Pervasive and Mobile Computing 20 (2015) 1–12

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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

A privacy mechanism for mobile-based urban traffic monitoring

Chi Wang[a], Hua Liu[b], Kwame-Lante Wright[b,*], Bhaskar Krishnamachari[b], Murali Annavaram[b]

[a] Microsoft Research, Redmond, WA, USA
[b] University of Southern California, Los Angeles, CA, USA

ARTICLE INFO

Article history:
Received 25 July 2013
Received in revised form 25 October 2014
Accepted 20 December 2014
Available online 27 December 2014

Keywords:
Mobile computing
Traffic monitoring
Privacy

ABSTRACT

In mobile-based traffic monitoring applications, each user provides real-time updates on their location and speed while driving. This data is collected by a centralized server and aggregated to provide participants with current traffic conditions. Successful participation in traffic monitoring applications utilizing participatory sensing depends on two factors: the information utility of the estimated traffic condition, and the amount of private information (position and speed) each participant reveals to the server. We assume each user prefers to reveal as little private information as possible, but if everyone withholds information, the quality of traffic estimation will deteriorate. In this paper, we model these opposing requirements by considering each user to have a utility function that combines the benefit of high quality traffic estimates and the cost of privacy loss. Using a novel Markovian model, we mathematically derive a policy that takes into account the mean, variance and correlation of traffic on a given stretch of road and yields the optimal granularity of information revelation to maximize user utility. We validate the effectiveness of this policy through real-world empirical traces collected during the Mobile Century experiment in Northern California. The validation shows that the derived policy yields utilities that are very close to what could be obtained by an oracle scheme with full knowledge of the ground truth.

1. Introduction

In existing sensor networks, power-constrained sensors are deployed in the targeted area and data is collected until the sensors run out of battery power or the collection time window expires. There are several disadvantages in such traditional sensor networks. First, the size of the sensor network is usually small. Second, most sensors are power constrained and hence may require replacing or recharging of their batteries; either of these tasks is intrusive to the sensing process and can sometimes be time consuming if the sensing environment is not easily accessible.

1.1. Motivation for participatory sensing

In order to overcome the shortcomings of traditional sensor networks, researchers have proposed projects such as *MetroSense* [1] and *Participatory Sensing* [2]. This new generation of sensing projects is based on the concept of "people-centric

[*] Corresponding author.
E-mail addresses: chiw@microsoft.com (C. Wang), hual@usc.edu (H. Liu), kwamelaw@usc.edu (K.-L. Wright), bkrishna@usc.edu (B. Krishnamachari), annavara@usc.edu (M. Annavaram).

http://dx.doi.org/10.1016/j.pmcj.2014.12.007
1574-1192/© 2015 Elsevier B.V. All rights reserved.

sensing'' at a large scale (e.g., campus, town, or metropolis). People are central to the sensing experience and represent the key architectural component in this new paradigm. In this category of sensing, human-carried sensors are brought into the environment that we are interested in sensing. The key element of such sensing is that people might be sensing their surroundings as they go about their daily activities without even making any explicit effort to sense. Mobile phones have become a key enabler for such *passive sensing*. Mobile phones are typically equipped with several integrated sensors, such as GPS, microphones, Bluetooth, and Wi-Fi. These features make the phones attractive for such participatory sensing projects.

In this paper, we focus on one particular participatory sensing application, namely urban traffic monitoring. In this traffic monitoring application, sensors, namely GPS, are integrated either into a mobile phone or into a user's vehicle. These sensing systems have the potential to radically improve the accuracy and timeliness of traffic information. In this application, several users driving on various road segments can use their GPS-enabled sensors to accurately determine their speed and position information. The measured information is then transmitted to a backend aggregation server. The aggregator collects segmented traffic reports from individual users and combines the reports to obtain complete traffic conditions of an entire road stretch. The global traffic information is in turn used by the aggregator to provide real-time traffic and travel time estimates to all the users in the system. Traffic sensing is an important application class where the accuracy of traffic estimation improves with the increasing number of participants.

1.2. Importance of privacy

While the motivation for traffic sensing using mobile phones is clear, the approach described above where the user reports speed and position information to the aggregator, potentially compromises the participant's privacy. In traditional sensor networks, since a sensor node is not associated with a particular individual the need for privacy is relatively low. However, in participatory sensing, when a mobile phone is being used as a sensor, the sensing device and the participant are closely tied together. A mobile phone identifies the sensor uniquely with a participant's identity. The data sensed is not only indicative of the participant's surroundings, but also reveals the participant's location and speed. Hence, we have to take the device holder's (application subscriber) privacy into account when designing the system. If the accurate location/speed information is intercepted by malicious attackers, the attackers can reveal the phone's identity by investigating the MAC layer packet headers. Once the identity of the device holder is revealed with *precise* location and speed information, the participant is exposed to the attacker. Imagine the day when an unwary traffic sensing participant gets a speeding ticket as an SMS message!

The goal of this paper is to study the privacy risks in traffic sensing. In order to protect users' privacy, we derived a utility based application method, which lets the users update the system with ''just enough'' information to the backend server, trading some data accuracy for improved user privacy. In this research, we consider the *location granularity* as a mechanism to obfuscate the users' precise location information. For instance, using a coarse location granularity the user can inform the aggregator that he/she is currently driving *somewhere* between two exits on a freeway without disclosing the precise location. Privacy is better protected but the system can still maintain reasonable service quality.

The paper is organized as follows. Section 2 describes the traffic monitoring application. Sections 3 and 4 depict our novel mathematical formulation of the problem, including the Markov-based road condition model and utility modeling. In Section 5, we propose a practical policy that suggests a near-optimal decision on maximizing a user's utility. Our experimental methodology and results are presented in Sections 6 and 7. Finally we present some related works in Section 8 and conclude our work in Section 9.

2. Application description

We believe that mobile based urban traffic monitoring systems will help relieve traffic conditions in the future and help application users estimate traffic conditions on the road with privacy taken into consideration.

In the simplest version of this application, we envision the use of virtual trip lines (VTLs) [3] to help coordinate data gathering. Virtual trip lines are GPS coordinates of a line that is *virtually* drawn on top of any road segment by the traffic application administrator. Mobile devices monitor their location using GPS and when they cross a VTL the device sends a raw update to a backend server with accurate position (VTL id) and speed information. The backend server aggregates the information obtained from multiple devices and uses it to estimate the current traffic conditions and provide accurate traffic and drive time estimates back to the mobile devices in real time. This information can then be used to alert the vehicle drivers about possible traffic congestions and even suggest alternate routes.

However, for the users on the road, the major privacy concerns are focused on users' exact location and speed. If the user's update information is overheard, or maliciously detected by eavesdroppers, the user's privacy is compromised by revealing the exact location and speed information. Even though the application may not need the user's identity when collecting the traffic condition updates, the MAC layer of the mobile devices implicitly reveals a user's identity by the MAC address. In this case, the simplest version for the traffic monitoring application does not preserve the user's privacy. We need to modify the application to provide better privacy protection. Therefore, we propose a utility based privacy preservation model for the traffic monitoring application. This modified application considers the tradeoff between the users' desire to protect privacy, and their requirement to have accuracy on traffic estimation error. It also provides a policy to optimize this tradeoff. That is, the improved traffic monitoring application allows the users to contribute to the system with "just enough" information to preserve privacy and meanwhile, make the use of the traffic estimation with adequate precision.

This modified traffic monitoring application (which is the focus of the remaining parts of this paper) consists of four message exchanges:

1. Application subscribers request an estimation of mean, standard deviation in speeds and road correlation factor for a certain stretch of road in a certain time interval.[1] For example, a user can send queries to a backend server by asking "what are the corresponding parameters for highway I-10 exit 31 to exit 33 at 4:00pm–4:30pm, July 4th?".
2. The backend server returns those parameter values (also referred to as model statistics in this paper) possibly based on the historical data, as well as an estimated number of users.
3. Users send out the optimized local information updates to the backend server, based on their utility-based privacy policy. This information includes an implicit spatial granularity and the user's current vehicle speed with a timestamp.
4. The backend server returns current traffic conditions on the road stretch to the application subscribers. The feedback information is real-time averaged traffic flow speed on the road with corresponding granularity, e.g., "the speed between VTL 34 and VTL 39 is 50 mph".

In the following section, we will focus on how the traffic is modeled, how to obtain the model statistics/parameters at the backend server, how to use the model to calculate the utility for each user, and how to calculate the optimized updates.

3. The Markov road model

In this paper we propose a Markov-based road model to measure the traffic estimation precision with a minimal number of parameters. The main purpose of this Markov-based traffic model is to characterize the impact of granularity on traffic estimation accuracy, so that we can measure the system's quality of service as a function of granularity. In this section, we present this novel model after describing the preliminaries, necessary assumptions, and notations used in the paper.

3.1. Preliminaries

Before we introduce the Markov-based road model, we first formally define the concept of spatial granularity which is an important parameter for our future analysis. Spatial granularity here is defined as an integer, each unit represents a length of road segment between two adjacent VTLs. In other words, each VTL interval implies a spatial granularity. For example, "between the 105th and 110th VTLs" implies granularity 5.

3.2. Assumptions and notations

We assume a complete road stretch as a line with length l. n VTLs are set along the road from one end to the other. The road is divided into sections evenly by the VTLs. The sections are continuous and non-overlapping road segments. Each section contains a length of $\frac{l}{n}$ where n is the number of VTLs.

Assume that the average vehicle speed in each section at a certain time slot is a random variable, denoted by $X_1, X_2, \dots, X_n$ from Section 1 the nth section. Considering the fact that the traffic flow at certain section on the road is directly affected by the traffic condition ahead of this section, we assume the speed X_i at the ith section is correlated with the speed X_{i+1} at the $(i+1)$th section. To model this correlation on the traffic flow, we employ a correlation factor α ($\alpha \in (0, 1]$). Road correlation factor α reflects the impact of the average speed in section i to the average traffic flow speed in section $i+1$. We will discuss how α is related to the traffic flow in the following section.

3.3. The Markov-based road model

Consider a road which is separated into n sections. Recall that X_i denotes the average traffic speed in the ith section, a first-order Markov-based equation is proposed as follows:

$$\begin{cases} X_n = x_n, \\ X_i = \alpha X_{i+1} + (1-\alpha)x_i, \quad i = 1, 2, \dots, n-1 \end{cases} \tag{1}$$

where x_i is a random variable corresponding to the speed fluctuation related to the vehicle location. $x_1, x_2, \dots, x_n$ are independent, with corresponding mean μ_i and variance σ_i^2. The distribution of x_i shows the exterior, isolated road conditions at every single location. As mentioned before, α is the road correlation factor between the speed at two neighboring sections. Specifically, the larger α is, the smoother the traffic flow is on this road stretch.

Note that the parameter α is dependent on the time of the day and the shape of the road. For instance, during peak hours X_i is likely to be dependent on fluctuations within its own segment (x_i) and is less dependent on the X_{i+1}. Hence, the optimal spatial granularity computed by our model (discussed later) is valid for a given road segment and for a given time interval.

[1] These parameters are used in the Markov model we proposed in Section 3 to measure the impact of traffic estimation accuracy for the application subscribers. We will discuss these parameters in detail in later sections.

3.4. Service quality measurement

As mentioned before, this application's service is provided by a centralized server. The quality of the service is measured by the accuracy of the estimated average traffic speed. Mathematically, the quality of the service is measured by a statistical variable: expected Mean Square Error (MSE).

Suppose the road stretch is divided into n sections. Given the average speed X_i at each section i at certain time, the Mean Square Error(MSE) of estimation can be calculated as:

$$\mathrm{MSE} = \frac{1}{g\left\lfloor \frac{n}{g} \right\rfloor} \sum_{i=1}^{g\left\lfloor \frac{n}{g} \right\rfloor} \left(X_i - Y_{\left\lceil \frac{i}{g} \right\rceil} \right)^2 \tag{2}$$

where $Y_k = \frac{1}{g} \sum_{i=(k-1)\cdot g+1}^{k\cdot g} X_i$ is the estimated speed collected by the service for section $(k-1)\cdot g + 1$ through $k \cdot g$. For instance, if spatial granularity is 5 (i.e. $g = 5$), Y_k measures the average speed over 5 consecutive road sections. $X_i - Y_k$ computes the error in estimation due to the reduced precision of location information in each section.

However, this only calculates the estimated error during a certain time interval for a given granularity g. To obtain the expectation for estimated error, we have to aggregate the error during each period and calculate the average.

Assume that the period of sampling is a constant t, the estimated error on speed will accumulate to $t\sqrt{\mathrm{MSE}}$ in each sampling time period.

We can derive the expectation of the estimated MSE from Eqs. (2) and (1). Given a fixed time interval, let μ_i denote the expectation of x_i and σ_i denote the standard deviation of x_i. Consider the following three equations:

$$E(x_i^2) = \mu_i^2 + \sigma_i^2$$
$$E(x_i) = \mu_i$$
$$E\left(\sum \left(x_i - \frac{\sum x_i}{g}\right)^2\right) = E\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{g}\right).$$

Let $E_k(g)$ represent the average MSE of the estimated speed in the road segment that contains section $(k-1)g+1$ to kg. We have

$$\begin{aligned} E_k(g) = \frac{1}{g} \Bigg\{ & (\mu_{kg}^2 + \sigma_{kg}^2)\left(\frac{1-\alpha^{2g}}{1-\alpha^2} - \frac{(1-\alpha^g)^2}{g(1-\alpha)^2}\right) + \sum_{i=1}^{g-1} (\mu_{(k-1)g+i}^2 + \sigma_{(k-1)g+i}^2)\left[\frac{1-\alpha}{1+\alpha}(1-\alpha^{2i}) - \frac{(1-\alpha^i)^2}{g}\right] \\ & + 2\sum_{1\le j\le g-1} \mu_{kg}\mu_{(k-1)g+j}\left[\frac{\alpha^{g-j}-\alpha^{g+j}}{1+\alpha} - \frac{(1-\alpha^g)(1-\alpha^j)}{g(1-\alpha)}\right] \\ & + 2\sum_{1\le j<i\le g-1} \mu_{(k-1)g+i}\mu_{(k-1)g+j}\left[\frac{1-\alpha}{1+\alpha}(\alpha^{i-j}-\alpha^{i+j}) - \frac{(1-\alpha^i)(1-\alpha^j)}{g}\right]\Bigg\}. \end{aligned} \tag{3}$$

The averaged MSE for the whole road stretch can be computed as $E(g) = \frac{1}{\lfloor \frac{n}{g} \rfloor} \sum_{k=1}^{\lfloor \frac{n}{g} \rfloor} E_k(g)$. Let $e(g)$ denote the estimated traffic speed error aggregated in the given time interval t. $e(g)$ can be calculated by:

$$e(g) = \sqrt{E(g)}t. \tag{4}$$

We have to emphasize that the MSE measurement using the Markov road model creates a bridge between service quality and the information granularity. It quantifies the impact of information granularity on traffic estimation accuracy and therefore makes user utility modeling of the tradeoff between service quality and privacy protection feasible. We will now show in Section 4 how we use $e(g)$ in the user's utility formulation.

4. Utility formulation

As we have mentioned in Section 2, we have to take people's privacy into account in this application to ensure broad user participation. This is a primary requirement for our approach to work in practice.

The accuracy of traffic estimation depends on the accuracy of the following three fundamental factors: location, time, and speed. The accuracy of traffic estimation is proportional to the information accuracy along these three dimensions. On the other hand, on intuitive grounds it is easy to see that a user's privacy is inversely proportional to the information accuracy. A user's privacy can be compromised if the combination of these three factors can be used to correlate a traffic update with a given user. In this work, we focus on the vehicle location dimension and assume trustfulness and safety of the other two dimensions: time and speed. We briefly discuss in the conclusions section how the frequency of sampling could affect privacy as well.

In our application, we give users the option to choose the granularity at which their locations are revealed, in order to increase the difficulty of malicious tracking. However, this needs to be carefully balanced with the system-wide goal of obtaining reliable traffic estimates.

The utility function of each application user has two parts: privacy protected (denoted as a function $p(g)$) and the expected feedback estimation error (derived in previous section as $e(g)$). We linearly combine these two parts with a weight factor β.

Privacy protection function $p(g)$ is modeled as a function of granularity index g. Intuitively, the users updated information can be explained as "I am driving between an interval with this speed". Increasing the length of the interval will generalize the location, decreasing the probability that a user's exact location can be determined. This means that $p(g)$ is a function that increases as g increases. There are different ways to model the privacy in this application. In our work, we use the following function to model the privacy protected[2]:

$$p(g) = l\frac{g-1}{g}. \tag{5}$$

The physical meaning of this equation can be explained as follows. Suppose that the car we are discussing is now somewhere along a road and the road's total length is l. If the user's uploaded information does not narrow down the search scope, the driver owns a private space with the length of l. That is, the probability of revealing the exact location of the user is uniformly distributed in this piece of road. On the other hand, if $g = 1$, then the exact location of the driver is revealed to the system. If the information is given with granularity g, according to the previous section, the malicious tracker has the chance $\frac{1}{g}$ to detect the exact location. That is, the driver's private space is reduced by $\frac{1}{g}$ compared to the best privacy reservation that can be achieved in this problem. Hence, the private space, i.e., the expectation of the length for which the vehicle runs without being detected, becomes $l(1 - \frac{1}{g})$.

The other part of the utility function is the estimation accuracy. As we described in the previous section, the accuracy loss is modeled as the expected speed estimate error. Mathematically, each single user's utility function $u(g)$ is defined as:

$$u(g) = -\beta \cdot e(g) + p(g). \tag{6}$$

An individual user's objective is to maximize his/her utility. We need to point out that the particular $p(g)$ we considered in this paper is a concave function that captures the significant impact of g's change when the granularity is already refined. For example, if g drops from 2 down to 1, it causes more significant loss of privacy than g dropping from 20 down to 19.

5. The privacy policy

Each user's decision now is an optimization problem that maximizes their own utility function. The only parameter that the users need to pick is spatial granularity g. The other parameters, including road correlation factor α, mean traffic flow speed μ_i and traffic flow speed standard deviation σ_i for a certain road stretch are obtained by querying a backend server. Since the road stretch we consider has a finite length, the number of VTLs on this road is also a finite number. Therefore, the number of feasible granularities, which can only be taken from integers, is bound by the number of VTLs. A straightforward way to compute the optimal granularity can be done by direct enumeration in the finite search space. For each tuple of (α, μ, σ), the optimal location update granularity is calculated.[3] We will validate the privacy policy in Section 7.

5.1. Impact of parameters on the optimal granularity

In this subsection, we analyze the impact of statistic parameters on the optimal granularity when applying a simplified model, where the random variables x_i are independent and identically distributed (i.i.d). With the i.i.d. assumption, the statistical parameters μ and σ are equal. The corresponding traffic satisfies the following property:

The expectation of the speed at every location is equal.

Theoretically, this traffic pattern appears when, for instance, the traffic moves smoothly during a segment in the middle of a freeway. However, in practice, there are small fluctuations of traffic and the deviation is within a range that can be estimated as having the same μ and σ. The validation in Section 7 will show that although the calculated optimized granularity is not exactly the same as the actual optimized granularity, the performance of this simplified model is very close to the real optimization point in terms of user utility.

Fig. 1 illustrates the impact of the analytical optimized granularity g^* in the simplified case for different α values when β changes. In the remaining part of this section, we discuss the parameters one by one to show their impacts on the analytical g^*.

[2] We are aware that privacy can be modeled in multiple rational ways in this traffic monitor application. We pick one example with meaningful physical implication here. The utility modeling methodology can be applied to diverse privacy functions. This methodology can also be applied to other utility functions that combine the estimation accuracy and privacy in an arbitrary non-linear fashion.

[3] The optimal granularity calculation could be performed on the client side but it would be far more advantageous for it to be performed by a backend server. Not only would the energy consumption of client devices be reduced, but the granularity only has to be calculated once for each road stretch and can be shared with all interested clients.

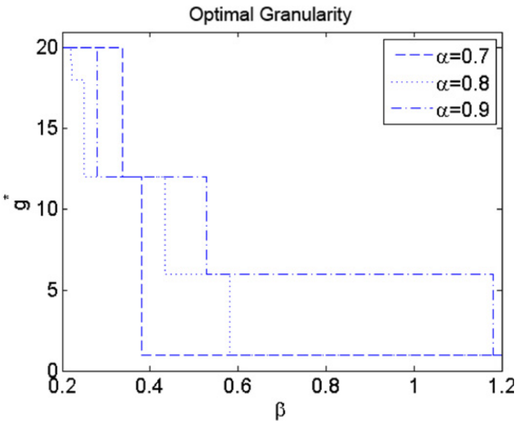


Fig. 1. Optimal granularity versus β when $\mu = 50$, $\sigma = 10$.

- General observations
  g^* decreases convexly with increasing β, regardless of α. The fact that g^* decreases with increasing β reflects the tradeoff between estimation accuracy and privacy protection. When users care more about the accuracy of the information they receive (service-concerned users), they will choose finer granularity to increase the overall utility. Intuitively, the more a user weights privacy, the more hesitant the user is to provide accurate information.
  A notable observation from Fig. 1 is that g^* is a convex decreasing function. Specifically, this means when β is relatively small, g^* drops more quickly than when β is larger. This phenomenon is consistent with our choice of privacy function (as discussed in Section 4). When the information granularity is coarse enough (i.e., g is sufficiently large), further increasing the granularity will not have a significant impact on the utility.
  Another implication is that only users who are extremely concerned with their privacy ($\beta < 0.4$ or so) will give very ambiguous information ($g > 10$). For most β values, g is kept to below 6. In practice, this observation implies that by using the proposed utility function, the service quality is maintained at an appropriate level.
- α's impact
  One observation from the plot is the corresponding β value strictly increases at the points where g^* drops to 1 (a user's most accurate information is revealed at this point) when α increases. We omit the proof here due to page constraints. Increasing α implies a smoother road traffic condition. When the traffic flow is smooth, privacy weighs more than traffic estimation concerns unless the user is very picky about the traffic estimation accuracy (where a larger β is required).
- μ's impact
  Fig. 1 is plotted using $\mu = 50$, $\sigma = 10$. μ almost has no impact on the analytical optimized granularity as long as $\sigma \ll \mu$. This fact implies that what matters for the calculation of the optimal granularity are the changes in speed on the road stretch, not their absolute value.
- σ's impact
  When σ increases, the same β yields a larger analytical optimal g^*. This fact is also intuitive. Note that σ reflects the fluctuation of x_i. When σ is large, the road condition changes more significantly than when σ is small. That means the users have to sacrifice privacy to receive better traffic estimations.

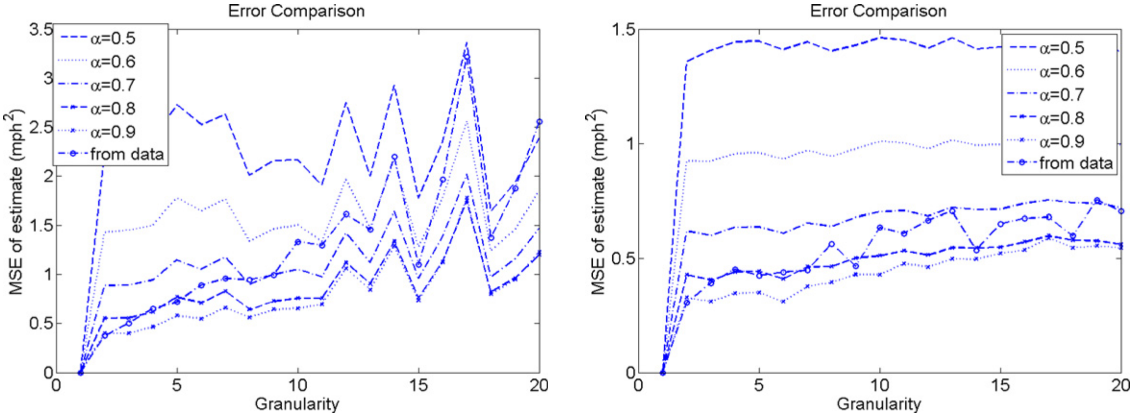
We will show in the next section that the simplified model can also approximate the optimal granularity for the complex model, where x_is are not assumed to be identical.

6. Experimental methodology

In this section, we first discuss how the real experiment was performed and then describe the structure and characteristics of the trace data set.

6.1. The mobile century experiment

A large scale experiment to measure the effectiveness of VTLs was conducted jointly by the Nokia Research Center and the University of California, Berkeley [4]. In this experiment 100 cars equipped with GPS-enabled Nokia N95 phones were driven by volunteer drivers for 8 h along a carefully constructed path in the San Francisco bay area. The location of this experiment was specifically selected because it featured both free flowing traffic, and congested, stop and go traffic. As the vehicles drove the travel route, the mobile devices continuously monitored their location using GPS. As the devices crossed a VTL they sent a traffic update to a backend database server. The traffic update is a tuple containing the VTL number, time, and speed.



(a) Northbound MSE curves.

(b) Southbound MSE curves.

Fig. 2. Comparison of MSE curves.

6.2. *Characterization of the trace data set*

This traffic data set contains diversity in terms of speed. It consists of fluid traffic flows (>50 mph smooth flows), congested flows, and stop and go traffic (traffic lights). There are 45 VTLs evenly placed to record the speed measurements from the vehicles. In our experiment, the feasible granularity set contains integers from 1 through 20.

The data set is separated into two independent cases: one of them includes data for all vehicles moving from south to north (a.k.a. northbound data) and the other contains data for vehicles moving from north to south (a.k.a. southbound data).

It is difficult to determine α directly from the distribution of speed (X_is) because the distribution of x_i is unknown. In order to estimate the road correlation factor α, we use the best curve fitting method on the MSE curve. Fig. 2(a) and (b) show the MSE curve fitting for the northbound data and the southbound data, respectively.

For the northbound data, curve error of estimation calculated from the data traverses the theoretical $E_\alpha(g)$ from $\alpha = 0.5$ through $\alpha = 0.9$, when g decreases. This implies that the analytical result fits with the data with different α for varying granularity. This result suggests that for the samples we take from northbound data, the first order Markovian model with a single correlation factor α may be too simple to characterize the traffic in that region. However, the theoretical curve matches almost perfectly to the empirical curve if we focus the granularity within a smaller range. For example, if we focus on $g = 10$ through $g = 16$, $\alpha = 0.6$ is a satisfactory fitting parameter.

For the southbound data, we find that the error curve is basically bounded by the $\alpha = 0.7$ and $\alpha = 0.8$ theoretical MSE curves. A single value α may still be insufficient to perfectly match the traffic pattern, however, the range of reasonable α is narrowed down to 0.7–0.8. We also noticed that α for southbound samples is larger than the northbound samples, which implies that for southbound samples, the correlation between adjacent locations is larger than in the northbound samples. The higher road correlation factor indicates a smoother distribution of speed.

7. Results

In this section, we focus on presenting how well the policy does on the real traces, compared to the real empirical optimum in terms of individual user's utility gain. According to the experiment settings, the empirical data set is split into two parts: northbound data and southbound data. We validate the analytical optimal policy in both data sets.

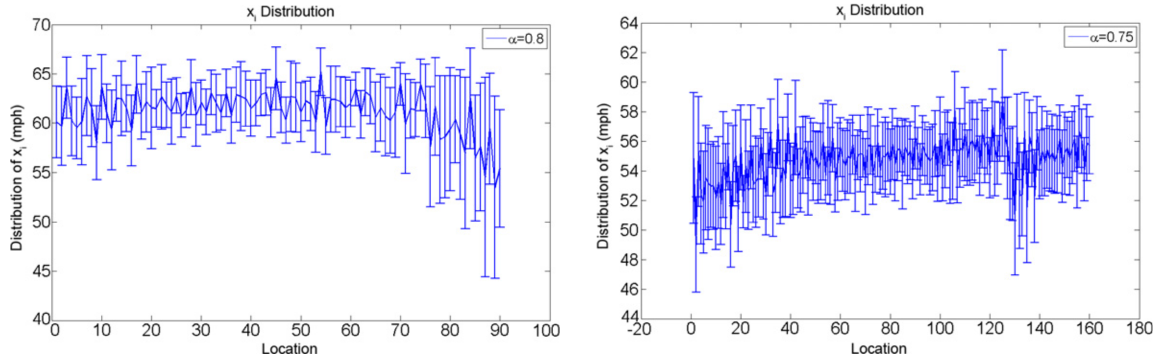
7.1. *Near optimal utility validation*

We will illustrate that although the optimal granularity calculated by our model does not always match the real optimal point in the trace data, its performance is close to the optimal point in terms of an individual user's utility gain.

7.1.1. *Northbound data set*

In our experiment, we let β vary from 0.2 to 5.0 with step size 0.2. For each given β, the utility at the empirical optimum and analytical optimal granularity g is compared in a pair of plots. In all the plots, the continuous curve represents the corresponding utility gained by individual users when granularity g changes. The vertical line is the analytical optimal granularity suggested by the traffic monitor application. The intersection of the red vertical line and the continuous curve is the actual utility gained for the application user if he takes the suggestion.

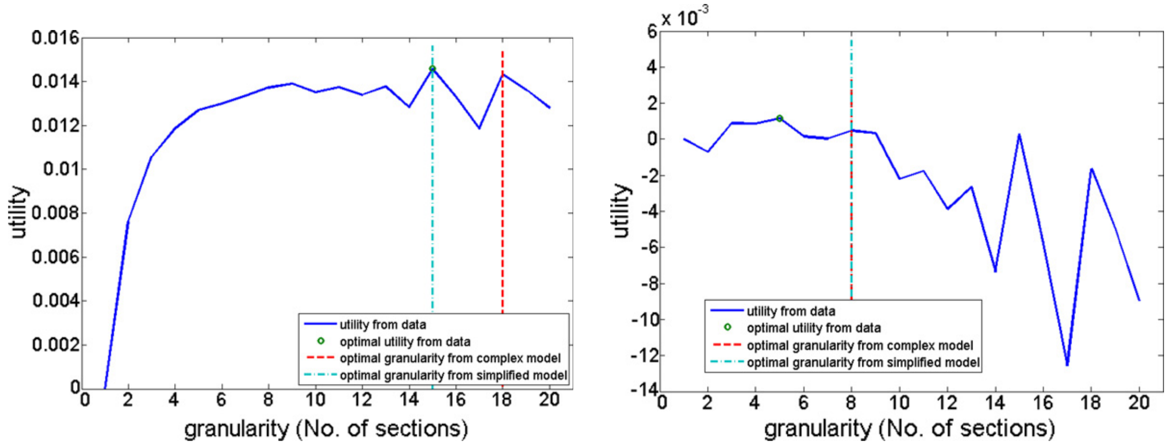
We have two models which are used to distinguish the method to compute statistical parameter μ (mean of vehicle speeds) and σ (standard deviation of vehicle speeds). In the "complex model", μs and σs are calculated from the real data



(a) Northbound distribution.

(b) Southbound distribution.

Fig. 3. x_i distributions.



(a) $\beta = 0.2$.

(b) $\beta = 1$.

Fig. 4. Utility comparison for northbound data set.

set for each section of the road. In the "simplified model", as discussed in Section 5.1, the i.i.d. assumption means identical μs and σs are applied for the whole segment.

It is worth noting again that the number of parameters (μ, σ, α) needed in a given stretch of road can vary, with more parameters needed for roads with more complex traffic. In our experiments, we found that 3 sets of μ, σ and α values are needed for the northbound traffic, while just 1 set of parameters gives good performance for the southbound traffic.

In Fig. 4, each single user's payoff is compared in three different cases with the actual utility curve: (a) the circle point is the optimal utility a user could get if he/she has oracle knowledge; (b) the dashed red vertical line is the utility obtained by applying the "complex model"; (c) the dotted blue vertical line is the user's utility computed with the "simplified model".

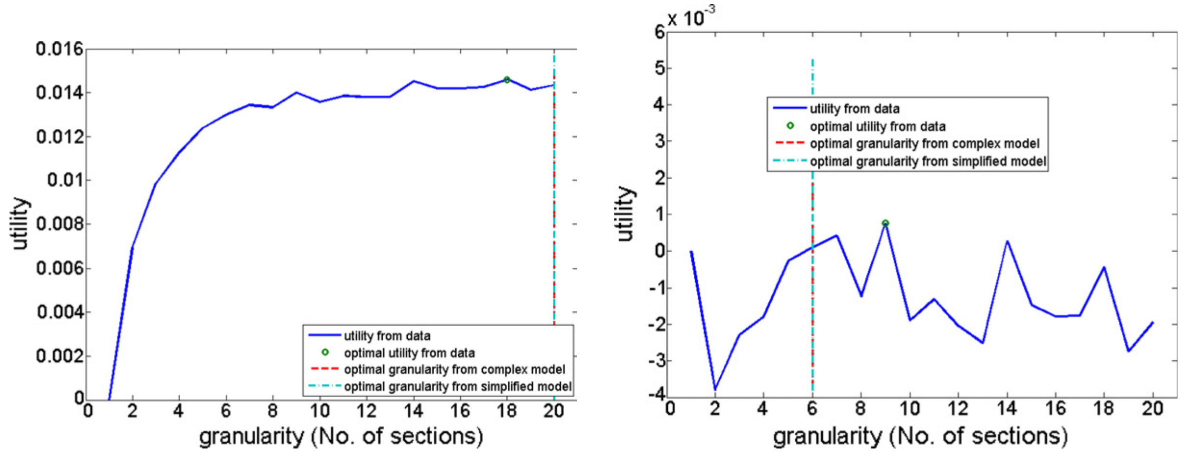
The plot shows that over all, the derived policy yields utilities that are very close to what could be obtained with an oracle scheme that has full knowledge of the ground truth in all cases. There also are some cases where the analytical granularity exactly matches the empirical optimal utility with oracle scheme. Another notable observation is that the "simplified model" performs no worse than the "complex model" in terms of utility gain for individual users. Notice that the "simplified model" is more practical than the "complex model" in real time traffic monitoring in the sense that the parameter values are estimated values.

In this northbound data set performance validation experiment, the accurate values we used in the "complex model" are illustrated in Fig. 3. The road correlation factor $\alpha = 0.8$ is chosen by the best curve fitting method for all the experiments.

We want to highlight this significant result from the utility validation experiment. *The results show that the utility-based privacy policy model we have developed performs near-optimally in all the cases.* Notice that true optimality can only be achieved by an oracle scheme that has accurate full knowledge of the ground truth. This is very impractical. Therefore, it is remarkable that our derived policy can perform very close to the empirical optimal.

7.1.2. Southbound data set

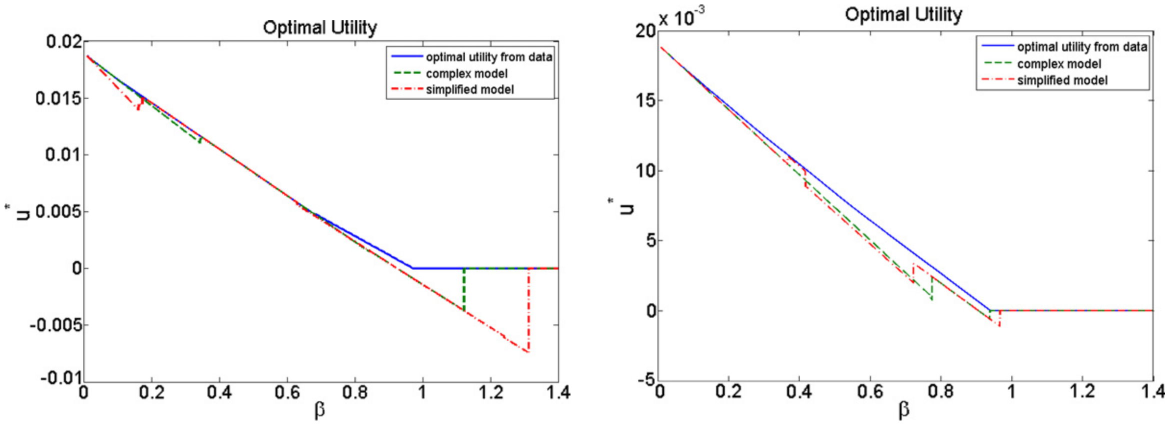
Similar performance validation experiments have been done to the southbound data set as well. For the southbound data set, the x_i's distribution used in the "complex model" is illustrated in Fig. 3(b). As we have mentioned before, one set



(a) $\beta = 0.2$.

(b) $\beta = 1$.

Fig. 5. Utility comparison for southbound data set.



(a) Northbound utility.

(b) Southbound utility.

Fig. 6. Utility performance for prediction with complex and simplified distributions.

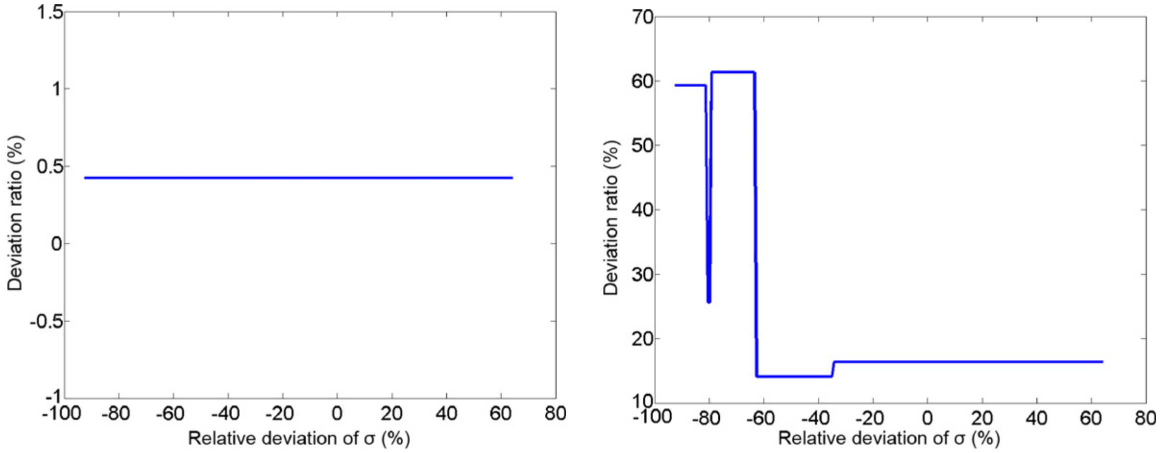
of parameters for the whole road stretch is good enough for the southbound data set. The α value is set to 0.75 for both "complex" and "simplified" models. Fig. 5 illustrates the utility versus different granularities obtained from the empirical southbound data set. The circle points are the maximum possible utility gained with oracle knowledge of the future. The vertical lines are the analytical optimal granularity for the complex and simplified models.

We also condense the comparison for different β in one plot each for the northbound and southbound data sets. In each plot, the dotted line shows the optimal utility that can be achieved. The solid lines are the utility that can be achieved under the theoretical optimal granularity, with assumptions of complex and simplified distributions respectively. For the northbound data set, shown in Fig. 6(a), α is chosen as 0.8 and for the southbound data set, shown in Fig. 6(b), $\alpha = 0.75$.

7.2. *Effects on statistical estimation error from historical data*

As we have mentioned before, the statistical mean, variation and road coordination factor are obtained from historical data. It is obvious that these parameters are different from current road conditions. In this subsection, we will show that even though the statistical data is estimated, our model to compute optimal granularity is robust. That is, we show a range of estimation error such that within this range, the optimal granularity computed by our model performs close to the true optimal in terms of user utility gains.

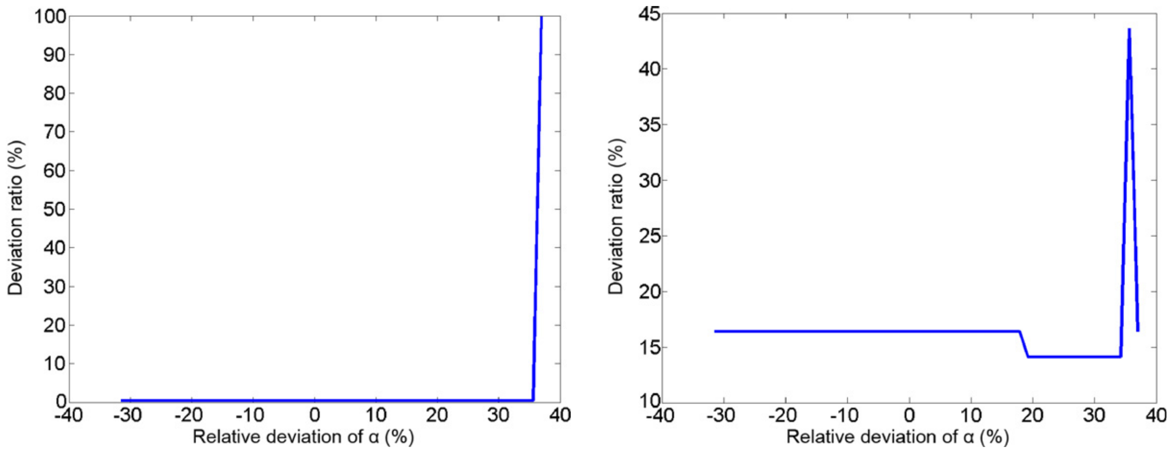
Assume we use i.i.d. to estimate the parameters. We first fix mean μ and road coordination factor α to investigate how the error of estimating deviation σ affects the result. When fixing $\mu = 55$ and $\alpha = 0.85$, Fig. 7 illustrates the effect of changing σ from 1 to 18 when $\beta = 0.1, 0.9$. The comparison foundation is $\sigma = 13.4$, corresponding to the zero point on the x-axis. The experiment shows that when estimation deviation is less than 50%, the performance of our algorithm is good.



(a) $\beta = 0.1$.

(b) $\beta = 0.9$.

Fig. 7. Effects of σ on the optimal utility gained.



(a) $\beta = 0.1$.

(b) $\beta = 0.9$.

Fig. 8. Effects of α on the optimal utility gained.

Fig. 8 demonstrates the cases where $\mu = 55$, $\sigma = 10$, when changing α from 0.5 to 1 and sampling $\beta = 0.1, 0.9$, the effects of error on estimating α to final optimal value. The comparison basis is $\alpha = 0.73$. The experiments suggest that the estimation on α needs to be more accurate. The estimated α deviation within 10% is tolerable.

We have also conducted some experiments on investigating μ's impact on the estimation error. However, the experiments show that μ's change does not affect the deviation of utility gained from the model computed value to the empirical value.

A problem we want to discuss here is the noise in the data set. In the experiment validation section where we considered the effects of estimation error, we observed a sudden drop in Fig. 7(b) (when the deviation is around −90) and a sharp rise in Fig. 8(b) (when the deviation is around 35). Due to the fact that there is only a single point deviating from other surrounding nodes, we think that one possible explanation is that there is noise in the raw data set. In our validation procedure, the noisy data has not been taken into account. Defining, identifying, and eliminating the noisy data from this huge data set is an open problem and is out of the scope of this paper.

8. Related work

There is a large body of research in the area of privacy preservation in traditional Internet based social networking applications [5–12]. However, bringing the concept of mobility to social networking magnifies these concerns immensely as compromised location privacy may lead to serious security concerns. One major difference between mobile social networking and traditional Internet social networking is related to a user's location information. In traditional Internet, unless a user is willing to reveal his/her location information (such as zip code or street address), his/her precise location privacy is preserved. However, in some mobile social networking applications, location information can be provided and used for social benefits such as the urban traffic monitor application discussed in this paper.

Several software solutions [13–16] have been proposed to protect privacy in mobile applications. Tang et al. [17] proposed a distributed method for storing personal information in mobile devices. Hong et al. [14,15] proposed Confab, a toolkit for mobile application developers and end users which supports a broad spectrum of privacy needs. Capra et al. [13] suggested a middleware architecture that provides privacy for mobile applications.

Several experimental systems [18–20] also built location based services where the location of a mobile device is hidden from the service provider for protecting privacy. However, in these systems, a basic assumption is that there exists a trusted central authority that is able to provide accurate information. Therefore, it is the authority's responsibility to preserve users' privacy. In those applications, privacy can be protected by applying k-anonymity [21] or l-diversity [22] mechanisms. In contrast, our traffic monitor system does not have a trusted central authority to secure a user's information. The application at the user side considers a user's tradeoff preference between privacy and traffic estimation accuracy. This preference is used to decide whether to report local traffic conditions with some ambiguity. Privacy is gained from the fact that data which is not transmitted cannot be compromised. Protecting privacy at the user end before sending out the sensitive information also removes the cost of maintaining a trusted central authority. Therefore this mechanism is considered more suitable for traffic monitoring, as a light-weight real-time social network mobile application.

It is only recently that human-involved sensor-embedded mobile applications have become one of the important streams of sensor network research [23–26]. Miluzzo et al. [25] proposed CenceMe, a large-scale deployment of sensor-equipped mobile phones to facilitate the sharing of "presence" information among friends. Eisenman et al. [26] investigates how personal recreation can benefit from sensing in the BikeNet projects. Reddy et al. [23] developed the Campaign framework for creating urban participatory sensing using mobile devices. In [2], Reddy et al. further develop a set of metrics to help participatory sensing organizers determine individual participants' fit with any given sensing project, and describe experiments evaluating the resulting reputation system. Shilton et al. [27] discussed the benefits and challenges of participatory design in participatory sensing settings, and outline a method to integrate participatory design into the research process. Hoh et al. [3] proposed a social network based traffic sensing application using the concept of spatial sampling with virtual trip lines. In these previous studies the focus is primarily on absolute user privacy rather than trading privacy with service quality. This paper specifically focuses on relative privacy where each single user can trade his/her privacy with expected traffic estimation accuracy by maximizing utility value.

9. Conclusions

In this work, we consider an urban traffic monitoring application in which a centralized server collects updates about locations and speeds from a population of sensor-embedded mobile devices belonging to application subscribers in order to estimate current traffic conditions. There is a major tension between privacy preservation and the service quality requirements of the users. Specifically, an individual user prefers to reveal as little of his/her own traffic information as possible while maintaining a certain level of traffic estimation accuracy.

In order to solve this problem, we propose a utility-based optimization policy. The trade-off from an individual user's perspective is modeled as a utility function that linearly combines the benefit of high quality traffic estimates and the cost of privacy loss. By using a novel Markov-based model, we are able to measure the traffic estimation quality so that it is feasible to mathematically derive an optimized information update policy to let the user contribute "just enough" local information to the backend server.

Furthermore, the efficiency of our proposed policy is validated through real-world empirical traces collected from a day-long 100-vehicle experiment on a highway in northern California. The validation demonstrates that the policy yields utilities for each user that are close to what could be obtained with an oracle scheme that has full knowledge of the ground truth.

We recognize that a major privacy concern with the proposed traffic monitoring application is the ability for a malicious server to reconstruct a user's trajectory using a set of location reports. For example, with two consecutive location reports one may be able to determine that a user is headed north in some particular area of a city. The more location reports that are acquired, the more possible it becomes to determine the exact route a car took. In a simple implementation that allows individual reports to be presented without any identity obfuscation, this has an unavoidable impact on privacy for the user.

We have some ideas to address this concern that could be explored more carefully in future work. The privacy impact due to combination of location reports can be minimized by reducing the frequency (number) of location reports sent by the user. The minimum reporting frequency will depend on factors such as the density of application users. A more sophisticated implementation of this protocol could take additional measures to anonymize (such as via aggregation or temporary identifiers) the identity of vehicles generating each report. In either case, users will have greater privacy when the number of reporting vehicles is higher.

Another possible direction is to relax the assumption that all the users in a given stretch of road should use the same information granularity, and investigate the case where different users can choose different information granularities by using game theoretic tools, possibly depending on their geographic location. This effectively allows a user to influence their location granularities based on personal preferences, allowing for greater privacy protection when in sensitive locations (e.g., near residence) and lower protection in other, more public, locations.

Acknowledgment

This work was funded in part by the National Science Foundation, under award number 0831545.

References

[1] MetroSense project. http://metrosense.cs.dartmouth.edu/.

[2] S. Reddy, K. Shilton, J. Burke, D. Estrin, M. Hansen, M. Srivastava, Evaluating participation and performance in participatory sensing, in: Proceedings of International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems—UrbanSense08, Raleigh, North Carolina, November 4, 2008.

[3] B. Hoh, M. Gruteser, M. Annavaram, Q. Jacobson, R. Herring, J. Ban, D. Work, J. Herrera, A. Bayen, Virtual trip lines for distributed privacy-preserving traffic monitoring, in: Proceedings of the 6th International Conference on Mobile Systems, Applications and Services, June, 2008.

[4] J.-C. Herrera, D. Work, X. Ban, R. Herring, Q. Jacobson, A. Bayen, Evaluation of traffic data obtained via GPS-enabled mobile phones: the Mobile Century field experiment, Transp. Res. C 18 (4) (2010) 568–583.

[5] A. Adams, Multimedia information changes the whole privacy ballgame, in: Proceedings of the Tenth Conference on Computers, Freedom and Privacy, 2000, pp. 25–32.

[6] S.G. Davies, Re-engineering the right to privacy: how privacy has been transformed from a right to a commodity, in: Technology and Privacy: The New Landscape, MIT Press, 1997.

[7] S. Patil, A. Kobsa, Uncovering privacy attitudes and practices in instant messaging, in: Proceedings of the 2005 International Conference on Supporting Group Work, 2005, pp. 109–112.

[8] S. Lederer, J. Mankoff, A.K. Dey, Who wants to know what when? privacy preference determinants in ubiquitous computing, in: Extended Abstracts on Human Factors in Computing Systems, 2003, pp. 724–725.

[9] S. Consolvo, I.E. Smith, T. Matthews, A. LaMarca, J. Tabert, P. Powledge, Location disclosure to social relations: why, when, & what people want to share, in: Proceedings of the Conference on Human Factors in Computing Systems, 2005, pp. 81–90.

[10] L. Barkhuus, A.K. Dey, Location-based services for mobile telephony: a study of users' privacy concerns, in: Proceedings of the 9th International Conference on Human–Computer Interaction, 2003.

[11] I. Smith, S. Consolvo, J. Hightower, J. Hughes, G. Iachello, A. LaMarca, J. Scott, T. Sohn, G. Abowd, Social disclosure of place: from location technology to communication practice, in: Proceedings of the International Conference on Pervasive Computing, May 2005.

[12] M. Annavaram, Q. Jacobson, J.P. Shen, HangOut: a privacy preserving social networking application, in: The Workshop on Mobile Devices and Urban Sensing, April, 2008.

[13] L. Capra, W. Emmerich, C. Mascolo, A micro-economic approach to conflict resolution in mobile computing, in: Proceedings of the 10th Symposium on Foundations of Software Engineering, 2002, pp. 31–40.

[14] J.I. Hong, J.D. Ng, S. Lederer, J.A. Landay, Privacy risk models for designing privacy-sensitive ubiquitous computing systems, in: Proceedings of the 5th Conference on Designing Interactive Systems, 2004, pp. 91–100.

[15] J.I. Hong, J.A. Landay, An architecture for privacy-sensitive ubiquitous computing, in: Proceedings of the 2nd International Conference on Mobile Systems, Applications, and Services, 2004, pp. 177–189.

[16] M.F. Mokbel, C. Chow, W.G. Aref, The new Casper: query processing for location services without compromising privacy, in: Proceedings of the 32nd International Conference on Very Large Data Bases, 2006, pp. 763–774.

[17] J. Tang, V. Terziyan, J. Veijalainen, Distributed PIN verification scheme for improving security of mobile devices, Mob. Netw. Appl. 8 (2) (2003) 159–175.

[18] G. Iachello, I. Smith, S. Consolvo, M. Chen, G.D. Abowd, Developing privacy guidelines for social location disclosure applications and services, in: Proceedings of the 2005 Symposium on Usable Privacy and Security, 2005, pp. 65–76.

[19] K.P. Tang, P. Keyani, J. Fogarty, 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 Conference on Human Factors in Computing Systems, 2006, pp. 93–102.

[20] M. Gruteser, D. Grunwald, Anonymous usage of location-based services through spatial and temporal cloaking, in: Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, 2003, pp. 31–42.

[21] L. Sweeney, *k*-anonymity: a model for protecting privacy, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10 (5) (2002) 557–570.

[22] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, *l*-diversity: privacy beyond *k*-anonymity, in: Proc. 22nd Intnl. Conf. Data Engg., ICDE, 2006, p. 24.

[23] S. Reddy, J. Burke, D. Estrin, M. Hansen, M. Srivastava, A framework for data quality and feedback in participatory sensing, in: Proceedings of the 5th International Conference on Embedded Networked Sensor Systems, 2007, pp. 417–418.

[24] N. Eagle, A. Pentland, Reality mining: sensing complex social systems, Pers. Ubiquitous Comput. 10 (4) (2006) 255–268.

[25] E. Miluzzo, N.D. Lane, S.B. Eisenman, A.T. Campbell, CenceMe—Injecting sensing presence into social networking applications, in: The 2nd European Conference on Smart Sensing and Context, EuroSSC'07, Lake District, UK.

[26] S.B. Eisenman, E. Miluzzo, N.D. Lane, R.A. Peterson, G.-S. Ahn, A.T. Campbell, The BikeNet mobile sensing system for cyclist experience mapping, in: Proceedings of the 5th International Conference on Embedded Networked Sensor Systems, Sensys'07, Sydney, Australia, November, 2007.

[27] K. Shilton, N. Ramanathan, V. Samanta, J. Burke, D. Estrin, M. Hansen, M. Srivastava, Participatory design of urban sensing networks: strengths and challenges, in: Participatory Design Conference, Bloomington, Indiana, October 1–4, 2008.