A Privacy Mechanism for Mobile-based Urban Traffic Monitoring

Chi Wang[§], Hua Liu[†], Bhaskar Krishnamachari[†], Murali Annavaram[†]
[§]Tsinghua University, Beijing, China
sonicive@gmail.com
[†]University of Southern California, Los Angeles, CA, USA
{hual, bkrishna,annavara}@usc.edu

ABSTRACT

Participatory sensing is a paradigm that allows each participant to sense, collect and transmit information about their surroundings to either other members in the group or to a centralized server. The information that is provided by the community of users is then combined to provide a useful service to all the participants. The focus of this work is one such participatory sensing application, namely mobile traffic monitoring. In this application each participant provides real time update on location and speed of the user's vehicle to a centralized server; information from multiple participants is then aggregated by the server to provide current traffic conditions to all participants. Successful participation in traffic monitoring application depends on two factors: the information utility of the estimated traffic condition, the amount of private information (speed and position) each participant reveals to the server. Each user prefers to reveal as little private information as possible, but if everyone withholds information, the quality of traffic estimation will deteriorate. We model these opposing requirements by considering each user to have a utility function that combines the benefit of high quality traffic estimate and the cost of privacy loss. Using a novel Markov-based 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 for this stretch of road to maximize user utility. We validate the effectiveness of this policy through real-world empirical traces collected from a day-long 100-vehicle experiment on a highway in Northern California, conducted in 2008. The validation shows that 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.

1. MOTIVATION FOR PARTICIPATORY SENSING

In existing sensor networks, power-constrained sensors are deployed in the targeted area and data is collected till the sensor runs out of battery 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 need either 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. In order to overcome these shortcomings of current sensor networks, researchers have proposed projects such as *MetroSense* [2] and *Participatory Sensing* [27].

This new generation of sensing projects are based on the concept of "people-centric 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 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 *silent sensing*. Mobile phones with several integrated sensors, such as GPS, audio, Bluetooth, and Wifi are increasingly being used in 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 such 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 condition on the 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 increasing number of participants.

1.1 Importance of Traffic Sensing

Population growth in the U.S. metropolitan areas has outgrown the transportation infrastructure. As a result, free-

way congestion is rapidly becoming a major economic hurdle. Estimates show that traffic congestion cost over 10 billion dollars in economic activity in 2003 [3], and burnt over 400 million gallons of excess fuel in Los Angeles metropolitan area alone. Commuters have turned their attention to real-time traffic monitoring and drive time estimation services [4] to avoid congested areas and to find alternate routes. These services all rely on traffic estimation based on *loop inductors* that are installed below the road surface on major freeways. Inductors provide the speed and density estimates based on vehicles that travel over the inductors. Many freeway inductors are connected to a centralized data server and send information to the server every time a car passes over an inductor. Data from these inductors is aggregated by the server to provide real-time traffic conditions. There are two disadvantages of loop inductors. First, loop inductors are expensive to install and maintain, and hence they are installed only on a few major road segments. Loop inductor installation is estimated to have cost 2.5 billion dollars already in the state of California. Second, majority of the installed inductors, except for those on some freeways, do not provide traffic updates to the data server. These inductors are primarily used for signal activation rather than traffic data collection.

In such contexts, we believe that GPS-embedded mobile devices will provide a cost effective alternative to provide real time traffic information, where they can augment inductor-based traffic sensor data by providing precise speed information at any arbitrary location, not just on freeways. In particular, mobile devices can provide traffic information even on secondary and tertiary roads where installing and managing inductor coils may be prohibitively expense. Mobile devices are integrating a variety of system components, such as on-board GPS receivers, that make them uniquely well suited for traffic sensing. They also have enough computing power to process the sensor data to make intelligent local decisions on when and how much traffic sensing information to update to a backend server. Finally, they can use their communication capabilities to instantly transmit that data to a backend data aggregator that can provide customized traffic service to an end user. The combination of device features, near universal availability, and wide coverage create new opportunities to dramatically change traffic sensing, traffic data aggregation, and most importantly real-time traffic estimation.

1.2 Importance of Privacy

While the motivation for traffic sensing using mobile phones is clear, the approach described above, where the user reports the speed and position information to the aggregator potentially compromises the participant's privacy. The simple sense and transmit approach totally ignores the device holder's privacy. Note that, 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 particularly, when a mobile phone is being used as sensor, the sensing device and the participant are closely tied together. A mobile phone identifies the sensor uniquely with a participant's identify. 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's) privacy into account when designing the system. If the accurate location/speed information is eavesdropped 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 that may tradeoff some data accuracy with 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. By coarse location information, privacy is protected while the system can still maintain reasonable service quality. In order to implement such utility based information update policy, we propose a novel Markov model to evaluate the impact of granularity on the accuracy of traffic estimation (i.e., the application service quality).

Specifically, in this paper we propose a policy which helps a single user to decide on the optimal information precision. We assume that the input to the policy is the mean, variance, and correlation information for a given road-stretch. A novel Markov-based model formulation is applied to the road traffic estimation accuracy measurement. Based on the Markov model, we propose a particular utility function that considers the tension between traffic estimation error and users' potential privacy loss. With the utility function, we are able to compute the optimal granularity for traffic information update on the corresponding road section. We validate this policy on a traffic update database. The traffic update database was generated from a recent study involving 10-hour 100-vehicles freeway real traffic experiment which is conducted on Feb 8th 2008 jointly by the Nokia Research Center and the University of California, Berkeley [1]. In this large scale study, the very first of its kind in the United States, 100 drivers provided over one million traffic updates to a backend database server.

It worth noting that the inherent trade-off between privacy and traffic estimation precision is the core for application design. One of our novel contributions is to formulate this tension as a utility optimization problem from the perspective of a single user, and derive a near-optimal policy that maps a set of available *a-priori* knowledge about traffic conditions to a deterministic decision about what spatial granularity the user must send information to the server.

This paper makes the following three contributions. First, we propose a Markov-based road model that takes into account the mean, variance, and correlation of traffic on a given stretch of road for traffic conditions. This model allows us to estimate the impact of granularity on estimation accuracy. Second, we formulate the decision making problem for an individual user (to decide the information granu-

larity to contribute to the society) as a utility optimization problem. The optimization problem assumes that the users are intelligent with characteristic of rationality and selfishness. A policy is derived based on the formulation which yields the optimal precision of information revelation for the corresponding road stretch. The information precision is optimal in the sense that if a user uses this granularity to reveal his/her local information, he/she can get optimal utility, which is a trade-off between privacy leak risk and social service quality. Third, extensive performance analysis of our proposed policy has been done on real experiment data consisting of more than one million traffic update records collected during a 10-hour 100-car experiment. Our analysis shows that a) our proposed policy is near optimal in all cases; b) the proposed policy is robust and it still yields good utility gain for users when the three parameters' estimations have errors.

The paper is organized as follows. Section 2 describes the traffic monitoring application. Section 3 and Section 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 user's utility. Our experiment methodology and encouraging results are presented in section 6 and 7. Finally we present some related works in Section 8 and conclude our work in Section 9.

2. APPLICATION DESCRIPTION

As we have discussed in the introduction section, we believe that the mobile based urban traffic monitoring system will help relieve the traffic conditions in future and help application users estimate traffic conditions on the road with privacy reservation concerns. A straightforward version of this urban traffic monitoring application is shown in Figure 1(a).

In the simplest version of this application, we envision the virtual trip line (VTL) sensors [22] as replacement for inductive loop sensors mentioned before. Virtual trip lines are GPS coordinates of a line that is *virtually* drawn on top of any road segment by a traffic administrator, such as US DOT (Department of Transportation). Virtual trip lines are stored in a database clustered by a geographic region. Reads to the database can be done by any mobile device client but updating the database can be done only by the traffic administrators. Any mobile device that enters a geographic region accesses the database and downloads the VTLs over the air. Mobile devices monitor their location using GPS and use the cached VTLs from a region to determine if they are crossing a VTL. 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 an 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 focusing on users' exact location and speed. If the user's update information is overheard, or maliciously detected by eavesdroppers, the user's privacy is leaked by revealing the exact location and speed information. Note that although the application may not need the user's identity when collecting the traffic condition updates, the MAC layer of the mobile devices implicitly reveals user's identity by using 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 do better privacy protection. Therefore, we propose a utility based privacy preservation model for the traffic monitoring application (see Figure 1(b)). This modified application considers the tradeoff between the users' desire to protect privacy, and their requirement to have accuracy on traffic estimation error and provides a policy to optimize this tradeoff. That is, the improved traffic monitoring application allow the users to contribute to the system with "just enough" amount of information to preserve privacy and meanwhile, make the use of the traffic estimation with proper precision.

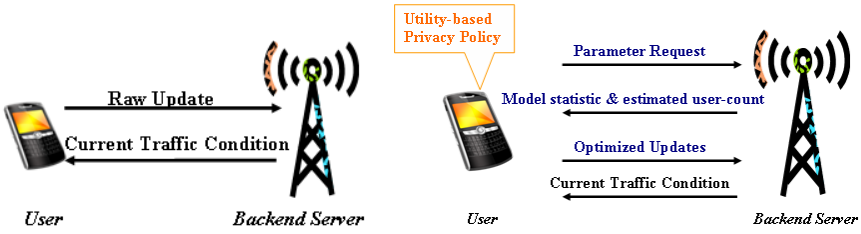
This modified traffic monitoring application (which is the focus of the remaining parts of this paper) consists of four message exchanges .

- First, 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 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?".

- Second, the backend server returns those parameter values (also referred as model statistics in this paper) possibly based on the historic data, as well as an estimated number of users.

- Third, users send out the optimized local information updates to the backend server. Upon receiving these model and estimated statistics, the application at the user side either computes or uses a look-up table to find an optimized update granularity. Note that in such a community-based application, the quality of collected global information depends on the quality of information contributed by individual end users who have the motivation to protect their privacy. Our proposed utility-based privacy policy formulates the tension between traffic estimation accuracy requirement, and user's desire about his/her own privacy into a utility function, then maximizes this utility function to obtain the optimal updates. With this modification, instead of reporting exact location information, as in the original simplest version, a user might vague his/her location information into a proper distance length such as "somewhere between VTL 34 and VTL 39". This information includes an implicit spatial granularity (user's location information with proper precision) and the user's current vehicle speed (i.e., traffic flow speed) with a timestamp.

[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.



(a) Original Application Model (b) Modified Application Model

Figure 1: Comparison of the original (intuitive) scheme and modified scheme

- Fourth, the backend server returns current traffic conditions on the road stretch to the application subscribers. According to the reported information from all users in the community, the application server is capable of estimating real-time averaged traffic flow speed on the road, which can help users monitor the traffic conditions for the interested road stretch.

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

We propose a Markov-based road model in this research to measure the traffic estimation precision with minimized number of parameters. The main purpose of this Markov-base traffic model is to characterize the impact of granularity on traffic estimation accuracy, so that we can measure the system 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 VTL. In other word, each VTL intervals implies a spatial granularity. For example, "between the 105th and 110th VTLs" implies granularity 5.

Both the user's local information submitted to the centralized server and the server to user's aggregated information feedback are a tuple of VTL intervals and speed. For example, $(100 - 105thVTLs, 30mph)$ is a valid information submittal.

It is straightforward that in this traffic monitoring application, the backend traffic estimation precision depends on the accuracy of the local sensing information contributed by individual users. On one hand, precise information submission yields better integrated traffic estimation but compromises on an individual's privacy. On the other hand, too conservative a contribution to the system will make the quality of traffic estimation suffer. In order to quantify the quality of service as a function of location information spatial granularity, we propose a novel Markov road condition model in this section. This Markov-based model only takes 3 statistical parameters into account and is able to quantify the estimation error with spatial granularity.

3.2 Assumptions and Notations

We assume a complete road stretch as a line with length l. n VTLs are settled on the whole road from left end to right end. The road is divided into sections evenly by the VTLs. The sections are continuous and non-overlap road segments. Each section contains a length of $\frac{l}{n}$ where n is the number of VTLs.

Suppose that the average vehicle speed in each section at certain time spot is a random variable, denoted by $X_1, X_2, \ldots, X_n$ from the first section to the n-th 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 i-th section is correlated with the speed X_{i+1} at the $(i+1)$-th section. To model this correlation on the traffic flow, we import 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 as shown in figure 2 which is separated into n sections. Recall that X_i denotes the average traffic speed in the i-th section, a first-order Markov-based equation is proposed as follows:

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

where x_i is a random variable corresponding to the speed fluctuation related to the vehicle location. $x_1, x_2, \ldots, x_n$ are independent, with corresponding mean μ_i and variance σ_i^2. Note that 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

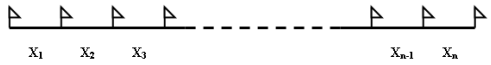


Figure 2: Illustration for a road with n sections

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.

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 feedback estimated averaged 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:

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

where $Y_k = \frac{1}{g}\sum_{i=(k-1)*g+1}^{k*g} X_i$ is the estimated speed collected by the service for section $(k-1)*g+1$ through $k*g$. 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 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 $t\sqrt{MSE}$ in each sampling time period. We derive the expectation of the estimated MSE from equations 2 and 1. Given a fixed time interval, let μ_i denotes the expectation of x_i and σ_i denote the standard deviation of x_i. Consider the following three equations:

$$E(x_i^2) = \mu_i^2 + \sigma_i^2$$

$$E(x_i) = \mu_i$$

$$E(\sum (x_i - \frac{\sum x_i}{g})^2) = E(\sum x_i^2 - \frac{(\sum x_i)^2}{g})$$

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

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

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, therefore makes user utility modeling on the tradeoff between service quality and privacy protection feasible. We will now show in section 4 how we use $e(g)$ while computing the user's utility formulation.

4. UTILITY FORMULATION

Privacy modeling is the difficult part for this application design. In previous section, we proposed a Markov road traffic model to quantify the quality of service with the spatial granularity of the location and speed information reported by on-road users. In this section, we focus on discussing how to quantify the privacy loss in terms of information precision.

As we have mentioned in section 2, we have to take people's privacy into account in this application. If a malicious attacker can obtain the application user's identity information with accurate location and speed information, the malicious attacker will be able to track the user's location and his/her movements, which is dangerous for the application user. Therefore, properly addressing the privacy concerns is the first step to ensure broad user participation which is a primary requirement for our approach to work in practice.

However, as inherited behavior for this class of applications, the accuracy of traffic estimation depends on the accuracy of following three fundamental factors: vehicle location information, time of VTL crossing, speed of VTL crossing. Accuracy of traffic estimation is proportional to the information accuracy along these three dimensions. Coincidentally, the notion of user's privacy is inversely proportional to the information accuracy. 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 dimension: time and speed.

Finally, if a series of traffic updates within a time window can be associated with a single user then it is easy to reconstruct the user's travel path. This information when combined with the road network knowledge can even predict the user's future trajectory of motion.

In our application, we use granularity to give users choice to vague their exact locations to increase the difficulty of malicious tracking. On the other hand, in order to obtain remarkable performance of the whole service, the users have to impose some information to the system to get useful system traffic estimation.

The utility function of each application user has two parts: privacy protected (denote 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 protected function $p(g)$ is modeled as a function of granularity g. Intuitively, the users updated information can be explained as "I am driving between an interval with this speed". Increasing length of the interval will vague the information so that decreases the probability that users exact location got detected. This fact means $p(g)$ is a function that increases when g increases. There are different ways to model the privacy in this application. Particularly, we use the following function to model the privacy protected [2]:

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

This function is proportional to g. The physical meaning of this equation can be explained as following. Suppose that the car we are discussing is now at somewhere of the road we considered and the road total length is l. If the user's upload information does not narrow down the search scope, the driver owns a private space with the length of l. That is, the probability to reveal the exact location of the user is uniformly distributed in this piece of road. On the other hand, if $g = 1$, we suppose the exact location of the driver is revealed to the system. If the information comes 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 $\frac{1}{g}$ compare to the best privacy reservation that can be reached 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 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:

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

[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.

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 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. We suppose the only parameters 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 the user by querying backend server, as discussed in Section 2. 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 granularity, 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), we can compute a pre-stored look-up table entry [3] for optimized g. This look-up table is pre-calculated and stored at the client side. Once a user obtains the model statistic parameter tuple, the application will retrieve the look-up table and find the best analytical strategy for information update granularity. We will validate this policy in the experiment results section. The validation shows that 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.

It is worth noting that the validation is taken from a single user's aspect of view. If the user can know (or the backend server can estimate and tell the user) that there are other $m-1$ subscribers in same road section with him. With the assumption that the optimal granularity g is same to them, it is a resource waste to collect m duplicated traffic condition records from every user in this road section. Also notice that we assume the wireless channel is open and the server broadcast the traffic condition estimation to all subscribed users at this road section. Therefore, we only need one of the m service subscribers to report to the server in order to collect the current traffic information of this piece of road section. An intuitive way to reduce duplications is that if the subscriber gets to know the (estimated) number of other users (suppose there are $m-1$) are within the same road section with him/her, he/she can use $\frac{1}{m}$ as the probability to contribute to the application. Note that if all the users are hiring this non-deterministic probability to report traffic conditions, we can only say that with high probability, the users will get service from the server. A timer for service tracking is needed in case no user reports the traffic condition in this section. If a user cannot receive feedback from the server before each timer is ticked, he/she will increase the

[3]Note that we can also compute the policy online. Using look-up table can improve response speed while occupies memory space. Compute policy in real-time saves the storage space but takes longer time in finding the optimized policy. Here, we only pick up the look-up table as an example.

probability to contribute his/her local knowledge to service. The detailed design issues, system parameter configurations, and performance evaluations for non-deterministic method are one of the focuses of our future work. In the remainder of this paper, we only consider the simplified case where the users are not able to estimate how many other users in the same road section to validate the utility-based policy we proposed in this paper. That is, after determining the spatial granularity, the users always report traffic conditions with the optimized granularity.

In the following of this section, we are going to discuss the effects of the parameters on the optimal strategy in a simplified case where we assume random variables x_i ($i = 1, 2, ..., n$) are identically distributed. The name "simplified" case is to contrast the "complex" case. In the complex case, we do *NOT* have the additional assumption that x_i and x_j (for $i, j = 1, 2, ..., n$ and $i \neq j$) are identically distributed. For the "complex" case, each road stretch may have different mean (μ_i) and variance (σ_i). The complex case is a closer approximation to the real road traffic since the statistical parameters are obtained directly from the historical data without approximations, while the simplified case decreases the computational overhead. Since this is a real-time power-constrained mobile device involved application, we believe the computational cost of the simple model is an attractive way to compute granularity. Moreover, as we show later on in section 7, the simplified model performs as well as the complex model.

5.1 Impact of Parameters on the Optimal Granularity Using a Simplified Model

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 independently identical distributed (i.i.d). With the i.i.d assumption, the statistical parameters μ and σ satisfy $\mu_1 = \mu_2 = \cdots = \mu_n$ and $\sigma_1 = \sigma_2 = \cdots = \sigma_n$. The corresponding traffic satisfies the following property:

The expectation of the speed at every location are equal.

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

Figure 3 illustrates the impact of analytical optimized granularity g^* in the simplified case for different α values when β changes. In the remaining part of this section, we discuss parameters one by one to show their impacts on analytical g^*.

- 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 had better choose finer granularity, to increase the overall utility. Intuitively, the more a user weights privacy, the more hesitate the user is to provide accurate information.

  A notable observation from figure 3 is that g^* is a convex decreasing function. Specifically, this means when β is relatively small, g^* drops more quickly than β is larger. This phenomenon is consistent with our chosen of privacy function (as discussed in section 4). When the information granularity is coarse enough (i.e., g is sufficiently large), a further ambiguous of the granularity will not have significant effect on the utility.

  Another implication is that only users who are extremely strictly 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 proposed utility function, the service quality is ensured in an appropriate level.

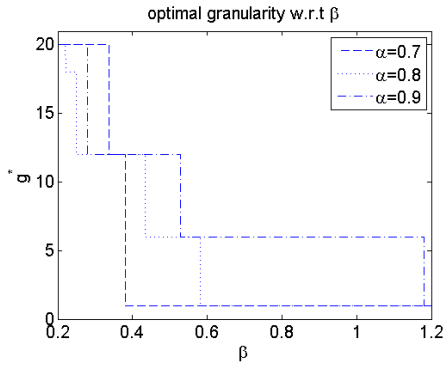


Figure 3: optimal granularity varying with β when $\mu = 50, \sigma = 10$.

- α's impact

  One observation from the plot is the corresponding β value strictly increases at the points where g^* drops to 1 (user's most accurate information is revealed at this point) when α increases [4] Increasing α implies a smoother road traffic condition. When the traffic flow is smooth, privacy weights more than traffic estimation concerns unless the user is very picky in the traffic estimation accuracy (where a larger β is required).

- μ's impact

  Figure 3 is plotted using $\mu = 50, \sigma = 10$. μ almost has no impact on analytical optimized granularity as long as $\sigma << \mu$. 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

[4]This is not a coincidence, it is provable. For the limit of the pages, we omit the proof here but intuitively explain the physical meaning of this phenomenon.

When σ increases, 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 trade better traffic estimations.

We will show in the experiment that simplified case can also approximate the optimal granularity comparing to the complex case where x_is are not assumed to be identical.

6. EXPERIMENTAL METHODOLOGY

In this section, we first describe how the real experiment has been done and what the data structural of the trace data base, followed by characteristics of the trace dataset.

6.1 The Experiments to Obtain the Trace

A large scale experiment to measure the effectiveness of VTLs was conducted on Feb 8th 2008 jointly by Nokia Research Center and University of California, Berkeley. In this experiment 100 cars equipped with Nokia N95 phones which are GPS enabled mobile devices were driven by volunteer drivers along a carefully constructed path in the San Francisco bay area. In this experiment vehicles were driven for 10 hours on an 8 mile section of I-880 south of Oakland CA. A real time screen snapshot taken during the live experiment is shown in Figure 4. The length of test road segment was chosen to have 1% to 2% penetration rate based on the number of participants and approximate round trip travel time. The location of this experiment was specifically selected because it featured both free flowing traffic at greater than 50 mph, and congested, stop and go traffic. At the beginning of the experiment all mobile devices downloaded the VTLs corresponding to the test site to their local cache. As the vehicles drove the travel route 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, speed>. For safety purposes during this experiment the mobile device also explicitly sent its identification information, a unique phone ID, to the backend server to respond to emergencies during the experiment. Sending a traffic update on crossing a VTL can be treated as spatial sampling. The alternative approach to traffic sampling is temporal sampling where the phone sends periodic traffic updates. In order to compare the effectiveness of spatial sampling over temporal sampling our mobile device client also sent periodic (once every 3 seconds) traffic updates to the backend server.

6.2 Characterize the Trace Dataset

This 10-hour experiment traffic dataset contains diversity in terms of speed. It is consisted of fluent traffic flows ($> 50mph$ smooth flows), congested flows and stop and go traffic (refer to a traffic lamp). There are 45 VTLs evenly placed to record the speed measurements from the 20 vehicles. In our experiment, the feasible granularity set contains integers from 1 through 20.

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

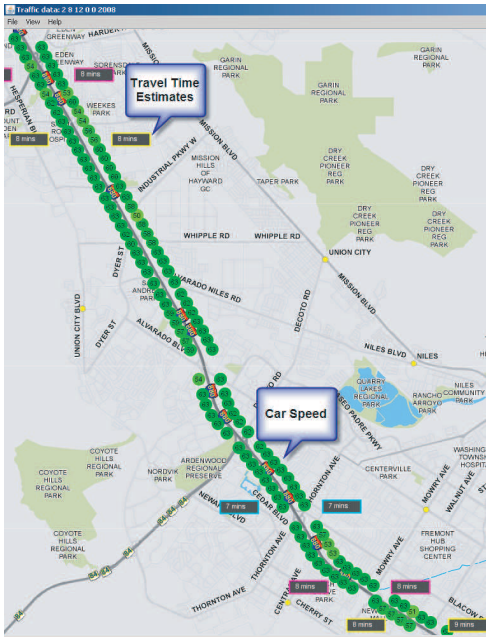


Figure 4: Traffic Experiment

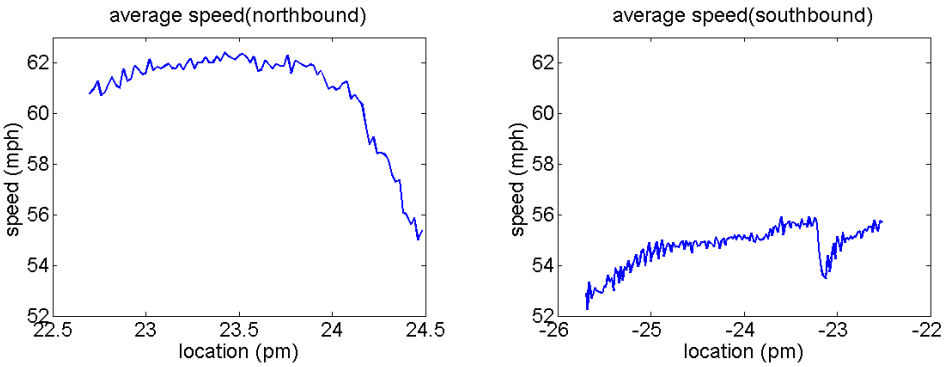
data for vehicles moving from north to south (a.k.a. southbound data).

In order to understand the features of the traffic dataset, figure 5(a) and figure 5(b) plot the average traffic speed between every VTL for the whole road. For the northbound data samples, we use time interval 45000s to 52000s (timestamp in the dataset) for the plot and the average is taken on all the vehicle speeds of passing a VTL in 70 seconds. For the southbound data, time interval is 40000s to 50000s and speed aggregation occurs every 100 seconds.

It is hard to determine α directly from the distribution of speed (X_is) because the distribution of x_i is unknown. In order to estimate road correlation factor α, we use best curve fitting method on the MSE curve. Figure 6(a) and figure 6(b) show the MSE curve fitting for northbound data and southbound data, respectively.

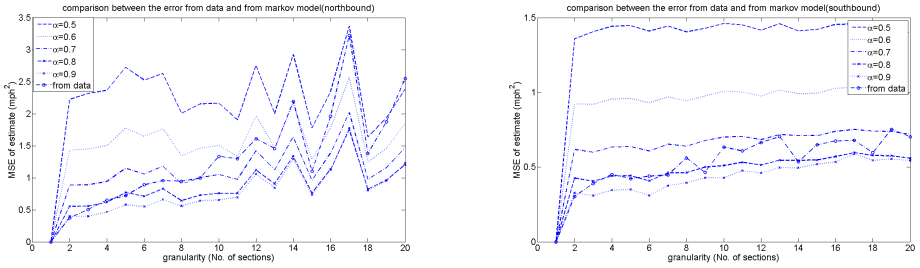
For the northbound data, curve error of estimation calculated from the data traverses the theoretical $E_\alpha(g)$ from $\alpha = 0.5$ through $\alpha = 0.9$, when g decreases. This fact implies that the analytical result fits with the data with different α for varying granularity. This result suggests 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 perfect to practical curve if we focus on the granularity within a smaller range. For example, if we research intensively on $g = 10$ through $g = 16$, $\alpha = 0.6$ is a good enough fitting parameter.

For southbound data, we find that the error curve is basically bounded by the two ($\alpha = 0.7$ and $\alpha = 0.8$) theoretical MSE curves. A single value α may still insufficient to perfect match the traffic pattern, however, the range of reasonable



(a) Northbound Data Set (b) Southbound Data Set

Figure 5: distribution of average speed at every section for a) northbound data set; b) southbound data set



(a) Northbound Data Set (b) Southbound Data Set

Figure 6: comparison of MSE for a) northbound data set; b) southbound data set

α is narrowed down to $0.7 \sim 0.8$. We also noticed that α for southbound samples is larger than the northbound samples, which implies that in the spacial and temporal condition for southbound samples, the correlation between adjacent locations is larger than in northbound samples. This conclusion is consistent with the information conveyed from the speed plot. The higher road correlation factor indicates more smooth distribution of speed (notice that the scales in two speed plots are different).

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 gained. According to the experiment settings, the empirical data set is split to two parts: northbound data and southbound data. We validate the analytical optimal policy in both data sets.

7.1 Near Optimal Utility Validation

In this subsection, we illustrate that although the optimal granularity calculated by our model is not always matching the real optimal point in the trace data, its performance is close to the optimal point in terms of individual user's utility gain.

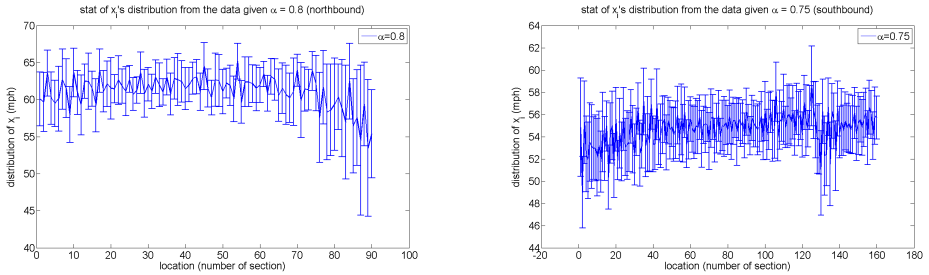
7.1.1 Northbound data set

We let β varies from 0.2 to 5.0 with step size 0.2 in this experiment. For each given β, 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 define two terminologies here we will use throughout the experiments. These two terminologies 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 real empirical data set for each section of the road, while in the "simplified model", as discussed in section 5.1, with identical distribution assumption on a certain segment of the road, simplified identical μs and σs applied for the whole segment.

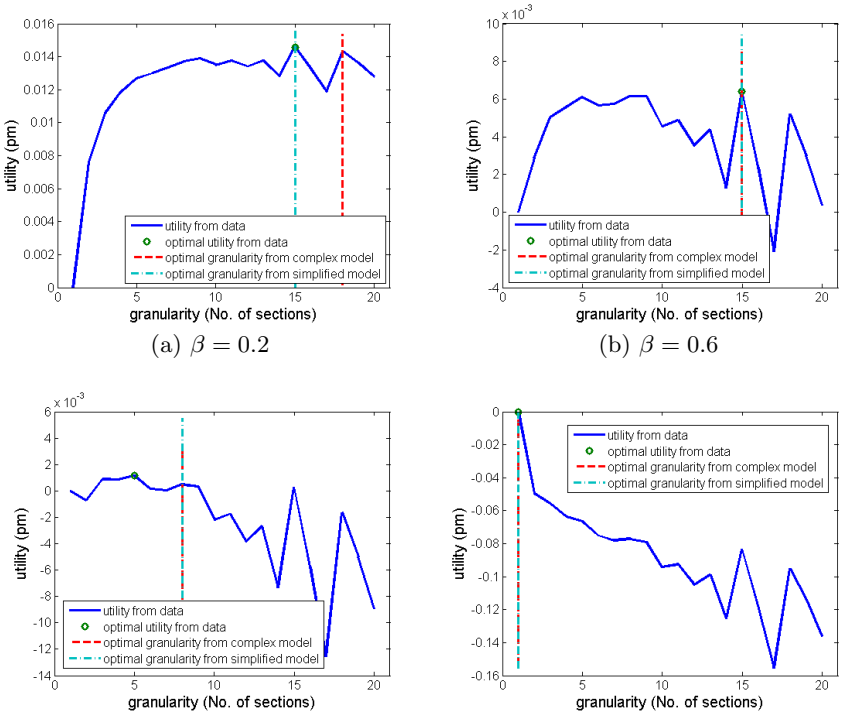
It is worthy noting again that the number of parameters (μ, σ, α) needed in a given stretch of road can vary, with more parameters needed for roads with more complex traffic, which needs to be determined empirically. In our experiments data set, we find empirically that 3 sets of μ, σ and α values are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

In figure 8, each single user's payoff is compared for three different cases with the actual utility curve: a) the circle point is the optimal utility a user could get if there has oracle knowledge; b) dashed red vertical line is the utility obtained



(a) Northbound Data Set (b) Southbound Data Set

Figure 7: x_i's distribution a) $\alpha = 0.8$ for northbound data set; b)$\alpha = 0.75$ for southbound data set



(a) $\beta = 0.2$ (b) $\beta = 0.6$

(c) $\beta = 1$ (d) $\beta = 5$

Figure 8: Utility comparison for northbound data set among empirical optimal utility with oracle knowledge, expected gained utility with complex distribution assumption and expected utility with simplified distribution assumption

by applying "complex model"; c) dotted blue vertical line is the user's utility computed with "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 exist some cases that 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 "simplified model" is more practical than "complex model" in real time traffic monitoring in the sense that the parameter values are estimation values.

In this northbound data set performance validation experiment, the accurate values we used in "complex model" are illustrated in figure 7(a) and the estimated parameters we used in "simplified model" are $\mu_1 = \mu_2 = \cdots = \mu_{50} = 65, \sigma_1 = \cdots = \sigma_{50} = 5, \mu_5 1 = \cdots = \mu_{89} = 62, \mu_{90} = 55, \sigma_1 = \cdots = \sigma_{90} = 7$. The road correlation factor $\alpha = 0.8$ is chosen by the best curve fitting method for all the experiments. Tradeoff preference parameter β varies as $0.2, 0.6, 1, 5$.

We want to highlight this significant result from the 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 the actually 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 also. For the southbound data set, the x_i's distribution used in "complex model" is illustrated in figure 7(b). As we have mentioned before, one set of parameters for the whole road stretch is good enough for the southbound data set. The parameters used in "simplified model" experiments are $\mu = 55$ and $\sigma = 6$. α value is set to 0.75 for both "complex" and "simplified" models. Figure 9 illustrates the utility for different granularity value curve obtained from the empirical southbound data set. The circle points are the maximum possible utility gained with oracle knowledge of future. The vertical lines are the analytical optimal granularity for complex and simplified models. The set of plots vary β as $0.2, 0.6, 1$ and 5. Similar results are observed for the southbound set as well.

We also condense the comparison for different β in one plot each for 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 distribution respectively. For the northbound data set 10(a), α is chosen as 0.8 and for the southbound data set 10(b), $\alpha = 0.75$.

7.2 Effects on Statistical Estimation Error from Historic Data

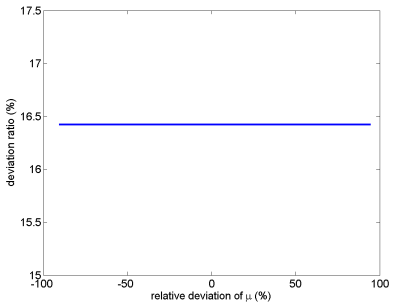


Figure 13: effects of μ to the optimal utility gained ($\beta = 0.9$)

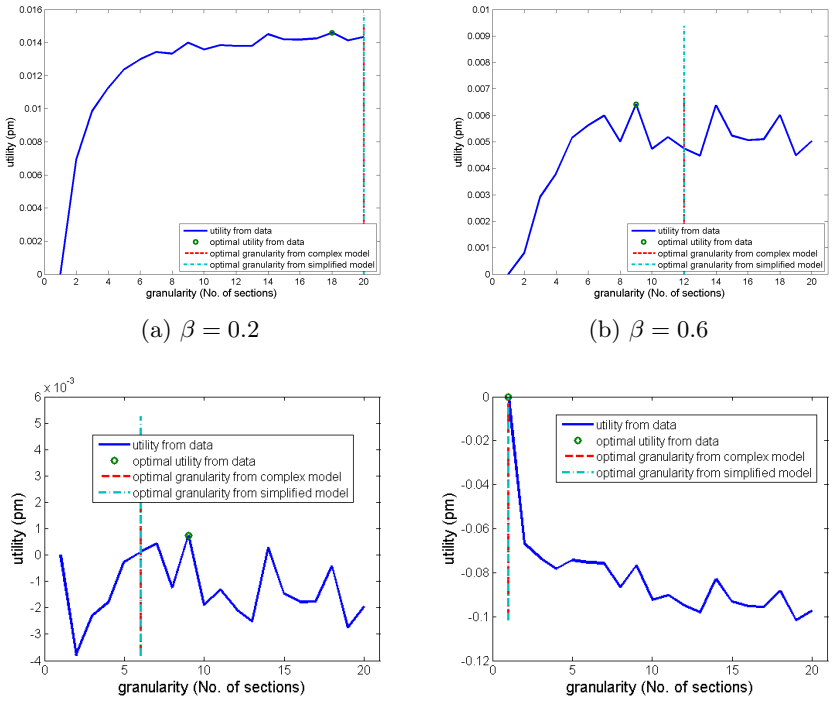
As we have mentioned before, the statistical mean, variation and road coordination factor are obtaining from historic data. It is obvious that these parameters from historic data are different with current road conditions. In this subsection, we will show that even the statistical data is estimated, using 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 still 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 error of estimating deviation σ affects the result. When fixing $\mu = 55$ and $\alpha = 0.85$, Figure 11 illustrates the effect of changing σ from 1 to 18 when $\beta = 0.1, 0.9, 5.0$ (β is sampled as low, medium and high values). The comparison foundation is $\sigma = 13.4$, corresponding to the zero point in x-axis. The experiment shows that when estimation deviation is less than 50%, the performance of our algorithm is good.

Figures 12 demonstrates the cases where $\mu = 55$, $\sigma = 10$, when changing α from 0.5 to 1 and sampling $\beta = 0.1, 0.9, 5.0$, 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 tolerative.

We have also conducted same experiments on investigating μ's estimation error effects. 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. We only show one plot here 13 (corresponding $\beta = 0.9$) and omit other plots for brevity.

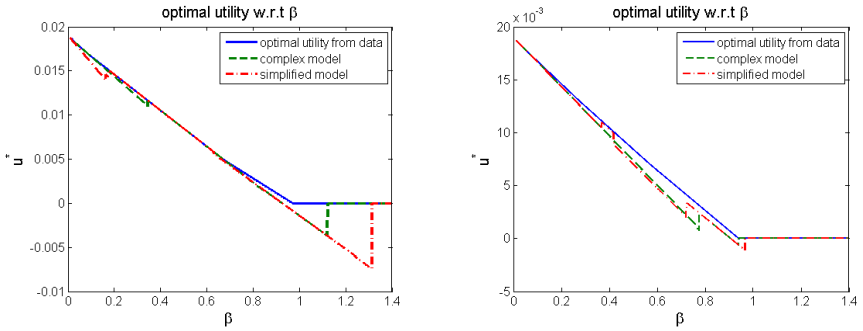
A problem we want to discuss here is the noise data in the empirical data set. In the experiment validation section when we were considering the effects of estimation error, we have observed the phenomenon of sudden drop (in figure 11(b) when x-axis value is around -90) or sharp rise (in figure 12(b) when x-axis value is around 35). Due to the fact that there is only a single point deviating from other surrounding nodes, we think that one possible reason of those points' appearance is that there exists noisy data in the raw data set. In our validation procedure, the noisy data hasn't been taken into account. How to define, identify and elimi-



(a) $\beta = 0.2$ (b) $\beta = 0.6$

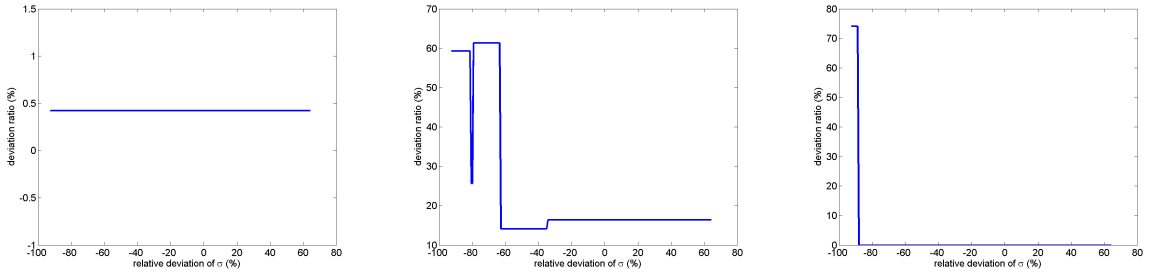
(c) $\beta = 1$ (d) $\beta = 5$

Figure 9: Utility comparison for southbound data set among empirical optimal utility with oracle knowledge, expected gained utility with complex distribution assumption and expected utility with simplified distribution assumption



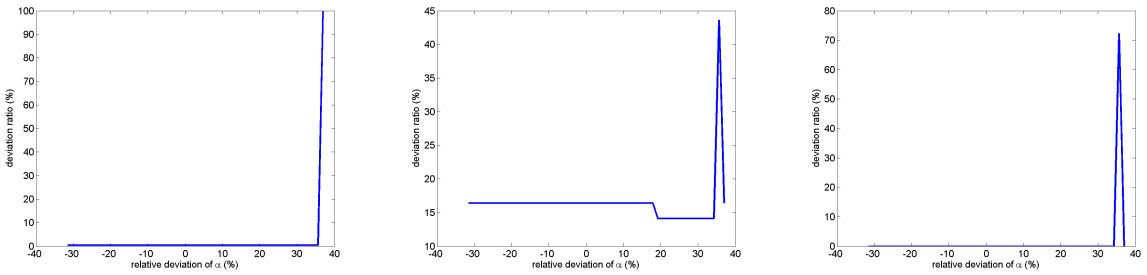
(a) Northbound Data (b) Southbound Data

Figure 10: Utility performance for prediction with complex and simplified distributions for northbound and southbound data sets



(a) $\beta = 0.1$ (b) $\beta = 0.9$ (c) $\beta = 5.0$

Figure 11: Effects of σ to the optimal utility gained



(a) $\beta = 0.1$ (b) $\beta = 0.9$ (c) $\beta = 5.0$

Figure 12: Effects of α to the optimal utility gained

nate the noisy data from this huge data set is an open problem and it is out of the scope of this paper.

8. RELATED WORKS

There is a large body of research in the area of privacy preservation in traditional internet based social networking applications [5, 10, 25, 20, 9, 7, 29]. However, bringing the concept of mobility to social networking magnifies these concerns immensely as compromising location privacy may lead to serious security concerns. One major difference between mobile social networking and traditional internet social networking is related to the user location information. In traditional internet, unless a user is willing to reveal his/her location information (such as zipcode, street address) subjectively, 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 [14, 8, 17, 18, 11, 24] have been proposed to protect privacy in mobile applications. Tang *et al.* [31] proposed a distributed method for storing personal information in mobile devices. Hong *et al.* [17, 18] proposed Confab, a toolkit for mobile application developers and end users which supports a broad spectrum of privacy needs. Desmet *et al.* [11] implemented a software architecture to allow the secure execution of third party applications on a Windows Mobile device. Capra *et al.* [8] suggested a middleware architecture that provides privacy for mobile applications.

Several experimental systems [19, 32, 15] 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 [30] or l-diversity [21] mechanisms. In contrast, our traffic monitor system does not has a trusted capable authority to secure user's information. The application at the user side consider user's tradeoff preference between privacy and traffic estimation accuracy, together with statistical traffic condition on certain stretch of road to decide reporting local traffic status with some ambiguity that can optimize the tradeoff between privacy preservation and quality of service. Protect privacy at the user end before sending out the sensitive information can not only reduce the overhead on protecting privacy in air, but also decrease the cost to maintain a trusted capable authority. Therefore this mechanism is considered more suitable for traffic monitor, as a light-weight real-time social network mobile application.

It is only recently that human-involved sensor-embedded mobile applications have come to one of the important streams of sensor network research [26, 12, 22, 13, 23]. Miluzzo *et al.* [22] proposed CenceMe, a large-scale deployment of sensor-equipped mobile phones to facilitate the sharing of "presence" information among friends. Eisenman *et al.* [13] investigate how personal recreation can benefit from sensing in the BikeNet projects. Musolesi *et al.* [23] are looking at novel ways to blend the virtual world and the sensed physical world together. Reddy *et al.* [26] developed Campaignr framework for creating urban participatory sensing using mobile devices. In [27], 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.* [28] 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.* [16] 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 service quality requirement for the users. Specifically, an individual user prefers to reveal as little of his/her own traffic information as possible while maintaining 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 estimate 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, conducted in 2008. 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.

There are multiple directions to our possible future works. As we mentioned in this paper, we need to take the effect of traffic density into account and design an application that allows non-deterministic traffic data aggregation scheme. 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 granularity by using game theoretic tools.

10. REFERENCES

[1] Researchers Test GPS-Cell Phone Navigation In South Bay, NBC News, February 2008, http://www.nbc11.com/news/15255056/detail.html.

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

[3] Rand California Traffic Congestion Statistics, http://www.ca.rand.org/stats/community/trafficcongestion.html

[4] Traffic and Driving Times, http://traffic.511.org/

[5] A. Adams. Multimedia information changes the whole privacy ballgame. In *Proceedings of the tenth conference on Computers, freedom and privacy*, pages 25–32, 2000.

[6] M. Annavaram, Q. Jacobson and J.P. Shen. HangOut: A Privacy Preserving Social Networking Application Invited paper. To Appear in *the workshop on Mobile Devices and Urban Sensing*, April, 2008.

[7] L. Barkhuus, and 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.

[8] L. Capra, W. Emmerich, and C. Mascolo. A micro-economic approach to conflict resolution in mobile computing. In *Proceedings of the 10th symposium on Foundations of software engineering*, pages 31–40, 2002.

[9] S. Consolvo, I.E. Smith, T. Matthews, A. LaMarca, J. Tabert, and 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*, pages 81–90, 2005.

[10] 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.

[11] L. Desmet, W. Joosen, F. Massacci, K. Naliuka, P. Philippaerts, F. Piessens, and D. Vanoverberghel. A flexible security architecture to support third-party applications on mobile devices. In *Proceedings of Workshop on Computer security architecture*, pages 19-28, 2007.

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

[13] 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).

[14] A.K. Ghosh and T.M. Swaminatha. Software security and privacy risks in mobile e-commerce. *Communications of ACM*, 44(2):51–57, 2001.

[15] M. Gruteser and 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*, pages 31–42, 2003.

[16] B. Hoh, M. Gruteser, M. Annavaram, Q. Jacobson, R. Herring, J. Ban, D. Work, J. Herrera, and A. Bayen. Virtual trip lines for distributed privacy-preserving traffic monitoring. To Appear in *Proceedings of the 6th international conference on Mobile systems, applications and services*, June, 2008.

[17] J.I. Hong, J.D. Ng, S. Lederer, and J.A. Landay. Privacy risk models for designing privacy-sensitive ubiquitous computing systems. In *Proceedings of the 5th conference on Designing interactive systems*, pages 91–100, 2004.

[18] J.I. Hong and J.A. Landay. An architecture for privacy-sensitive ubiquitous computing. In *Proceedings of the 2nd international conference on Mobile systems, applications, and services*, pages 177–189, 2004.

[19] G. Iachello, I. Smith, S. Consolvo, M. Chen, and G.D. Abowd. Developing privacy guidelines for social location disclosure applications and services. In *Proceedings of the 2005 symposium on Usable privacy and security*, pages 65–76, 2005.

[20] S. Lederer, J. Mankoff, and A.K. Dey. Who wants to know what when? privacy preference determinants in ubiquitous computing. In *extended abstracts on Human factors in computing systems*, pages 724–725, 2003.

[21] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. *l*-diversity: Privacy beyond k-anonymity. In *Proc. 22nd Intnl. Conf. Data Engg. (ICDE)*, page 24, 2006.

[22] E. Miluzzo, N. D. Lane, S. B. Eisenman and A. T. Campbell. CenceMe ÍC Injecting Sensing Presence into Social Networking Applications. In *the 2nd European Conference on Smart Sensing and Context, EuroSSC'07*, Lake District, UK

[23] M. Musolesi, E. Miluzzo, N.D. Lane, S. B. Eisenman, T. Choudhury, and A. T. Campbell. Integrating sensor presence into virtual worlds using mobile phones. In *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*, (Raleigh, NC, USA, November 05 - 07, 2008).

[24] M.F. Mokbel, C. Chow, and 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*, pages 763–774. 2006.

[25] S. Patil and A. Kobsa. Uncovering privacy attitudes and practices in instant messaging. In *Proceedings of the 2005 international conference on Supporting group work*, pages 109–112, 2005.

[26] S. Reddy, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. A framework for data quality and feedback in participatory sensing. In *Proceedings of the 5th international conference on Embedded networked sensor systems*, pages 417–418, 2007.

[27] 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.

[28] K. Shilton, N. Ramanathan, V. Samanta, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Participatory Design of Urban Sensing Networks: Strengths and Challenges. *Participatory Design Conference*, October 1-4, 2008, Bloomington, Indiana.

[29] I. Smith, S. Consolvo, J. Hightower, J. Hughes, G. Iachello, A. LaMarca, J. Scott, T. Sohn, and G. Abowd. Social Disclosure of Place: From Location Technology to Communication Practice. In *Proceedings of the International Conference on Pervasive Computing*, May 2005.

[30] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557ÍC570, 2002.

[31] J. Tang, V. Terziyan and J. Veijalainen. Distributed PIN verification scheme for improving security of mobile devices. In *Mobile Networks and Applications*, 8(2), pages 159–175, 2003.

[32] K.P. Tang, P. Keyani, J. Fogarty, and 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*, pages 93–102, 2006.