among all of her observed locations, based on her observable events both in the past and in the future. The complexity of this attack is $O(TM^2)$ [26] for one trace, where T is the number of time instants and M is the number of locations in the area of interest (i.e., the monitored area).

3.4 Utility metrics

The utility of participatory sensing applications is crucial to their emergence and economic sustainability. The utility in this context depends on the data quality, the data relevance to the application and the data availability. Here, we focus on data quality and availability aspects, namely the data accuracy, the data completeness and the area coverage. We analyze the effect of privacy protection on utility, based on the aspects explained below:

- *Data Accuracy*: As the users report imprecise or coarse-grained location (and/or time) information in their sensed data, an error is introduced in the measurements of other locations(and/or time instants). We measure the data inaccuracy by means of the average absolute error (L_1 norm) introduced to the sensed data due to location/time obfuscation. We express the average absolute error as a percentage of the data range.
- *Data Completeness*: One important factor that affects data availability is data loss; some of the sensed data collected by the users might not be emitted (data hiding) to the AS due to privacy concerns. We define the data completeness as the percentage of the sensed data received by the AS.
- *Area Coverage*: Another component of the data availability is the percentage of the area of interest, where sensor measurements are done by users. Various privacy-enabling techniques differently affect the size of the total monitored area: e.g. while data hiding tends to decrease it, location obfuscation tends to increase it, as observed by the AS. As higher data hiding and larger obfuscation negatively affect the data availability, we define the area coverage as the fraction of the areas in which data is sensed over all areas where data is reported by the

[1] $D(loc_1, loc_2)$ can be the absolute distance function, in which case the expected distortion would be in km or meters. We choose to normalize it for the sake of presenting results with a uniform upper bound on the privacy level.

users. Note that this metric is maximized at 1, i.e., all areas where data is reported correspond to real points of sensor measurements.

4 Our adaptive scheme

In this section, we introduce a simple, yet effective location-privacy protection scheme, that is built upon the existing privacy-preserving techniques of location obfuscation and hiding. The main idea is that before each user submits data, she should be able to estimate locally her expected privacy-level and configure the protection mechanisms accordingly. This requires us to emulate an adversary's attack on user devices; however, due to limitations on processing power and also battery capacities, we need to achieve this by implementing a light-weight approach. The location-privacy quantification framework proposed by Shokri et al. [25, 26] is very comprehensive and useful, but it is computationally heavy, as explained in Section 3.3. Thus, we employ the distortion-based metric [24] and a Bayesian-network approach on the user-side, in order to calculate locally an estimate of user privacy-level on mobile devices. Note that in the remainder of this section, the term 'node' is used to refer to the users' mobile devices, because it is user's devices where the scheme runs and users do not take action.

We employ location obfuscation for confusing the aggregation server (AS) about the actual location of the sensed data. Location obfuscation is the generalization of the fine-grained location information; we designate its granularity with λ, which is the obfuscation parameter. As stated in Section 3, a location is a grid cell, and therefore, an obfuscated area is a set of grid cells. In our strategy, a reasonable upper bound λ_{max} on λ is assumed, so that the sensor data remains useful for the participatory sensing application.

In our scheme, we want to let people have the privacy protection level they desire. In order to provide this, we define θ, the *personal privacy threshold*, which expresses the desired level (i.e., the lower bound) of expected distortion (i.e., distance) from the actual user location. This privacy threshold depends on the user's sensitivity about her privacy at a particular location, and it can be chosen by a user-specific function of the desired absolute distance from the sensitive location (cf. Fig. 1).

The algorithm for determining the obfuscation is as follows. When a node has data to submit, it calls the location obfuscation module with the lowest λ, i.e., $\lambda = 1$. Then, it provides the output of this module—a set of locations constituting the obfuscation

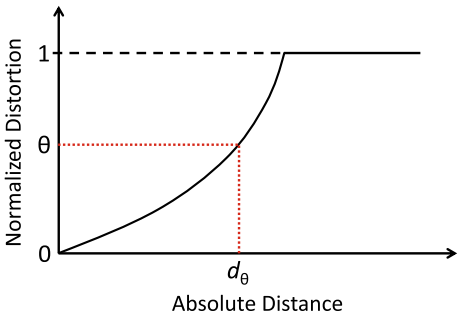

Fig. 1 Absolute distance vs. distortion. θ is the desired privacy threshold and d_θ is the corresponding absolute distance to achieve θ

area—to the privacy level estimation module. The estimation is then compared against the node's privacy threshold θ. If θ is reached, then the node submits the data to the AS with the last generated obfuscation area. Otherwise, it increases λ and repeats the process. If λ_{max} is reached, but not θ, then the data is not submitted.

Our obfuscation algorithm randomly determines the obfuscation area as explained in Section 4.1. A randomly chosen obfuscation area might be ineffective, whereas another obfuscation area of the same size can provide sufficient privacy protection. Finding the optimal obfuscation area would be time and energy consuming, hence we introduce a limit (i.e., by means of a counter) on the number of obfuscation areas we try: α_{max} per λ level. α is the number of obfuscation areas that have been tried for satisfying θ with the same λ value. As long as θ is not reached and $\alpha < \alpha_{max}$, another obfuscation area of the same size is generated and privacy level is estimated based on this new area. Otherwise, if $\lambda < \lambda_{max}$, then λ is incremented and the process is repeated. Figure 2 shows this adaptive privacy protection strategy as a flowchart.

We explain, in Section 4.1, the obfuscation mechanism we employ and in Section 4.2 how local estimation is done.

4.1 Location obfuscation mechanism

The location obfuscation mechanism we employ in our proposed scheme and in the static mechanisms takes two inputs: the obfuscation level λ and the actual location l that is subject to obfuscation. Since the area of interest is discretized, the obfuscation area to be generated consists of a set of grid cells including the actual location/cell l.

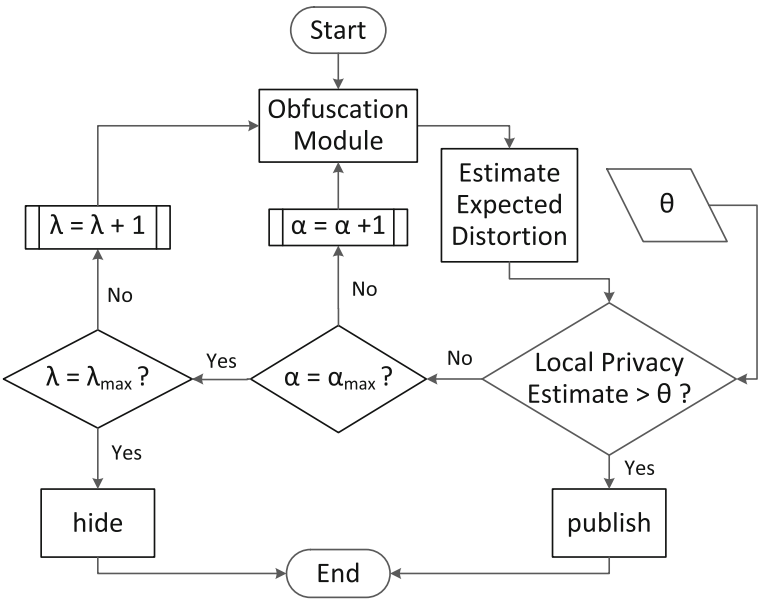

Fig. 2 Adaptive location-privacy protection system. Expected distortion estimation keeps track of user history in case of active adversary assumption

λ actually encodes the size of the obfuscation area in terms of cells. First, we determine the size $s_x \times s_y$ of the obfuscation area according to λ as follows:

$$s_x := 1 + \lceil \lambda/2 \rceil$$
$$s_y := 1 + \lfloor \lambda/2 \rfloor ,$$

where $\lceil . \rceil$ and $\lfloor . \rfloor$ are the ceiling and floor operations, respectively. Then, the area of size $s_x \times s_y$ cells is randomly positioned over the actual location l. Note that any deterministic choice for this positioning would render the area generalization ineffective in terms of privacy, because the adversary can find the actual location by trying different obfuscation areas iteratively. Randomization avoids adversary from finding the actual location l, because in this case any of the locations in an obfuscation area is equally likely the actual location without any a priori knowledge. Note that some of the locations in an obfuscation area may be infeasible to reach from observed location in the previous time instant due to the maximum speed constraint. Such constraints are taken into account in the local privacy-level estimation described in the next subsection.

Figure 3 shows an example of this obfuscation mechanism over a gridded area, where the actual location l is in the center and $\lambda = 6$. The obfuscation area can be positioned in the bounding box in this figure, so long as l remains part of it. The two 4×4 squares in the figure represent the top-left and bottom-right possible obfuscation areas with $\lambda = 6$.

4.2 Local privacy-level estimation

We calculate locally the expected privacy-level at the node, as explained below. We maintain locally an event linkability graph at each node, as the one depicted in Fig. 4. Each vertex in this graph represents an event observable by an adversary at the corresponding time instant. An observable event corresponds to a data item associated to a particular location, which is sent to the AS. The linkability graph helps us to identify the trajectories that the AS observes and also to estimate its belief about their authenticity.

In order to build the linkability graph, a node needs to know the geographical topology of the area, i.e., it needs to know the potential connectivity among different locations. It also needs to know the assumptions made by the AS for inferring user trajectories. To this end, one important assumption made by the AS is the maximum possible speed of a mobile node, which also determines the maximum

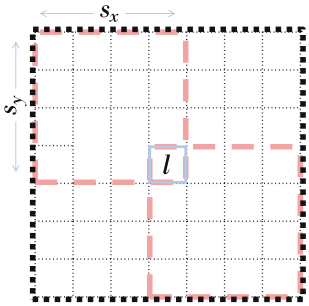

Fig. 3 The area in which an obfuscation area of size 4×4 ($s_x = s_y = 4$) can be positioned based on actual location l given $\lambda = 6$

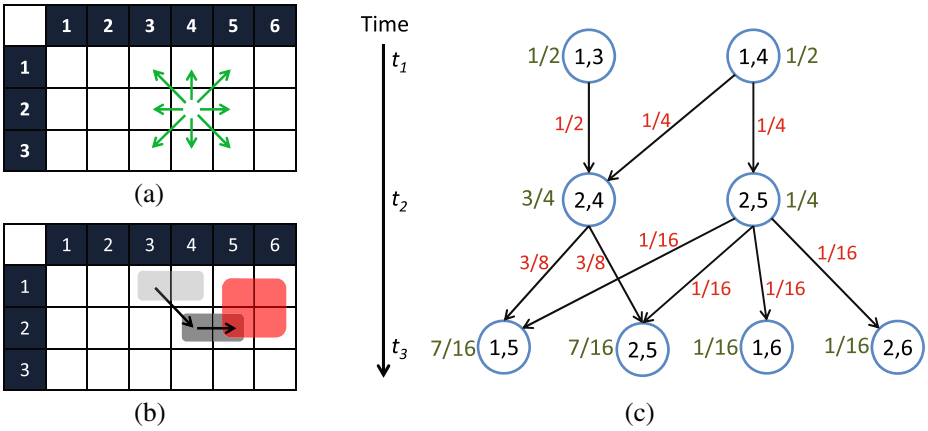

Fig. 4 Example of estimation of probabilities of possible trajectories for 3 time steps of a node. **a** Possible moves in a single step for a node on the given area. **b** Real trace of a node and its obfuscation decisions at each time instant. **c** Inferred linkability graph, on which the probabilities of being there are assigned to each edge and vertex. Each vertex on this graph represents an observed event from the node and the indices on the vertices are the location ids w.r.t. the area in Fig. 4a and b

possible distance between sequential vertices in time. A node can extract its own maximum speed from its traces, but it is not practical for the AS to know this value for each individual node. Nevertheless, he can make a global estimation on the average maximum speed and choose it as the upper bound for all the nodes he wants to attack. A node can construct, based on this knowledge, its linkability graph by connecting the vertices (i.e., the observable events) that are adjacent in time and space.

Since the user may have to continuously disclose her data, hiding at a specific time instant does not provide her with full privacy protection. Technically, hiding, as perceived by the adversary, produces yet another obfuscation area that is the maximum feasible one based on the maximum speed and the user's previous reported locations. In practice, for the current time instant, this yields all the locations that are reachable from the locations in the previous time instant, and we consider all such locations as observable events in the privacy-level estimation.

The linkability graph is progressively constructed as new events are produced over time. The vertices corresponding to the current time-instant are connected to the vertices from the previous time-instant, based on the feasibility of being adjacent in space and time. If there are vertices with no children in the previous time instants, then these vertices are identified to be impossible and are removed. The same is applied to the vertices with no parents in the current time-instant. Note that the elimination of vertices needs to be propagated in the whole graph because some vertices in older time instants might lose all their children, which suggests that they are no longer probable locations of the node.

We use the linkability graph and employ the Bayes rule to calculate the probability that an observed event corresponds to the actual location of the node. The first time observed events are inserted to the graph, a uniform probability $1/k$ is assigned to each vertex, as dictated by the k-anonymity employed by the chosen location obfuscation level, where k is the number of vertices. As new vertices are added at a subsequent time-instant, they can only be children of those in the previous

time-instant and their probabilities of being genuine are calculated according to the Bayes rule. Also, after the elimination of impossible events, the probabilities assigned to the siblings or parents of these events are updated and these updates are propagated in the graph. The probability of an event being genuine is depicted in Fig. 4 as a label beside its corresponding vertex. We explain the calculation of these probabilities following the example of Fig. 4. Initially, at time t_1, locations $(1, 3)$ and $(1, 4)$ are reported by the node and thus $\Pr(loc_{t_1} = (1, 3)) = \Pr(loc_{t_1} = (1, 4)) = 1/2$. Then, at time t_2, the node reports two locations to the AS, namely $(2, 4)$ and $(2, 5)$. We calculate the probability that location $(2, 4)$ is genuine as follows:

$$\begin{aligned}
&\Pr(loc_{t_2} = (2, 4)) \\
&\quad = \Pr(loc_{t_2} = (2, 4) | loc_{t_1} = (1, 3)) \cdot \Pr(loc_{t_1} = (1, 3)) \\
&\qquad + \Pr(loc_{t_2} = (2, 4) | loc_{t_1} = (1, 4)) \cdot \Pr(loc_{t_1} = (1, 4)) \\
&\quad = 1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}
\end{aligned}$$

The same approach is applied for location $(2, 5)$ and for the 4 observed locations reported by the node at time t_3. [2]

After having calculated the probability of each leaf vertex being genuine, a node calculates its expected distortion (ED) according to Eq. 2. For example, the ED of the node in our example at time t_3 is calculated as follows:

$$\begin{aligned}
ED(u, t) &= \sum_{\Psi} D(\text{actual}(u, t), \Psi(t)) \cdot \Pr(\Psi, t) \\
&= 1 \cdot \frac{7}{16} + 0 \cdot \frac{7}{16} + \sqrt{2} \cdot \frac{1}{16} + 1 \cdot \frac{1}{16},
\end{aligned}$$

where $D(.)$ stands for the Euclidean distance in this example.

Background information So far, we have explained how the privacy leakage is estimated locally by a mobile node, under the assumption of no background information about the node's mobility at the adversary side. Now, we consider that some background information on node's mobility is possessed by the adversary. Specifically, we assume that the adversary has, for each mobile node, a prior spatio-temporal probability distribution with PDF $\pi(X, t)$ over the locations X at time t that is built based on partial leakage of location information. As the mobile node does not know the exact leakage of its mobility, it samples its mobility history and builds a similar prior distribution, in order to accurately estimate its privacy leakage to the adversary by its emitted data. The prior distribution is employed to calculate the transition probabilities between successive locations.

[2]Note that the size of the obfuscation area at time t_3 is 2×2 (as shown in Fig. 4b), therefore there are 4 vertices corresponding to 4 reported locations at this time instant.

For example, in Fig. 5, assume a prior spatio-temporal distribution as follows:

$$\pi(X=(1,3),t=t_1)=1/16,\ \pi(X=(1,4),t=t_1)=1/8$$
$$\pi(X=(2,4),t=t_2)=1/10,\ \pi(X=(2,5),t=t_1)=1/20$$
$$\pi(X=(1,5),t=t_3)=1/10,\ \pi(X=(2,5),t=t_3)=1/5$$
$$\pi(X=(1,6),t=t_3)=1/10,\ \pi(X=(2,6),t=t_3)=1/20\,.$$

By employing this prior distribution for calculating the transition probabilities, we derive that:

$$\Pr(loc_{t_1}=(1,3))=\frac{\pi(X=(1,3),t=t_1)}{\pi(X=(1,3),t=t_1)+\pi(X=(1,4),t=t_1)}=1/3$$
$$\Pr(loc_{t_1}=(1,4))=\frac{\pi(X=(1,4),t=t_1)}{\pi(X=(1,3),t=t_1)+\pi(X=(1,4),t=t_1)}=2/3$$
$$\Pr(loc_{t_2}=(2,4)\mid loc_{t_1}=(1,3))=1$$
$$\Pr(loc_{t_2}=(2,4)\mid loc_{t_1}=(1,4))$$
$$=\frac{\pi(X=(2,4),t=t_2)}{\pi(X=(2,4),t=t_2)+\pi(X=(2,5),t=t_2)}=2/3$$
$$\Pr(loc_{t_2}=(2,5)\mid loc_{t_1}=(1,4))$$
$$=\frac{\pi(X=(2,5),t=t_2)}{\pi(X=(2,4),t=t_2)+\pi(X=(2,5),t=t_2)}=1/3$$

Then, based on these transition probabilities, $\Pr(loc_{t_2}=(2,4))$ can be calculated again as follows:

$$\begin{aligned}&\Pr(loc_{t_2}=(2,4))\\&=\Pr(loc_{t_2}=(2,4)|loc_{t_1}=(1,3))\cdot\Pr(loc_{t_1}=(1,3))\\&\quad+\Pr(loc_{t_2}=(2,4)|loc_{t_1}=(1,4))\cdot\Pr(loc_{t_1}=(1,4))\\&=1\cdot\frac{1}{3}+\frac{2}{3}\cdot\frac{2}{3}=\frac{7}{9}\end{aligned}$$

Therefore, the background information can significantly affect the expected distortion that can be achieved by a privacy-protection strategy.

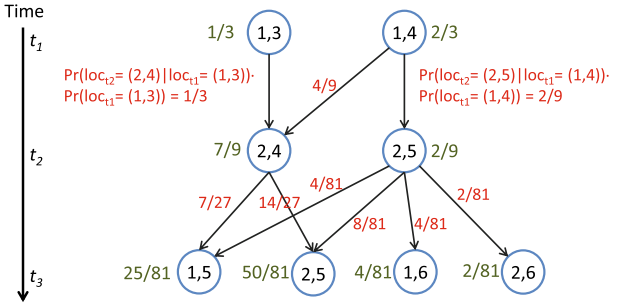

Fig. 5 Inferred linkability graph when background information is assumed to be available at the adversary. The prior spatio-temporal distribution is employed to find transition probabilities between successive locations

Complexity analysis The complexity of our algorithm is dominated by the maintenance of the linkability graph. Each time a data submission is about to be made, the obfuscation module generates a maximum number of L locations constituting an obfuscation area. This operation has time complexity $O(L)$. The pairwise connectivity check between the locations in consecutive time instants takes $O(L^2)$ times. Later, the probabilities assigned to the current observed events are calculated in $O(L)$. Therefore, the time complexity of the whole process is $O(L^2)$. Note that this process has to be repeated until θ is met. The number of repetitions, however, is bounded by a constant maximum obfuscation parameter λ, thus, the total time complexity of the whole estimation and protection operation remains $O(L^2)$. This is a better performance than the LPM, as we explain in the following example: Given an area of interest of 20×25 grid cells (i.e. $M = 500$) and a maximum obfuscation parameter $\lambda = 10$, our complexity is $O(36^2)$, whereas the LPM complexity is $O(500^2)$, i.e., almost 250 times slower.

Our approach is lightweight in terms of space requirements as well. In addition to map topology—which would be required by any client-side location-privacy protection mechanism—our scheme only stores the linkability graph, where each vertex has a probability value and location information. This results in $O(TL)$ vertices and $O(TL^2)$ edges in the worst-case, where T is the number of elapsed time instants. These storage requirements can be easily handled by modern mobile devices, that presumably have several GBs of storage capacity.

5 Evaluation

In this section, we assess the performance of our adaptive approach for protecting location-privacy and compare its effectiveness with that of static protection policies in terms of utility and privacy. To this end, we perform simulation experiments, with not only artificial data sets, but also real data traces (explained in Section 5.1). The estimate of the privacy level of a user, as observed by the AS, is measured by the LPM [25, 26]. This software tool provides an objective estimate of the privacy level of users, and its output belongs in [0, 1], with 0 meaning no privacy protection and 1 meaning maximum protection.

We replayed real data traces in a simulation environment, that we developed in C++, and ran experiments using artificial data traces (cf. Section 5.2). We implemented our adaptive strategy, along with static protection mechanisms of obfuscation and hiding, which works with fixed λ and hiding probability $\Pr_h$, respectively. $\lambda_{\max}$ was set to 10, which means that the largest possible obfuscation area is of size 6×6 grid cells. For expected distortion computation, we used Euclidean distance.

For the scenario in which some background information is assumed to be available at the adversary, we let the AS monitor all user presence in 25 random locations. Given a total of 500 grid cells in the sensed area, the expected number of node events observable at the adversary is given by the formula below:

$$\sum_{i=1}^{500} \frac{25}{500} \cdot (\#\text{ events generated in } l_i) \tag{3}$$

In our dataset, each node has around 20,000 events on the average. Given the above formula, the expected number of exposed events of a node corresponds to 1 % of

Table 1 Parameters λ, Pr_h of the Avg and Max static policies experimentally found to satisfy the various privacy thresholds on average and most of the time respectively

θ	Avg Static				Max Static			
	w/out BK		w/ BK		w/out BK		w/ BK	
	λ	Pr_h	λ	Pr_h	λ	Pr_h	λ	Pr_h
0.1	1	0	1	0.2	1	0.1	2	0.3
0.2	1	0.1	3	0.3	1	0.2	3	0.5
0.3	2	0.1	3	0.5	2	0.2	4	0.6
0.4	4	0.1	4	0.6	4	0.2	5	0.7
0.5	4	0.2	6	0.7	4	0.3	7	0.8
0.6	8	0.2	6	0.8	8	0.3	8	0.4
0.7	8	0.4	7	0.9	7	0.5	8	0.5
0.8	9	0.6	9	0.6	9	0.7	9	0.8
0.9	10	0.8	10	0.9	10	0.9	10	0.98

its generated events. The mobile node does not know which locations are monitored by the adversary. Although, knowing that the expected total number of its leaked events to the adversary is 1 %, it can consider a random 1 % of its generated events as the background knowledge available at the adversary. This gives the node a chance to take into account the adversarial background knowledge in the local inference module.

First, we compare our adaptive strategy to combinations of the aforementioned static mechanisms and experimentally prove the ineffectiveness of static policies at satisfying user privacy requirements. Then, we demonstrate that our simple local estimation of privacy is an accurate measurement. Subsequently, we analyze the trade-off between utility (i.e., accuracy, area coverage, data completeness) and privacy for different static policies and our adaptive privacy-enabling policy. To this end, we define two different static policies for a given θ:

- *Avg Static:* This policy defines fixed Pr_h and λ that meet θ on the average; privacy violations are allowed from time to time.
- *Max Static:* This policy defines fixed Pr_h and λ so that θ is met most of the time. This is a rather conservative privacy-protection policy.

Note that these static policies employ the obfuscation mechanism described in Section 4.1 and apply this mechanism statically with the predefined parameters. They do not consider past or future events of the node when obfuscating the actual location.

Table 1 shows the experimentally identified static privacy-protection policies corresponding to each privacy threshold adapted in simulations, and Table 2 shows the parameters we have used for the experiments.

Table 2 Experiment parameters

	Adaptive	Static
$\boldsymbol{\theta}$	0.1–0.9	N/A
λ	Adaptive	1–10
$\mathbf{Pr_h}$	N/A (Adaptive hiding)	0–0.9
# of nodes	20	
Monitored area	25 × 20	
$\boldsymbol{\lambda_{max}}$	10	

5.1 Real data trace

During the Lausanne Data Collection Campaign (LDCC) [20], run by Nokia Research Center (Lausanne), a dataset of around 200 users collected. The data was collected over a year from 2009 to 2011, from smart-phones that were provided to the participants. We utilize 20 time-continuous user traces and we consider an area of $1.25 \times 1.00\,\text{km}^2$ from this dataset and partition it into 25×20 grid cells. The traces we used in our simulations are one-day long and the time is slotted into 40 instants. We fixed the maximum possible speed to 4 grid cells per time instant after analyzing the maximum speed achieved in the real traces. Finally, for electrosmog measurements, we employ the logged signal strength in dBm from the campaign.

5.2 Artificial data trace

To facilitate the comparison of our results with artificial data to those obtained with real data, we assume an area of the same size (25×20 grid cells) as in the case of the experiments with real data. We assume 20 mobile nodes that move around with the random waypoint mobility model. The maximum speed is assumed to be 4 grid cells per time slot. At each time slot, a mobile node senses an electrosmog measurement (i.e., the signal strength) and submits through privacy protection mechanisms.

We model the electrosmog generation for our simulations with artificial data as follows. The transmission power of base stations ranges from 10 W to 40 W, depending on the network characteristics; we choose 20 W as the base station transmission power in our setting. The frequency of channel is set to 900 MHz as in GSM. We implement free space path loss on this value for each grid cell. There is one base station centered in the area of interest and it covers the whole area in our simulation. We also apply the Rayleigh fast-fading model upon the free-space path loss to simulate a realistic urban area electromagnetic field distribution. Equation 4 shows the free-space path loss PL, where f is frequency in MHz and d is distance in meters. Equation 5 shows the Rayleigh distribution, where R is the power in Watt, and σ is the parameter of the Rayleigh distribution; we use the Rayleigh simulator proposed by Komninakis [13] to apply Rayleigh fading in this setup. Note that, for different frequencies, the characteristics of the electrosmog change, but currently the other existing frequencies in use are greater than 900 MHz, which means that the path loss will be much higher. Therefore, the measurements will yield lower values of electrosmog as the distance increases. In this sense, the loss of generality is negligible in regard to our choice of channel frequency.

$$PL = 20\log(f) + 20\log(d) - 27.55 \tag{4}$$

$$\Pr(R) = \frac{R}{\sigma^2} e^{-\frac{R^2}{2\sigma^2}} \tag{5}$$

5.3 Results

5.3.1 Ineffectiveness of static policies

As explained in Section 3.2, as nodes move and emit sensor data, the spatio-temporal correlation between events can occasionally violate user θ when a static privacy policy is employed. For example, we consider the time series of electrosmog

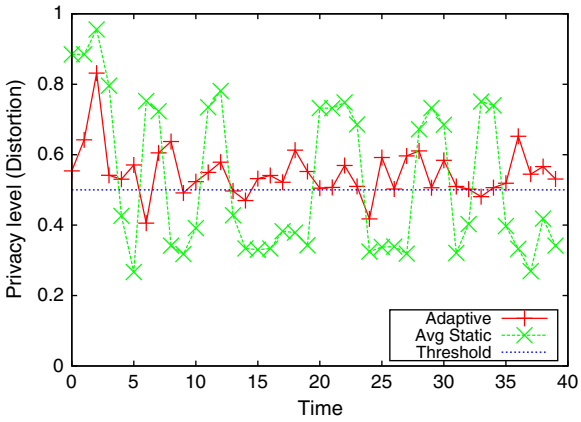

Fig. 6 Privacy levels measured by the LPM over time (40 time instants in this example) for part of one real trajectory in cases of adaptive and average static policies given $\theta = 0.5$

measurements emitted by a certain real user. We assume that $\theta = 0.5$ for this user. As depicted in Fig. 6, a static protection policy, which satisfies θ on the average, often results in significant privacy violations. Our adaptive privacy-protection strategy, on the contrary, dynamically adjusts location obfuscation and hiding behavior to almost always meet θ. Note, at this point, that we measure user privacy in an objective way from the AS point of view by employing the LPM in this figure. The LPM tends to be a bit more conservative than the privacy level estimated locally at the node (although highly correlated as shown later), which does not violate the user privacy requirement by definition, as long as hiding is not chosen (hiding is the last resort for a node to protect privacy. If θ is not met with even the largest λ, then it is possible that hiding is also not enough). Another interesting aspect in Fig. 6 is that our adaptive privacy policy meets θ as minimally as possible, given the employed techniques for location obfuscation.

For all nodes from the real-data traces, the adaptive privacy policy needs to use a number of different obfuscation levels and hiding probabilities in order to meet different θ values, as depicted in Fig. 8. Evidently, due to the fluctuations of the privacy exposure of the users caused by their mobility patterns, a wide spectrum of parameters has to be used for achieving different privacy thresholds. This result is also experimentally verified by artificial data traces. In addition, as shown in Fig. 7,

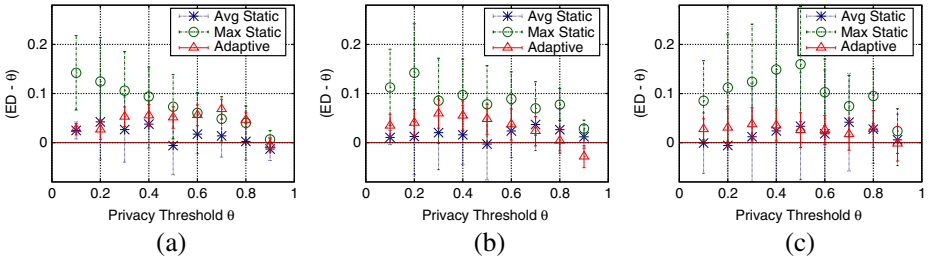

Fig. 7 Level of privacy achieved by adaptive vs. static policies with **a** real and **b** artificial data traces. **c** Level of privacy achieved in the case of background information available to the adversary with real data traces

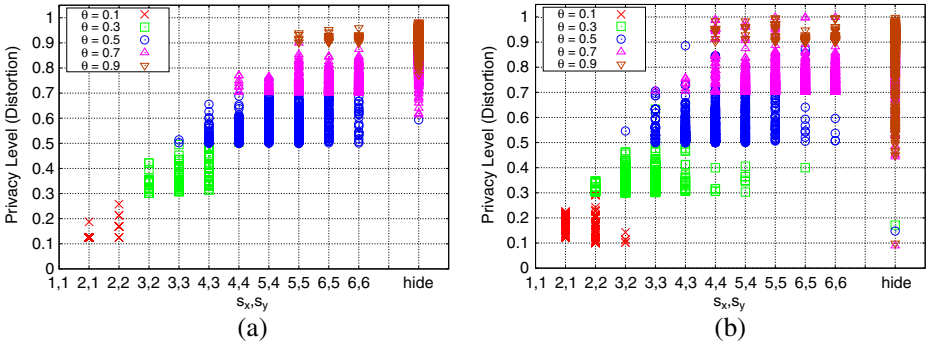

Fig. 8 Parameters ($\lambda = s_x + s_y - 2$) chosen by the adaptive strategy for all (*real*) users over all time steps vs. the local estimations of privacy levels **a** without background information and **b** with background information

the "Avg Static" policy violates thresholds almost half of the time, whereas the adaptive strategy almost always meets them for both real and artificial data. Average values over all users and over all times are plotted in this figure, with a confidence of interval 95 %. Note that meeting $\theta = 0.9$ is very strict and sometimes infeasible with the employed location privacy-enabling techniques, as Figs. 7 and 8 show.

5.3.2 Local estimation of privacy

Figure 9 shows the privacy levels achieved by the adaptive strategy, as estimated locally at the nodes and externally by the LPM for different θ values. These results represent average values and confidence intervals over all nodes. As shown for both real- and artificial-data traces, privacy estimations by our simple approach are highly correlated to the estimations by the LPM (i.e. Pearson correlation > 0.5) for all θ values. Therefore, our simple approach is accurate enough to locally estimate the level of location-privacy of mobile users.

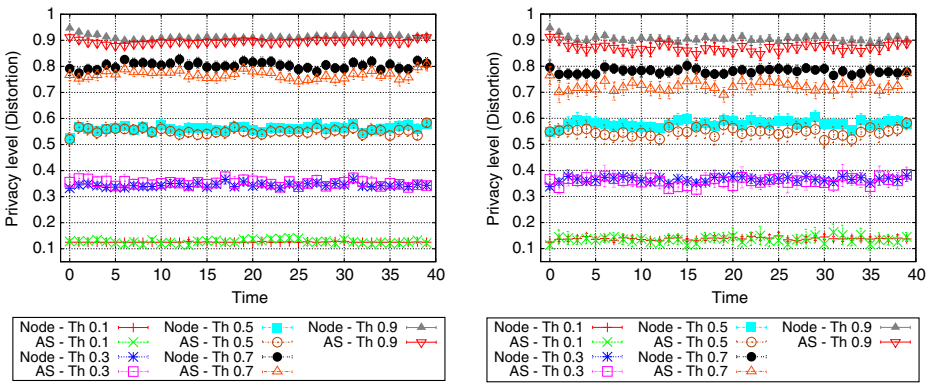

Fig. 9 Comparison of local privacy estimation (*Node*) to measurements by LPM (*AS*) over time with **a** real and **b** artificial data traces

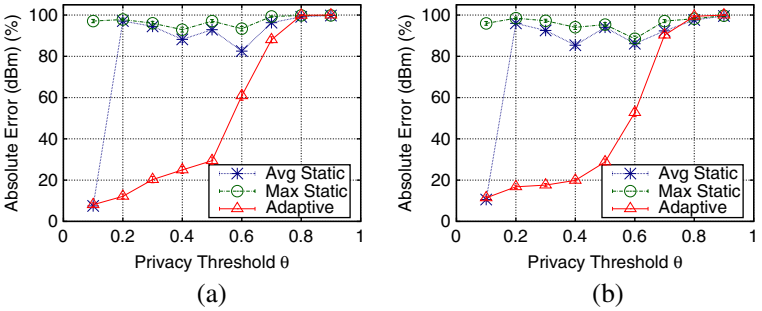

Fig. 10 Percentaged absolute error (dBm) for **a** real and **b** artificial traces

5.3.3 Utility vs. privacy

In Fig. 7, observe that the adaptive protection strategy meets the various privacy thresholds more narrowly, as compared to the "Max Static" policy. As a result, the adaptive strategy is expected to deteriorate the utility of the participatory sensing application less than any static one, while satisfying the privacy requirements of the users. Indeed, the absolute error as a percentage of the data range introduced by the adaptive strategy is lower than the respective errors by the two static policies, as shown in Fig. 10. The results in this and the subsequent figures are average values over all data items from all users and over all times with a confidence interval of 95 %. Note that the results of the real- and the artificial-data traces are similar, despite the significant difference in the mobility behavior of the users.

Moreover, we show the data loss from the two static policies and our adaptive policy in Fig. 11. Notice that the data loss is significantly lower for reasonable privacy requirements of the users, i.e., lower than 0.8 for real data and always for artificial data. Also, the data loss for $\theta \leq 0.6$ is almost insignificant ($\sim$15 % or less) for the adaptive policy, and it is double or more for the two static policies for $\theta \geq 0.2$. This was expected, as static policies need to employ a non-zero $\Pr_h$ throughout the sensing process in order to satisfy even low θ values, as opposed to our adaptive strategy that hides sensor data only when needed.

We measure the deterioration of the area coverage by the ratio of the actual sensed area over the total area reported as sensed. As shown in Fig. 12, this utility

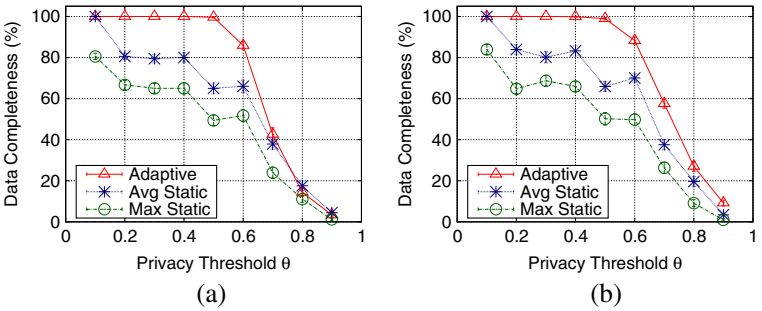

Fig. 11 Data completeness with **a** real and **b** artificial traces

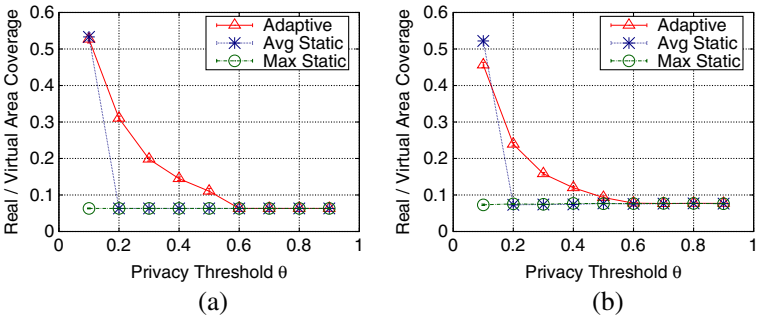

Fig. 12 Area coverage with **a** real and **b** artificial traces

metric deteriorates significantly with high θ values. Although, the area coverage degrades smoothly when the adaptive strategy is employed, as opposed to the static policies. Note that in Figs. 10 and 11 there are small fluctuations; this is due to mobility patterns of the users and also the probabilistic nature of data hiding for static policies.

Overall, Figs. 10, 11, and 12 clearly demonstrate the *trade-off* between utility and privacy. Our results can be employed to derive feasibility conditions on the application-utility and user-privacy requirements for the realization of a participatory sensing application. Our adaptive privacy-protection strategy dominates any static strategies that involve the same location-privacy protection techniques in terms of utility for any user-privacy requirements that render the participatory sensing application feasible.

5.3.4 Adversary background information

Here, we run experiments with the threat model that involves background information at the adversary. The impact of adversarial background information on the chosen privacy parameters by the adaptive privacy-protection strategy is depicted in Fig. 8b. As observed therein, our adaptive strategy performs almost as well as in the case of no background information (cf. Fig. 8a). In fact, the prior background information causes the adversary to be biased and therefore enable the nodes to choose parameters lower than before. For example, for $\theta = 0.3$, some nodes chose strategy 2, 2 in Fig. 8b, as opposed to the case in Fig. 8a. Also, as depicted in Fig. 7, the user loses some privacy due to the adversary background information, but the amount of loss is negligible. Our adaptive approach is still adaptive enough to protect the user privacy in this threat model, despite the existence of background information at the adversary.

Regarding utility, comparing the data completeness in the cases with background information at the adversary (cf. Fig. 13) and without background information (cf. Fig. 11), we observe that they exhibit similar trends. Thus, introducing some background information to the adversary does not cause utility deterioration when the adaptive privacy-protection strategy is employed. However, notice that both static privacy-protection strategies significantly deteriorate utility, since they need to employ larger static parameters than before, in order to meet the various privacy thresholds. Similar findings are observed for the other utility parameters, namely

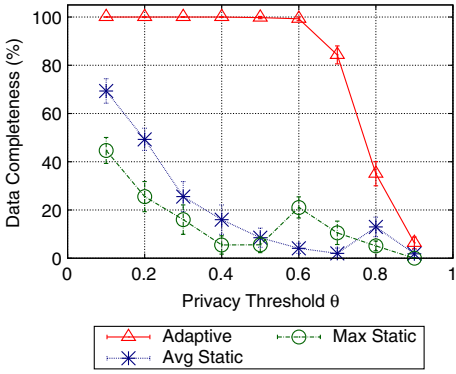

Fig. 13 Data completeness in percentage in the case of adversarial background information

data accuracy and area coverage, when background information is available at the adversary. Therefore, in the presence of background information at the adversary, the employment of an adaptive privacy-protection scheme becomes more important.

5.4 Discussion

In our work so far, we have considered the maximum speed of users, known real identities of sensor-data sources, and additional background information on user mobility history as background knowledge at the adversary. Repetitive trajectories of a user can be another type of background knowledge, which, when accumulated at the adversary, can pose additional privacy threats for the user: People generally move in regular patterns (i.e., on a daily basis), for example from home to work in the morning. If a user generates a different obfuscated trajectory each time he moves along the same trajectory, then an adversary can find the real trajectory in time. However, a simple modification to our adaptive privacy-preserving scheme can eliminate this threat. Specifically, a mobile user has to keep track of her privacy preserving actions along repetitive trajectories (i.e., the obfuscated area for each actual location, per repetitive trajectory) and reuse them in the future. In this way, the privacy leakage due to repetitive trajectories would be limited. Moreover, in order to keep storage overhead bounded, a LRU replacement policy could be employed. The experimental evaluation of this approach is left for future work.

Also, an adversary might employ other background information, such as location semantics, which could be of great help for identifying the real user traces and the user activities. This kind of background information can be modeled as probability distributions over space and time for each user and included in Bayesian inference model employed on the user device.

Another issue that needs addressing is the usability of our approach. We designed our scheme with an automated software tool embedded in localization modules in mind. The average mobile user would not like having to interact with his device every time his location is used by an application. Therefore, our system should be triggered automatically whenever an application asks for the user's location. Such an automation will provide users with a peace of mind in terms of privacy protection. Furthermore, it might not be straightforward for mobile users to interpret the privacy levels, i.e., the expected distortion values, and hence their θ inputs. Here, the privacy

levels need to be conveyed to users in the form of "average confusion from actual location in meters", which can be done using the normalization factor used in the estimation. For example, in our experiments, the normalization is done by 4-hop distance (i.e., the max-speed) which is around 200 meters. In this case, a privacy level of 0.7 yields a confusion of around 140 meters from the actual location.

Last but not least, our system can be extended with location sensitivities, where users input different θ values for their sensitive locations. This would result in an even more dynamic and personalized privacy-protection system. Such an extension can even take into account location semantics, which would enable the batch setting of the user privacy thresholds.

6 Conclusion

In the context of participatory sensing, we have defined a simple, yet effective, adaptive location-privacy protection scheme. Our approach is based on estimating locally in real-time the expected location-privacy level at the user-side, which enables her to adapt her privacy parameters with respect to her mobility, in order to satisfy an *individual* privacy constraint. We have experimentally showed the accuracy of our approach for privacy estimation and the effectiveness of our adaptive privacy-protection strategy, as opposed to static ones. Our adaptive approach achieves more application utility than static policies, and satisfies the individual privacy requirements of the users in case whether background information on the user's mobility history is available to the adversary or not. Furthermore, we have demonstrated the *trade-off* between application utility and user privacy in the context of participatory sensing. As experimentally found, our adaptive privacy-protection scheme is able to maintain high data utility, while satisfying the user privacy requirements. Our results can be used to derive feasibility conditions on the desired application utility and user privacy requirements. The proposed approach is easy to deploy on current mobile devices and supports continuous and sporadic location dissemination by users.

As future work, we plan to consider the existence of application-related background information and topological information that constrains the location obfuscation of the user; we also plan to include additional privacy-enabling techniques in our adaptive policy.

References

1. Canetti R, Feige U, Goldreich O, Naor M (1996) Adaptively secure multi-party computation. In: Proc. of Symposium on Theory of Computing (STOC)
2. Christin D, Reinhardt A, Kanhere SS, Hollick M (2011) A survey on privacy in mobile participatory sensing applications. J Syst Softw 84(11):1928–1946
3. Christin D, Rosskopf C, Hollick M, Martucci LA, Kanhere SS (2012) IncogniSense: an anonymity-preserving reputation framework for participatory sensing applications. In: Proc. of IEEE conference on Pervasive Computing and Communications (PerCom)
4. Das T, Mohan P, Padmanabhan VN, Ramjee R, Sharma A (2010) PRISM: platform for remote sensing using smartphones. In: Proc. of conference on Mobile Systems, Applications, and Services (MobiSys)
5. De Cristofaro E, Soriente C (2011) Short paper: pepsi—privacy-enhanced participatory sensing infrastructure. In: Proc. of 4th ACM conference on Wireless Network Security (WiSec)

6. Diaz C, Seys S, Claessens J, Preneel B (2002) Towards measuring anonymity. In: Proc. of conference on Privacy Enhancing Technologies (PET)
7. Dua A, Bulusu N, Feng WC, Hu W (2009) Towards trustworthy participatory sensing. In: Proc. of USENIX conference on Hot Topics in Security (HotSec)
8. Dwork C (2006) Differential privacy. In: International colloquium on automata, languages and programming. Springer, pp 1–12
9. Gedik BLL (2008) Protecting location privacy with personalized k-anonymity: architecture and algorithms. IEEE Trans Mob Comput 7(1):1–18
10. Groat MM, Edwards B, Horey J, He W, Forrest S (2012) Enhancing privacy in participatory sensing applications with multidimensional data. In: Proc. of IEEE conference on Pervasive Computing and Communications (PerCom)
11. Hu H, Xu J (2009) Non-exposure location anonymity. In: Proc. of IEEE International Conference on Data Engineering (ICDE)
12. Jadliwala M, Freudiger J, Aad I, Hubaux J-P, Niemi V (2011) Privacy-triggered communications in pervasive social networks. In: Proc. of IEEE international symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM)
13. Komninakis C (2003) A fast and accurate Rayleigh fading simulator. In: Proc. of IEEE Global Telecommunications Conference (GLOBECOM)
14. Krause A, Horvitz E, Kansal A, Zhao F (2008) Toward community sensing. In: Proc. of international conference on Information Processing in Sensor Networks (IPSN)
15. Krumm J (2009) A survey of computational location privacy. Pers Ubiquit Comput 13(6):391–399. doi:10.1007/s00779-008-0212-5
16. Lu H, Pan W, Lane ND, Choudhury T, Campbell AT (2009) SoundSense: sound sensing for people-centric applications on mobile phones. In: Proc. of conference on Mobile Systems, Applications, and Services (MobiSys)
17. Minami K, Borisov N (2010) Protecting location privacy against inference attacks. In: Proc. of ACM Workshop on Privacy in the Wlectronic Society (WPES)
18. Mun M, Hao S, Mishra N, Shilton K, Burke J, Estrin D, Hansen M, Govindan R (2010) Personal data vaults: a locus of control for personal data streams. In: Proc. of ACM Conference on Emerging Networking Experiments and Technologies (Co-NEXT)
19. Mun M, Reddy S, Shilton K, Yau N, Burke J, Estrin D, Hansen M, Howard E, West R, Boda P (2009) PEIR, the personal environmental impact report, as a platform for participatory sensing systems research. In: Proc. of conference on Mobile Systems, Applications, and Services (MobiSys)
20. Nokia Research Center: Lausanne data collection campaign. http://research.nokia.com/page/11367. Accessed 7 Apr 2012
21. Pingley A, Yu W, Zhang N, Fu X, Zhao W (2009) CAP: a context-aware privacy protection system for location-based services. In: Proc. of IEEE International Conference on Distributed Computing Systems (ICDCS)
22. Serjantov A, Danezis G (2002) Towards an information theoretic metric for anonymity. In: Proc. of conference on Privacy Enhancing Technologies (PET)
23. Shankar P, Ganapathy V, Iftode L (2009) Privately querying location-based services with SybilQuery. In: Proc. of conference on Ubiquitious Computing (UbiComp)
24. Shokri R, Freudiger J, Jadliwala M, Hubaux J-P (2009) A distortion-based metric for location privacy. In: Proc. of ACM Workshop on Privacy in the Electronic Society (WPES)
25. Shokri R, Theodorakopoulos G, Danezis G, Hubaux J-P, Le Boudec J-Y (2011) Quantifying location privacy: the case of sporadic location exposure. In: Proc. of Privacy Enhancing Technologies Symposium (PETS)
26. Shokri R, Theodorakopoulos G, Le Boudec J-Y, Hubaux J-P (2011) Quantifying location privacy. In: Proc. of IEEE symposium on Security and Privacy (S&P)
27. Vu K, Zheng R, Gao J (2012) Efficient algorithms for K-anonymous location privacy in participatory sensing. In: Proc. of IEEE conference om computer communications (IEEE INFOCOM)
28. Westin AF (1967) Privacy and freedom. Atheneum
29. World Health Organization: Electromagnetic fields and public health. http://www.who.int/mediacentre/factsheets/fs304/en/index.html (2006). Accessed 10 Apr 2012
30. Xiao X, Tao Y (2006) Personalized privacy preservation. In: Proc. of ACM SIGMOD conference on management of data, SIGMOD '06
31. Yan Z, Chakraborty D, Parent C, Spaccapietra S, Aberer K (2011) SeMiTri: a framework for semantic annotation of heterogeneous trajectories. In: Proc. of international conference on Extending Database Technology (EDBT)

Berker Agir is a Ph.D. student at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in the Laboratory for Communications and Applications. He received his B.Sc. degree (2010) in Computer Science and Engineering from Sabanci University, Turkey. His main research interests are in security and privacy in networks, mobile applications and systems, with special focus on location privacy.

Thanasis G. Papaioannou is a postdoctoral fellow at the Distributed Information Systems Laboratory of Ecole Polytechnique Fédérale de Lausanne (EPFL). He received his B.Sc. (1998) and M.Sc. (2000) in Networks and in Parallel/Distributed Systems from the Department of Computer Science, University of Crete, Greece, and his Ph.D. (2007) from the Department of Computer Science, Athens University of Economics and Business (AUEB). From spring 2007 to spring 2008, he was a Visiting Professor in the Department of Computer Science of AUEB, teaching (i) Distributed Systems and (ii) Networks—Network Security. His research interests are in data stream processing, mechanism design for online environments, cloud resource management, privacy, trust and reputation, sensor networks, and QoS.

Rammohan Narendula obtained his PhD at Ecole Polytechnique Fédérale de Lausanne (EPFL) in 2013. He obtained his Masters (M.S—by Research) from High Performance Computing and Networking Laboratory (HPCN) Indian Institute of Technology Madras (IITM), India. His main research activities are on privacy preserving indexing of content and decentralized privacy-aware social networking solutions.

Karl Aberer is a professor at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, since 2000. His particular interest is on decentralized system architectures, self-organization mechanisms, and emergent structures in information systems. He also serves as the director of the Swiss National Research Center for Mobile Information and Communication Systems (NCCR-MICS). Aberer has a PhD in mathematics from Eidgenössische Technische Hochschule in Zürich. He is a member of the IEEE and the ACM.

Jean-Pierre Hubaux is a professor at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. His current research activity is focused on privacy protection mechanisms, notably in pervasive communication systems and online social networks. In particular, he is involved in projects related to location and data privacy, privacy in pervasive communications, and also to privacy and security of online advertising. He has recently started a research activity in genomic privacy, in close collaboration with geneticists. He is one of the seven commissioners of the Swiss FCC. He is a Fellow of both ACM and IEEE.