Toward Community Sensing

Andreas Krause[1] Carnegie Mellon
Eric Horvitz Microsoft Research
Aman Kansal Microsoft Research
Feng Zhao Microsoft Research

Abstract

A great opportunity exists to fuse information from populations of privately-held sensors to create useful sensing applications. For example, GPS devices, embedded in cellphones and automobiles, might one day be employed as distributed networks of velocity sensors for traffic monitoring and routing. Unfortunately, privacy and resource considerations limit access to such data streams. We describe principles of community sensing that offer mechanisms for sharing data from privately held sensors. The methods take into account the likely availability of sensors, the context-sensitive value of sensor information, based on models of phenomena and demand, and sensor owners' preferences about privacy and resource usage. We present efficient and well-characterized approximations of optimal sensing policies. We provide details on key principles of community sensing and highlight their use within a case study for road traffic monitoring.

1. Introduction

Privately owned sensors such as cameras, GPS devices, cell-phones, and home weather stations are bountiful in the world. In principle, the data from large populations of such sensors could be harnessed to provide valuable services. For example, GPS devices, which are becoming popular as integral components of smartphones and automobiles, could provide real-time traffic monitoring services with extensive coverage. In practice however, privately-held sensor data is rarely shared because of privacy concerns of the sensor owners or others whose privacy might be violated by such sensing. Beyond privacy considerations, applications depending on authorization via real-time requests for data could be disruptive and annoying to owners. Furthermore, owners may not wish to donate battery and networking resources required for sensing and transmitting data.

We focus on principles of *community sensing* that are aimed at unlocking the wealth of data and inferences available via privately held sensors and content. Our methods provide the machinery that allows sensing applications to make contracts with users about the resources used on their devices and the frequency and nature of incursions in privacy. We seek optimal sensor polling policies given the characteristics of the target phenomenon being sensed, the demands of the application, and availabilities of sensors, based on privacy preferences and expected locations.

By building on and extending structural results of sensor selection problems [12], we show how we can compute a well-characterized approximation to optimal sensing policies by acting in accordance with the preferences of sensor owners and considering the overall utility of sensed data for the sensing application. In our approach, sensed data is drawn from a population of sensors by computing the context-sensitive value of information, given statistical inferences about the availability and informativeness of different sensors. Personal data sharing policies are overlaid on the expected value as costs and constraints, allowing for mechanisms that exchange sensed data with tangible reimbursements of value to providers, such as access to aggregated data and inferences.

To motivate the core concepts of community sensing, we shall focus on a case study on the challenge of using privately held location sensors for traffic monitoring. Traffic sensing systems are growing in popularity in major cities via the use of fixed sensors deployed on major highways at significant expense. Beyond sensing and routing based on current traffic conditions, applications have been developed to characterize traffic [31] and perform forecasts about future traffic situations [8] contextual. Unfortunately, even when traffic sensing is available for key aspects of a highway system, there is typically little coverage of flows on important arterials and sidestreets. Several research and commercial prototypes have harnessed mobile *probe data* to obtain real-time or historical GPS flow information, often via special studies [17] or via contracts with the operators of GPS-monitored fleets [30]. Some services have attempted to leverage cell-tower signals for coarse positioning with mixed success [29].

GPS devices are becoming more popular as part of car-based and portable navigation systems and cellphones. The data collected by these GPS receivers are becoming increasingly more available via on-demand network connectivity.

[1]This work was performed during an internship at Microsoft Research.

In the absence of concerns about privacy and network resources, data about the driver's location and velocity might be readily shared with applications, via infrastructures such as SenseWeb [9]. However, privacy concerns are expected and reasonable [24, 30, 29, 20]. Beyond general anxieties about the sharing of location and velocity data, studies have demonstrated that, even with significant attempts at obfuscation, home and work locations of drivers can be inferred from GPS tracks [16]. Based on this consideration, in our case study we will focus on location privacy, i.e., favor sensing policies which avoid continuous monitoring of a user's location. Further, a sensor owner may wish to limit the amount of system resources used for such sensing and sharing, *e.g.*, to conserve battery power.

We develop community sensing methods that promise to unleash privately owned sensors for multiple sensing applications. Our main contributions are:

- An integrated approach to community sensing that jointly employs computations of the context-sensitive value of information for modeling phenomena, a model of the distribution of needs in a population, a forecast of the configuration of sensors, and constraints based on sharing preferences.
- A theoretical analysis of community sensing, which allows us to devise provably near-optimal sensing policies, maximizing the utility of the acquired information, while satisfying resource and privacy constraints.
- A case study applying those principles to a traffic monitoring problem.
- An empirical evaluation based on data from a deployed prototype.

2. Community Sensing Challenge

The key goal of community sensing is to continue to select the best subset of privately owned sensors so as to estimate a complex spatial phenomenon, in strict accordance with constraints on sensor usage. We consider the following tightly integrated factors to identify the best subset of sensors: We construct a *model for the phenomenon* being sensed. We rely on existing information about the phenomenon to make context-sensitive sensor selections that promise to provide the maximal value of information. Second, we consider the ultimate uses of data or inferences, and take a *utilitarian* approach: we assert that a sensing application should weight its information needs based on the expected demand for information by multiple people and organizations. Thus, we seek sets of observations which most improve those aspects of the phenomenon model that have the highest *demand*. Third, the availability and reliability of sensors may vary significantly, especially since the sensors are privately held rather than owned by the sensing application. Thus, we employ a principled approach for taking advantage of *uncertain sensor availability*. Finally, we overlay with care a *model of user preferences* about sensor access and resource usage.

2.1. Formalization of Community Sensing

Phenomenon modeling. We model the spatiotemporal phenomenon by a stochastic process, with a random variable $\mathcal{X}_s$ for each location[2] $s \in \mathcal{V}$. After observing values at a small number of locations $\mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}$, this process allows us to predict the phenomenon values at the unobserved locations $\mathcal{V} \setminus \mathcal{A}$, *e.g.,* by using conditional expectations $\mathbb{E}[\mathcal{X}_{\mathcal{V}\setminus\mathcal{A}} \mid \mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}]$. Hereby, for $\mathcal{A}' \subseteq \mathcal{V}$, we shall use the notation $\mathcal{X}_{\mathcal{A}'}$ in order to refer to the vector $(\mathcal{X}_s)_{s\in\mathcal{A}'}$ of random variables. As predictions are uncertain, we use our model to predict the variance at each location $s \in \mathcal{V}\setminus\mathcal{A}$,

$$\mathrm{Var}(\mathcal{X}_s \,|\, \mathcal{X}_\mathcal{A}{=}\mathbf{x}_\mathcal{A}) = \mathbb{E}[(\mathcal{X}_s{-}\mathbb{E}[\mathcal{X}_s \,|\, \mathcal{X}_\mathcal{A}{=}\mathbf{x}_\mathcal{A}])^2 \,|\, \mathcal{X}_\mathcal{A}{=}\mathbf{x}_\mathcal{A}].$$

To quantify the value of the sensor locations, we use the *reduction* in the predicted variance,

$$\mathrm{Var}(\mathcal{X}_s) - \mathrm{Var}(\mathcal{X}_s \mid \mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}),$$

after observing $\mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}$. Based on the phenomenon model alone, we could select a set $\mathcal{A}$ of locations which maximally reduce the variance at the unobserved locations.

Demand modeling. Instead, in order to ensure that predictions are most accurate where they are needed most [29], we take a utilitarian approach to compute the information value of sensing at any set $\mathcal{A}$ of locations. Hence, we aim to achieve the highest reduction in variances at locations s which are most frequently queried. More formally, we define a non-negative spatial process, $\mathcal{D}_s$, called the *demand process*, over all locations $s \in \mathcal{V}$. We can then consider the expected *demand-weighted* variance reduction,

$$R(\mathcal{A}) = \sum_{s\in\mathcal{V}} \mathbb{E}\left[\mathcal{D}_s(\mathrm{Var}(\mathcal{X}_s) - \mathrm{Var}(\mathcal{X}_s \mid \mathcal{X}_\mathcal{A}))\right]$$

The expectation is taken with respect to the observations $\mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}$ and the demands $\mathcal{D}_s = d_s$.

Models of availability and privacy. When making decisions about polling the community of sensors, we will generally *not* be able to sample the phenomenon at locations $\mathcal{A}$ directly, as we may not have access to sensors at these locations because of uncertainty regarding the current sensor locations. Thus we must additionally learn and integrate models of the uncertainty in sensor availability, which will be a key focus of this paper.

Instead of choosing locations to observe, we assume that we will have a set $\mathcal{W}$ of possible *observations* we can choose among. Each observation might, *e.g.*, correspond to a person porting a GPS system walking through a city, or a GPS-ready car traveling through a road network. Hence,

[2] In the case of spatiotemporal planning, we would associate a random variable with each location-time pair for a set of time points (*c.f.*, [21]).

each observation $w \in \mathcal{W}$ corresponds to a *distribution* over possible sensor locations, and any particular subset $\mathcal{B} \subseteq \mathcal{W}$, induces a distribution $P(\mathcal{A} \mid \mathcal{B})$ over subsets $\mathcal{A}$. The model of sensor availability incorporates a notion of *selection noise*. Also, note that we distinguish between the possible observations $\mathcal{W}$ we can select from, and the *measurements* obtained after a subset $\mathcal{B} \subseteq \mathcal{W}$ of observations has been selected.

Our final informational objective to maximize is

$$F(\mathcal{B}) = \mathbb{E}_{\mathcal{A}|\mathcal{B}}[R(\mathcal{A})] = \sum_{\mathcal{A}} P(\mathcal{A} \mid \mathcal{B}) R(\mathcal{A}),$$

i.e., the expected demand-weighted variance reduction, where the expectation is taken over the set of observed locations $\mathcal{A}$, the measurements at these locations $\mathcal{X}_{\mathcal{A}} = \mathbf{x}_{\mathcal{A}}$, and the demands $\mathcal{D}_{\mathcal{V}}$ at all locations[3]. This utility function $F(\mathcal{B})$ effectively quantifies the expected *value of information* of the observations $\mathcal{B}$, which is *context-sensitive* and specific to the particular application at hand.

In addition to uncertainty which is *inherent* in the community sensing application (e.g., the location of a car and GPS sensor is uncertain at query time), we will explore methods for artificially *introducing* selection noise in order to achieve privacy. Since privacy is a highly application specific notion, we will not attempt to define a general notion of privacy for community sensing. Instead, in Section 4.3, we describe a privacy-aware community sensing approach for the traffic monitoring problem. We explore two notions of achieving privacy through injection of selection noise. The first approach achieves location privacy by *spatial obfuscation*, a technique that introduces selection noise. The second approach is based on *sparse querying*, where the number of queries to each sensor owner is minimized.

Models of resource usage and preferences. Central in a model of community sensing is the coupling of the overall information utility to an application, with constraints defined by users on sensor sharing prefences and resource usage. We achieve this coupling by introducing a cost function C which associates each set $\mathcal{B}$ of observations with a nonnegative cost $C(\mathcal{B})$.

For example, *resource usage* can be modeled by associating a nonnegative cost $c(s)$ with each observation $s \in \mathcal{W}$, and defining

$$C(\mathcal{B}) = \sum_{s \in \mathcal{B}} c(s).$$

When we consider each reading to have unit cost, *i.e.,* $c(s) = 1$, the cost $C(\mathcal{B}) = |\mathcal{B}|$ is equal to the number of selected sensors. This cost model aims to minimize the number of collected readings over the whole community sensor network, or effectively, the network load.

[3]Note that this formulation implies the assumption that the selection noise is independent of demand and phenomenon, which we make for clarity of presentation in this paper.

By defining more complex cost functions, we can make community sensing conform to more general, expressive policies, including individualized preferences. Such policies can include specification about the minimal inter-probe interval, the maximum number of probes per unit of time, and the allowed times or locations of probes (*e.g.*, sensor device ignores probes when the owner is within specified range of some location, such as their home or office). Other preferences might include the requirement that the owner authorize probes before information is transmitted.

For example, we might limit the number of queries to individual users based on their preferences. In this case, we can partition the set of possible observations $\mathcal{W}$ into a collection of subsets $\mathcal{W}_1 \cup \cdots \cup \mathcal{W}_m$, each $\mathcal{W}_i$ corresponding to all possible observations which could be made by continuously monitoring user i. We can then, for example, model cases where the cost of querying a user increases as more and more observations are made:

$$C(\mathcal{B}) = \sum_i g_i(|\mathcal{B} \cap \mathcal{W}_i|), \tag{2.1}$$

where g_i are monotonically increasing convex functions with $g_i(\emptyset) = 0$, and $g_i(\ell)$ models the cost of querying user i exactly ℓ times. We can also require that each user be queried at most ℓ times by setting

$$g_i(j) = 0 \text{ if } j \leq \ell, \infty \text{ otherwise.} \tag{2.2}$$

2.2. Problem Specification

Based on the components detailed above, our goal is to select a set of observations $\mathcal{B}^*$ such that

$$\mathcal{B}^* = \operatorname*{argmax}_{\mathcal{B}} F(\mathcal{B}) \text{ subject to } C(\mathcal{B}) \leq L. \tag{2.3}$$

Hereby, L specifies the budget that we can spend on observations (e.g., applications "pay" owners on a per probe basis). The solution set $\mathcal{B}^*$ could, for instance, determine the set of GPS-equipped automobiles considered by the system. By requiring that the cost satisfies the constraint $C(\mathcal{B}) \leq L$ (or other constraints or cost functions per more general preference models for probing), *contracts* are implemented between the sensing application and the sensor owners.

For clarity of presentation, we will focus on the *static* selection problem. In this setting, we only consider the state of the world (such as road speeds) at the current time step, and do not plan ahead to select observations which may become useful in the future. However, techniques similar to those presented by Meliou et al. [21] can be used to extend our approach to the dynamic, spatiotemporal planning setting.

3. Optimizing Community Sensing

We now present algorithms for the community sensing optimization problem specified in Section 2.2. Our goal will

be to obtain measurements from a subset of the available observations to optimize value of information subject to preference and resource constraints discussed earlier. Problem (2.3) requires solving a difficult (NP-hard) discrete optimization problem, and finding the exact solution is typically intractable. In this section, we will prove that Problem (2.3) carries structure, which will allow us to obtain provably near-optimal solutions.

Let us first consider Problem (2.3) for the unit cost case $C(\mathcal{B}) = |\mathcal{B}|$ and the budget $L = k$. We shall later discuss how more complex constraints can also be handled in the optimization. Rather than search through all prohibitively many $\binom{n}{k}$ subsets, we use a greedy algorithm for selecting the k best observations:

Algorithm 3.1. *Start with the empty set of selected observations, $\mathcal{B}_0 = \emptyset$. In the j-iteration, find the observation*

$$w_j = \operatorname*{argmax}_{w \in \mathcal{W}} F(\mathcal{B}_{j-1} \cup \{w\}),$$

and then set $\mathcal{B}_j := \mathcal{B}_{j-1} \cup \{w_j\}$. Continue to iteratively expand the set $\mathcal{B}_j$ until $|\mathcal{B}_j| = k$. Output $\mathcal{B}_G = \mathcal{B}_k$.

Performance Guarantee. We can show that this algorithm is guaranteed to find a set $\mathcal{B}_G$ which is close to the optimal solution. The key to analyzing the performance of the greedy algorithm is the following intuition: adding an observation helps more if we have made few observations so far, and less, if we already have made many observations. This intuition is formalized by the combinatorial concept of submodularity. A set function $f : 2^{\mathcal{W}} \to \mathbb{R}$ is called *submodular*, if for all $\mathcal{B} \subseteq \mathcal{B}' \subseteq \mathcal{W}$ and $w \in \mathcal{W} \setminus \mathcal{B}'$ it holds that $f(\mathcal{B} \cup \{w\}) - f(\mathcal{B}) \geq f(\mathcal{B}' \cup \{w\}) - f(\mathcal{B}')$. Also, f is called *nondecreasing*, if for all $\mathcal{B} \subseteq \mathcal{B}' \subseteq \mathcal{W}$ it holds that $f(\mathcal{B}) \leq f(\mathcal{B}') \leq f(\mathcal{W})$. A fundamental result by Nemhauser et.al. [22] states that the set $\mathcal{B}_G$ of size k obtained by the greedy algorithm applied to some nondecreasing submodular function f is guaranteed to achieve at least a constant fraction $(1 - 1/e) \approx 0.63$ of the optimal solution:

$$f(\mathcal{B}_G) \geq (1 - 1/e) \max_{|\mathcal{B}| \leq k} f(\mathcal{B}).$$

This performance guarantee will apply to Algorithm 3.1 if the objective function $F(\mathcal{B})$ used in the algorithm is submodular. In the literature, submodularity has been established for several important objective functions (*c.f.*, [12] for an overview). In [4], it was shown, that under a natural condition of *conditional suppressor-freeness*, which is often satisfied in practice, the expected local variance reduction $\mathbb{E}[\mathrm{Var}(\mathcal{X}_s) - \mathrm{Var}(\mathcal{X}_s \mid \mathcal{X}_\mathcal{A})]$ is nondecreasing and submodular in $\mathcal{A}$. The key question is thus whether the complex objective function $F(\mathcal{B})$, integrating phenomenon, demand and availability models, is still submodular. In the following, we establish this result, which will allow us to unlock a wealth of techniques for optimizing submodular functions for the community sensing domain.

We first observe that, since the class of submodular functions is closed under nonnegative linear combinations, the demand-weighted variance reduction $R(\mathcal{A})$ is nondecreasing and submodular as well. Now, to establish submodularity of $F(\mathcal{B})$, the key step is to ensure that the introduction of uncertainty in sensor availability, *i.e.*, taking the expectation of $R(\mathcal{A})$ over locations $\mathcal{A}$, does not diminish submodularity of $F(\mathcal{B})$. The following lemma establishes this result.

Lemma 3.1. *Whenever $R(\mathcal{A})$ is submodular and nondecreasing, then $F(\mathcal{B}) = \mathbb{E}_{\mathcal{A}|\mathcal{B}}[R(\mathcal{A})]$ is submodular and nondecreasing.*

The proof of this Lemma is provided in our technical report [15]. Lemma 3.1 proves that the introduction of selection noise does not destroy submodularity. However, even though our value of information objective $F(\mathcal{B})$ is submodular, it is not clear how we can efficiently implement the Algorithm 3.1: First of all we need to be able to efficiently evaluate the demand weighted variance reduction $R(\mathcal{A})$. In Section 4, we show how this computation is possible in closed form in many important applications. However, even if we can efficiently compute $R(\mathcal{A})$, in order to evaluate $F(\mathcal{B})$, we need to compute the expectation

$$\mathbb{E}_{\mathcal{A}|\mathcal{B}}[R(\mathcal{A})] = \sum_{\mathcal{A}} P(\mathcal{A} \mid \mathcal{B}) R(\mathcal{A}),$$

which is a sum over an intractable number of terms. However, we can approximate this expectation by Monte Carlo sampling. For any set $\mathcal{B}$ of observations, we draw N independent samples $\mathcal{A}_1, \ldots, \mathcal{A}_N$ from the distribution[4] $P(\mathcal{A} \mid \mathcal{B})$, and approximate $F(\mathcal{B})$ by its sample average, *i.e.*,

$$F(\mathcal{B}) \approx \frac{1}{N} \sum_{i=1}^{N} R(\mathcal{A}_i).$$

As $R(\mathcal{A})$ is bounded between 0 and 1, we can use Hoeffding's inequality to bound the number of samples we need in order to approximate $F(\mathcal{B})$ to required precision ε with probability at least $1 - \delta$:

Lemma 3.2. *For any $\varepsilon > 0$ and $\delta > 0$, we need*

$$\left\lceil \frac{1}{2\varepsilon^2} \log \frac{1}{\delta} \right\rceil$$

samples in order to ensure that

$$P\left(\left| F(\mathcal{B}) - \sum_{i=1}^{N} R(\mathcal{A}_i) \right| \leq \varepsilon \right) \geq 1 - \delta.$$

Using a similar analysis as presented in [11], we can show that the result of Nemhauser *et.al.* about the performance of the greedy algorithm still holds, even if the objective function F is evaluated with small additive error ε:

[4]For many practical settings, e.g., if the locations of each sensor is independent, this sampling is tractable.

Theorem 3.3. *For any $\varepsilon > 0$ and $\delta > 0$, under the conditions described in [4] (or with any submodular, nondecreasing objective function $R(\mathcal{A})$ bounded above by 1) and k allowed observations:*
Accuracy: *The greedy solution $\mathcal{B}_G$ satisfies*

$$F(\mathcal{B}_G) \geq (1 - 1/e) \max_{|\mathcal{B}|=k} F(\mathcal{B}) - \varepsilon,$$

with probability at least $1 - \delta$.
Computation Load:*The greedy algorithm finds this solution $\mathcal{B}_G$ using at most*

$$\mathcal{O}\left(k^3 n \frac{1}{\varepsilon^2} \log \frac{kn}{\delta}\right)$$

independent samples of $R(\mathcal{A})$.

The proof of this Theorem is provided in our technical report [15]. Note, that the number of independent samples of $R(\mathcal{A})$ needed for estimating the objective $F(\mathcal{B})$ (*c.f.*, Lemma 3.2) grows only as $\mathcal{O}(n \log n)$ with the number $n = |\mathcal{W}|$ as possible observations to select from. In practice, as seen in the next section, k is much smaller than n, and only a small number of observations are typically needed to satisfy the sensing application.

Handling more complex cost functions. In the above discussion, we focused on the challenge of selecting the best k observations. In practice, we can use the submodularity of our value of information objective $F(\mathcal{B})$ as established in Lemma 3.1 to near-optimally solve much more complex optimization problems.

In practice, different observations could have different costs. For example, one might be able to choose between placing high-quality fixed sensors at a higher cost, or query via the community sensing system at lower cost, but at the risk of uncertain availability, as considered in Section 5.4. Similarly, requesting different types of contributions from privately held sensors (e.g., audio, video imagery, etc.) might require providing different incentives to users. In this case, the cost $c(w)$ of each observation would vary, and the system would have a constraint on the budget which can be spent on observations. An algorithm similar to the one presented in [19] can be used to solve this more complex optimization problem.

As discussed in Section 2, we can employ more general user preferences, such as a policy that asserts that each user can be queried only ℓ times within a specified time window (*c.f.*, Eqn. (2.2)). In this case, the optimization problem is

$$\mathcal{B}^* = \operatorname*{argmax}_{\mathcal{B}} F(\mathcal{B}) \text{ s.t. } |\mathcal{B}| \leq k \text{ and } \forall i : |\mathcal{B} \cap \mathcal{W}_i| \leq \ell,$$

where $\mathcal{W}_i$ is the set of observations which could potentially be made by continuously monitoring user i. It can be shown that the set $\mathcal{I} \equiv \{\mathcal{B} \subseteq \mathcal{V} : |\mathcal{B}| \leq k \text{ and } \forall i : |\mathcal{B} \cap \mathcal{W}_i| \leq \ell\}$ defines independent sets of a *matroid*. In [6], it is shown that a greedy algorithm provides a $\frac{1}{2}$ approximation even in this more general setting, i.e., the greedy solution $\mathcal{B}_G$ satisfies

$$F(\mathcal{B}_G) \geq \frac{1}{2} \max_{\mathcal{B} \in \mathcal{I}} F(\mathcal{B}).$$

In Section 2, we also considered the case where the cost of querying a user can increase nonlinearly in the number of queries (2.1). Notice that, in contrast to the diminishing returns property of our value of information objective, this cost function has an *accelerating cost* property: Adding a new observation increases the cost the more observations we have already made. In fact, we can show:

Lemma 3.4. *The cost function* (2.1) *is* supermodular.

A set function $C(\mathcal{A})$ is called supermodular if and only if $-C(\mathcal{A})$ is submodular. This observation allows us to use techniques described by Krause and Horvitz [14] for finding a near-optimal solution to the community sensing problem.

4. Case Study: Traffic Monitoring

Let us now consider in more detail the community sensing application for traffic introduced in Section 1. In order to instantiate the community sensing problem for traffic, we need to choose appropriate models for the measured phenomenon, application demand, sensor availability, and probing cost. The application goal is to provide normalized road speeds (*i.e.*, average speed of vehicles across road segments, divided by the posted speed limits for those segments) over road segments within the sensed road network.

4.1. Phenomenon Model

We model the joint distribution of the normalized road speeds $\mathcal{X}_s$ for all locations $s \in \mathcal{V}$ using a Gaussian Process (GP, *c.f.*, [25]), defined over the road network[5]. Such a model is fully specified by a mean function $\mathcal{M}(s)$ and covariance function $\mathcal{K}(s,t)$. For each set $\mathcal{A} = \{s_1, \ldots, s_n\} \subseteq \mathcal{V}$ of locations, this GP induces a multivariate normal distribution $\mathcal{X}_\mathcal{A} \sim \mathcal{N}(\mu_\mathcal{A}; \Sigma_{\mathcal{A}\mathcal{A}})$, where $\mu_\mathcal{A}$ is the prior mean vector $\mu_\mathcal{A} = (\mathcal{M}(s_1), \ldots, \mathcal{M}(s_n))$, and $\Sigma_{\mathcal{A}\mathcal{A}} = (\mathcal{K}(s_i, s_j))_{i,j}$ is the prior covariance matrix, obtained by evaluating the kernel function at all pairs of points. In Section 5, we study the fit of the GP to the traffic data, and explain how we can estimate the kernel from training data. Based on this phenomenon model, we can evaluate the variance reduction in closed form [25]:

$$\mathrm{Var}(\mathcal{X}_s) - \mathrm{Var}(\mathcal{X}_s \mid \mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}) = \Sigma_{s\mathcal{A}} \Sigma_{\mathcal{A}\mathcal{A}}^{-1} \Sigma_{\mathcal{A}s} \quad (4.1)$$

Here, $\Sigma_{s\mathcal{A}} = (\mathcal{K}(s, s_1), \ldots, \mathcal{K}(s, s_n))$ is the vector of cross-correlations between $\mathcal{X}_s$ and $\mathcal{X}_\mathcal{A}$. Note that (4.1) does *not* depend on the observed values $\mathbf{x}_\mathcal{A}$, an interesting property of GPs. The expected variance reduction $\mathbb{E}[\mathrm{Var}(\mathcal{X}_s) - \mathrm{Var}(\mathcal{X}_s \mid \mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A})]$ can thus be evaluated in closed form.

[5]Normal distributions have been considered for traffic modeling by [2].

4.2. Demand Model

For each road segment $s \in \mathcal{V}$, we define its *demand* $\mathcal{D}_s$ by the number of cars traveling over each road segment and model it using a Poisson random variable, with mean λ_s. Even though, in practice, demand and phenomenon are correlated, for computational considerations, we shall assume independence. Hence,

$$R(\mathcal{A}) = \sum_{s \in \mathcal{V}} \mathbb{E}\left[\mathcal{D}_s\right] \mathbb{E}\left[(\mathrm{Var}(\mathcal{X}_s) - \mathrm{Var}(\mathcal{X}_s \mid \mathcal{X}_\mathcal{A}))\right]$$
$$= \sum_{s \in \mathcal{V}} \lambda_s \Sigma_{s\mathcal{A}} \Sigma_{\mathcal{A}\mathcal{A}}^{-1} \Sigma_{\mathcal{A}s},$$

which can be computed efficiently in closed form, in time $\mathcal{O}(nk^2 + k^3)$, where $k = |\mathcal{A}|$ and $n = |\mathcal{V}|$. Note, that when selecting observations, k is much smaller than n. It can be shown that $R(\mathcal{A})$ is nondecreasing in $\mathcal{A}$, i.e., for $\mathcal{A} \subseteq \mathcal{A}' \subseteq \mathcal{V}$, it holds that $R(\mathcal{A}) \leq R(\mathcal{A}') \leq R(\mathcal{V})$. We can thus normalize $R(\mathcal{A})$ to be between 0 and 1 by dividing by $R(\mathcal{V})$. Based on this objective, we can identify the observations of road segments that promise to most efficiently improve the demand-weighted prediction accuracy.

4.3. Models of Availability and Privacy

Due to the nature of the traffic monitoring application, selection noise is inherent in the community sensing problem. Even if we continuously monitor a user (car), there is uncertainty in their motion and location. For each possible observation $b \in \mathcal{B}$ (such as each car we can query), we model its *availability distribution* (e.g., distribution over the car's location) as a multinomial distribution. In principle, we could estimate such a distribution from training data gathered from continuously monitoring the users. However, such monitoring would raise reasonable privacy concerns. Instead, our goal is to devise sensing policies which guard against such possible incursions in privacy.

Several different models of privacy have been proposed in prior work (*c.f.*, [28, 5]). Rigorously formalizing privacy requires strong assumptions about the capabilities of the adversary who attempts the intrusion in privacy. In the traffic monitoring domain, our goal is to develop a privacy-aware community sensing approach which guards against incursions in location privacy (e.g., identification of a person's home) through inference attacks as considered by Krumm [16]. In his study, Krumm found that such inference attacks can be made difficult by artificially introducing a large amount of noise to the sensor locations. Motivated by these findings, we explore techniques for *introducing* selection noise (i.e., artificially increase the uncertainty in the availability distributions $P(\cdot \mid b)$). We give two concrete examples for such availability distributions. Both sample availability models are sensitive to privacy considerations as they avoid constant monitoring of all sensor locations.

Spatial obfuscation: In a spatial obfuscation approach, we minimize privacy intrusion by giving up the ability to specifically select and address an individual sensor. Instead, we divide the space into a set of cells, and only specify which region we want to query. The sensor is then selected according to a probability distribution defined over all sensors associated with the cell. Once selected, the sensor would reveal its exact location as well as its measurements, but it would not reveal any identifying information. In addition to being sensitive to privacy considerations, the spatial obfuscation approach has computational advantages, in that optimization is computed for a smaller set of potential observations.

In the traffic monitoring application, such an approach could be realized by employing a trusted arbitrator. A possible example would be a cell phone network provider, who is already required by law to monitor locations for emergency response purposes. This arbitrator could perform the sampling (i.e., the random selection of a car in a network cell), and then, via cryptographic techniques, transmit the obtained measurements without any identifying information to the traffic monitoring service.

Sparse querying: Another approach is to directly keep track of a pool of drivers volunteering to participate in the traffic service. However, instead of constantly monitoring each driver, which would be highly privacy invasive, we only very sparsely and infrequently pose queries. In this model, we do not initially know where a car is located when it is queried. Once it receives the query, it will provide its location along with its measured road speed. Based on previous queries and the provided location information, for each driver a time-dependent availability distribution can be estimated. This distribution can then be used to inform the decision of which cars should be queried. Alternately, owners of sensors may allow a service to monitor them for a predefined monitoring phase and, thus, learn a predictive model in advance. Techniques such as those described in [14] could be employed in order to avoid sampling which would lead to too accurate predictive models of the location, by requiring a minimum amount of uncertainty in the availability distribution $P(\cdot \mid \mathcal{B})$.

Such a sparse monitoring approach could be implemented by allowing the user to specify preferences about access to their sensor data (i.e., a limit on the number of queries per week as described in Section 2, and the exclusion of sending measurements in proximity to certain private locations using a spatial cloaking approach as described by Krumm [16]).

5. Experiments

The following data sets were used for the traffic monitoring case study.

Phenomenon Data: In our experiments, the phenomenon being measured by the sensing application is the set of road speeds on Greater Seattle Area highways. We use real traffic data from 534 highway segments distributed across interstates I-5, I-405, I-90 and state highways 520, 518, 167 and 525, using sensors deployed by Washington Dept. of Transportation. Our goal is to model the normalized road speeds. For each road segment, the normalized road speed is defined as the ratio of average speed measured by sensors on that road segment to the posted speed limit. We use sensor data logged between October 2006 to December 2006, in 15 minute time steps. We use the first 2/3 of the data as training set, and the remaining 1/3 as a test set. Based on the training data, we used the sample mean and covariance to define our covariance model. In Section 5.4, we explain how a model can be learned from less training data as well. In our analysis, we limited ourselves to highway segments as historical speed data is available for them, and we can compare the phenomenon model generated by our sensing algorithm to the ground truth.

Demand Data: In order to estimate the demand model, we use 3166 route planning requests obtained from users of a context-sensitive routing prototype used by volunteers at Microsoft during 2006 and 2007. Users input a start and destination and the system provides a best route. We randomly split the route requests using 2/3 of the data as a training library and holding out 1/3 of the data for testing. We decompose each recommended route into a set of highway segments. For each of the 534 chosen highway segments, we count the number of recommended routes in which it is contained, for both the training and test sets. Based on these counts, we estimate the parameters of the demand distribution. Figure 2(c) presents an example of the demand process during rush-hour.

Availability Data: As explained in Section 4.3, we consider two separate availability models. For the *spatial obfuscation* model, we discretize the road network into a set $\mathcal{W}_c$ of cells of varying diameter Δ, according to their geographic location (latitude / longitude). Potential observations correspond to the possible cells $w \in \mathcal{W}_c$. For each cell, we define the availability distribution $P(\cdot \mid w)$ as the uniform distribution over all road segments contained in w.

Our *sparse querying* model is based on three years of GPS traces from the Microsoft Multiperson Location Survey (MSMLS) [17]: 252 voluntary drivers used a Garmin Geko 201 GPS receiver to record 10,000 timestamped GPS location readings when driving. We considered only those traces intersecting with the 534 chosen highway segments. Furthermore, we only considered traces from users for whom data is available for at least 6 days, 85 in total. For each user, we select the first 2/3 of the days as training, and hold out the last 1/3 for testing. For the training and test set, we estimate the multinomial parameters of the availability distributions by counting how often a given user was present on each road segment.

5.1. Experimental Setup and Evaluation

To evaluate the community sensing algorithm, we use the following methodology. We first train the phenomenon, demand, and availability models based on the training data, as described above. These models are then provided to the sensing algorithm 3.1. The algorithm is then executed for each time step for which the ground truth data is available, *i.e.*, the time instances corresponding to the last $1/3$ of each data set involved in the experiment. The algorithm yields a set $\mathcal{B}$ of sensors selected for observation. A real-world community sensing infrastructure for collection and aggregation of sensor data would seek to probe these sensors. However, the actual readings obtained would depend on sensor availability. For each set $\mathcal{B} \subseteq \mathcal{W}$ of selected observations, for a given test sample $\mathbf{x}_\mathcal{V}$ describing the normalized speeds of all road segments at the current time step of the highway data, we draw a set $\mathcal{A}$ of road segments according to the availability distribution $P(\mathcal{A} \mid \mathcal{B})$ estimated from the availability test set. The set $\mathcal{A}$ forms the set of road speed sensor readings actually obtained.

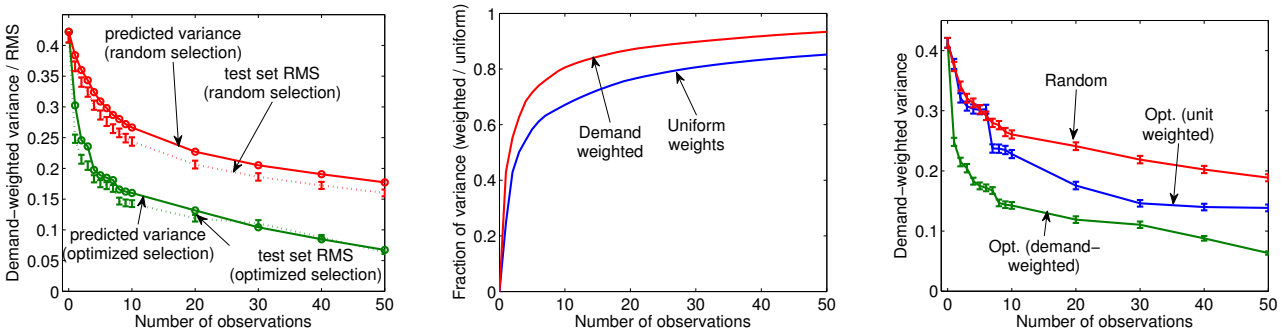
To evaluate the performance of the sensing algorithm based on the actual demand and sensor availability, we compare the phenomenon values predicted (using the phenomenon model) based on the small number of observations obtained through the sensing algorithm with the ground truth phenomenon values. The sensor probe budget is varied by changing the number of observations that our algorithm is allowed to make. The error metric used for the comparison is the *demand-weighted expected RMS error*, computed as follows: We predict the normalized road speeds on the unobserved road segments using the phenomenon model, and, for each road segment, then compute the residual $r_s = \mathbf{x}_s - \mathbb{E}[\mathcal{X}_s \mid \mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}]$ as the difference between true and predicted speeds. We then compute the demand-weighted root mean squared error,

$$DRMS = \sqrt{\frac{1}{n}\sum_s \lambda'_s r_s^2},$$

where λ'_s are the demands estimated from the test set. Finally, we report the mean $DRMS$ error where we average over all test samples (*i.e.*, all of time steps in the hold-out set), and use $\widehat{F}(\mathcal{B})$ to refer to the reduction of this average error.

5.2. Experiments without selection noise

In our first set of experiments, we explore the phenomenon and demand modeling only. We choose the road segments deterministically, *i.e.*, we set $\mathcal{A} = \mathcal{B}$, and hence, $F(\mathcal{B}) = R(\mathcal{B})$.



(a) *Comparing $F(\mathcal{B})$ and $\widehat{F}(\mathcal{B})$* (b) *Uniform vs. weighted RMS* (c) *Optimizing for different metrics*

Figure 1. Experiments without selection noise. (a) compares predicted with measured error as more observations are selected. (b) compares the scores for greedily chosen sets of increasing size, when optimizing uniform vs. demand-weighted variance reduction. Both objectives show strong diminishing-returns behavior. (c) compares reduction of demand-weighted error if road segments are chosen at random and when optimizing uniform and demand-weighted variance reduction.

Evaluation of model accuracy: In the first experiment, we verify how well the predicted scores $F(\mathcal{B})$ match the test set error $\widehat{F}(\mathcal{B})$. We generate observation selections $\mathcal{B}$ of increasing size. Figure 1(a) presents the results of this experiment, where the observations are selected naïvely: uniformly at random (the red curves, using 1000 independent trials) and using our optimized sensing algorithm. We can see that in both cases, the predicted error $F(\mathcal{B})$ (indicated by the solid lines) and the measured test set error $\widehat{F}(\mathcal{B})$ (indicated by the dotted lines) match very well. This result indicates, that our models do not overfit the training data, and the objective function $F(\mathcal{B})$ models the generalization error $\widehat{F}(\mathcal{B})$ effectively.

We also observe that the optimized selection reduces the prediction error significantly more quickly than the uniform sampling. Using optimized selection, the error is reduced by 50% compared to the baseline when using only 5 observations. Uniform sampling requires 40 observations in order to achieve the same performance.

Demand-weighted vs. uniform coverage: In our next experiment, we compare the effects of including demand on the coverage metric. We optimize one sequence of observation selections of increasing size using the training set demands, and another sequence setting the demands uniformly to 1 over all road segments. Figure 1(b) presents the corresponding production curves. Both curves are normalized between 0 and 1, where 1 means that the error has been reduced to zero. We can see that, for the demand-weighted objective, the coverage approaches its maximum far more quickly. This is encouraging, as it implies that far fewer observations are needed to achieve a specified level of coverage. For example, in the demand-weighted model, 10 observations achieve 80% reduction in error. Under the uniform metric, 30 observations are needed to achieve the same reduction in error. The concavity of the curves in Figure 1(b) also empirically corroborates the submodularity of the variance reduction: the scores $F(\mathcal{B})$ initially increase rapidly, and then quickly flatten out.

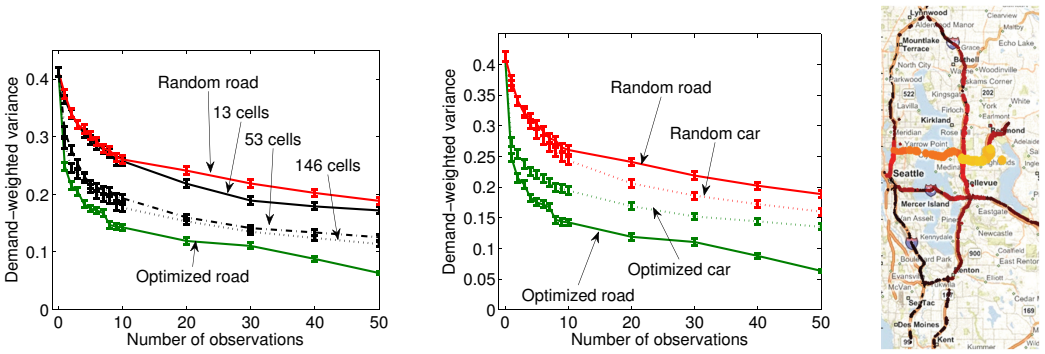
We also compare how well the demand-weighted error is reduced when optimizing for the uniform variance reduction as specified in Equation (4.1) instead of for $F(\mathcal{B})$. Figure 1(c) shows, that, although optimizing for the uniform variance reduction, shown by curve Opt. (unit weighted), reduces the demand-weighted error more quickly than random sensor selection, it decreases the error far more slowly than when using the demand model for sensor selection, shown by curve Opt. (demand weighted). The biggest difference appears in the initial part of the curve, in which the error is reduced most quickly. This suggests that, for this application, demand models should be used if available.

5.3. Experiments with selection noise

In the next series of experiments, we consider the different sources of selection noise, as introduced in Section 4.3.

Spatial obfuscation: We first consider the spatial obfuscation model. In this model, we discretize the space into a collection of cells, and we decide which cells we should query. The availability distributions are chosen as uniform over all road segments intersecting with each cell. Note, that a long road segment can be contained in several cells.

We first generated discretizations of varying granularity, by varying the diameter of the cells. The different versions of the problem contain 13 (most coarse), 53 (intermediate), and 146 (finest) cells. For each granularity, we generate sequences of observation selections, where we allow for querying the same cell multiple times, in which case multiple road segments are selected uniformly and independently. Figure 2(a) presents the results of this experiment.



(a) *Spatial obfuscation* (b) *Sparse querying* (c) *Demands)*

Figure 2. (a) shows of the influence of the spatial obfuscation on error reduction. Even relatively coarse discretization allows drastic improvements over purely random selection. (b) shows impact of sparse querying, i.e., compares selecting cars vs. road segments. (c) Example of demands during rush hour (higher intensity indicates higher demand).

As we decrease the amount of spatial obfuscation, the performance of the optimized observation selection changes from the uniform random sampling (corresponding to the case of 1 single cell) to fully optimized selection (corresponding to 534 cells, one for each road segment). We can see, that, while coarse discretization (13 cells) does not perform much better than random, even the intermediate discretization (53 cells) performs quite well. In the important steep initial part of the curve, the spatial obfuscation scheme performs only slightly worse than the fully optimized selection. The analysis with the finest discretization (146 cells) does not perform significantly better than the one using the intermediate discretization. We believe the reason for this effect is that road segments are contained in multiple cells according to their spatial extent, and hence, in the finer discretizations contain a significant amount of randomness.

The results of these experiments have several implications. First of all, the results indicate that, for this sample application, the computationally more intensive reasoning at the road-segment level of granularity does not provide much benefit over reasoning over a coarser spatial discretization. This is encouraging, as it indicates that the approach presented can likely be scaled up to consider significantly larger geographies. Secondly, from a privacy point of few, a priori knowledge about precise locations of cars is not necessary, and strong spatial obfuscation (as suggested in [16] as one possible countermeasure to inference attacks on GPS traces), does not severely limit the value of these sensors for traffic monitoring applications.

Sparse querying: Our second model for availability implements the notion of sparse querying. Here, each observation $w \in \mathcal{W}_p$ corresponds to a participant of the MSMLS study. For each participant ("car"), we estimate a training and test availability distribution $P(v \mid w)$ over all road segments $v \in \mathcal{V}$. Based on the training distributions, our goal is to select a subset of the participants to query. Figure 2(b) shows how the test set error decreases, as we select more and more cars. We compare optimized selection vs. uniform random selection, both of cars and specific road segments. Figure 2(b) shows that, in the initial, steep part of the curve, the difference between optimized selection and random sampling is most drastic. For example, in order to reduce the error by 50% over the baseline, we need 10 queries when optimizing the selection, and 30 queries when sampling at random. We also notice that random selection of cars performs better than randomly selecting road segments. The reason for this effect is correlation between demand and availability—on roads with higher demand, users are more likely present. Furthermore, as we increase the number of observations, the discrepancy between selecting roads and cars increases, and the discrepancy between optimized and random selection of cars decreases. We speculate that the reason for this effect is that the availability models are limited by the MSMLS data.

5.4. Further experiments

Learning curve: In the previous experiments, we used historical data from all 534 highway sensors in order to estimate the phenomenon model. In many community sensing scenarios, such dense instrumentation is not available, and a model would have to be learned from a small amount of training data. In order to understand how the prediction performance of chosen observations depends on the amount of training data used to learn the phenomenon model, we performed an experiment. Intuitively, the correlation of road speeds at segments s and t should depend on the geodesic (i.e., shortest path) distance of these segments with respect to the road network topology. In spatial statistics, such models in which correlation only depends on geographic distance are called *isotropic*. In order to take the graph topology into account, we performed *multidimensional scaling* (MDS) [18]. In this approach, a set of points $\mathcal{S} = \{x_1, \ldots, x_n\}$ with pairwise dissimilarity $d(x_i, x_j)$ are

embedded in a low-dimensional Euclidean space such that $||\rho(x_i) - \rho(x_j)|| \approx d(x_i, x_j)$. The mapping $\rho : \mathcal{S} \to \mathbb{R}^d$ is chosen as to minimize the squared loss ("stress"),

$$\rho^* = \operatorname*{argmin}_{\rho} \sum_{i,j} |d(x_i, x_j) - ||\rho(x_i) - \rho(x_j)|||^2$$

We then defined our isotropic process with respect to the Euclidean embedding of the road segments, by using a Gaussian kernel,

$$\Sigma^{(iso)}_{s_i s_j} = \theta_1 \exp\left(-\frac{(\rho(s_i) - \rho(s_j))^2}{\theta_2^2}\right),$$

with respect to the Euclidean embedding $\rho(s)$ of the road segments $s \in \mathcal{V}$[6]. We trained the hyperparameters θ_1 and θ_2 by maximizing the marginal likelihood [25].

We first compared the uninformed, isotropic kernel $\Sigma^{(iso)}$ with the rich correlation structure $\Sigma^{(full)}$ of the historical training data. In order to compare the two kernels, we repeatedly randomly sampled sets of 10 road segments to observe, and used the kernel to predict the road speeds at the unobserved locations. In both cases, we used the mean velocities estimated from the training set to define the Gaussian process. Figure 3(a) displays a comparison of demand-weighted prediction accuracy for both kernels. Although the fully informed kernel $\Sigma^{(full)}$ leads to an error of 0.27, the uninformed kernel $\Sigma^{(iso)}$ shows much higher error of 0.43.

In practice, it would be unrealistic to assume knowledge about the full correlation structure of all road segments. However, some knowledge is typically available about how specific parts of the road network are correlated. For example, a small number of fixed sensors (such as traffic cameras, etc.) might be permanently deployed, and their data can be used for training the phenomenon model. In order to understand the benefit of partial knowledge about road speed correlations, we performed the following experiment. We repeatedly randomly picked a set $\mathcal{A}$ of road segments of fixed size k. Our goal was then to learn a kernel $\Sigma^{(k)}$ which observes the correlation structure of $\mathcal{A}$, *i.e.*, $\Sigma^{(k)}_{\mathcal{A}\mathcal{A}}$ should correspond to the empirical covariance $\Sigma^{(full)}_{\mathcal{A}\mathcal{A}}$ of the selected sensors $\mathcal{A}$. Locally however, the kernel should behave similarly to the isotropic kernel $\Sigma^{(iso)}$, i.e., the pairwise correlation should decay according to the geodesic distances on the road network graph. Hence, by varying k from 0 to 534, we would effectively interpolate between the uninformed isotropic kernel and the fully informed empirical covariance. In order to achieve this interpolation, we use an approach described by Nott and Dunsmuir [23]. There, it is shown that the (nonstationary) kernel defined by

$$\Sigma^{(k)}_{st} = \Sigma^{(iso)}_{s\mathcal{A}} \Sigma^{(iso)\,-1}_{\mathcal{A}\mathcal{A}} \left(\Sigma^{(full)}_{\mathcal{A}\mathcal{A}} \Sigma^{(iso)\,-1}_{\mathcal{A}\mathcal{A}} - I\right) \Sigma^{(iso)}_{\mathcal{A}t} + \Sigma^{(iso)}_{st}$$

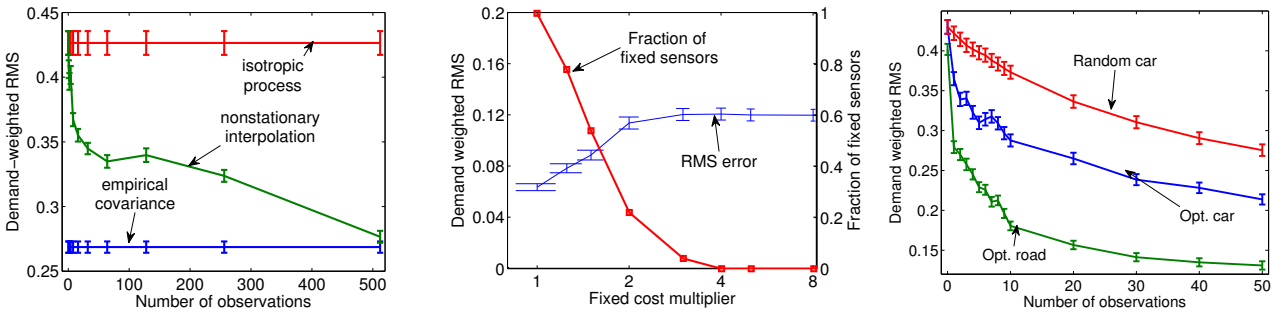
satisfies the desired property. Figure 3(a) shows the decrease in prediction error as more data is used for training (*i.e.,* k is increased from 0 to 534), when using only 10 randomly chosen observations for prediction. We can see that there is a rapid initial drop—even if only $k = 16$ training sensors are available, the error is reduced from 0.43 to 0.36. Hence, a small number of initial sensors greatly helps to "coordinate" the correlation.

Stationarity vs. mobility: As we have seen in the previous experiments, the ability to select individual road segments (the case of no selection noise) results in lower error than when selection noise is present. The case of no selection noise corresponds to a static deployment of sensors at a specified set of locations. The situation of only placing static sensors, or only querying moving objects are two ends of a spectrum. We performed an additional experiment to explore this spectrum in more detail. In this experiment, we fix the cost of the community provided "mobile sensors" to 1, and additionally allow applications to maintain dedicated fixed sensors at a cost $c \geq 1$. For a fixed budget $L = 50$, we then try to find the optimal balance between static and mobile sensors, maximizing the error reduction, while spending at most our budget L. We use the CELF algorithm described in [19] to find this balance. This algorithm also exploits submodularity, and is guaranteed to find a solution achieving at least a constant fraction of $\frac{1}{2}(1 - 1/e)$ of the optimal score.

Figure 3(b) presents the results of this experiment. For each cost multiplier c, we plot the fraction of static sensors contained in the approximate solution, as well as the demand-weighted error incurred by this solution. For $c = 1$, only static sensors are selected. However, even for $c = 2$, only a small fraction of static sensors are selected, which indicates that the mobile sensors very quickly become more cost-effective. For $c \geq 4$, no static sensors are selected.

Spatiotemporal querying: In our last experiment, we consider the spatiotemporal observation planning problem. Here, our goal is not just to decide *which* car to query, but also *when*. More specifically, we consider an autoregressive model for three timeslices, 15 minutes apart. In this setting, our phenomenon model captures the correlation of road speeds across both space and time. Formally, we associate a random variable $\mathcal{X}_{s,t}$ with each road segment s and each timestep t. We train our phenomenon model by grouping subsequent triples of training examples $(\mathbf{x}^{(t)}_{\mathcal{V}}, \mathbf{x}^{(t+1)}_{\mathcal{V}}, \mathbf{x}^{(t+2)}_{\mathcal{V}})$, and consider the empirical covariance of this data as the kernel of our GP. Because of the sparseness of data, we consider demand $\mathcal{D}_{s,t} = \mathcal{D}_s$ and availability distributions as constant over time, and identical to the single time step variant. Similarly to the road segments, we replicate the possible observations $\mathcal{W}$ for each timestep: a single observation now is a pair (w, t) of cell/user w and timestep t. When selecting an observation

[6]We also experimented with a class of graph kernels called *diffusion kernels* [10] based on the pairwise or adjacency matrices. This approach however resulted in slightly worse generalization error.



(a) *Learning curve* (b) *Stationarity vs. mobility* (c) *Spatiotemporal querying*

Figure 3. (a) Accurate phenomenon models can be learned just using topology information of the road segment and knowledge about the correlation of only a few road segments. (b) cost-benefit analysis between purely static sensors (no selection noise) and mobile sensors (selection noise). If more than twice as expensive, static sensors become less cost-effective. (c) random and optimized selection of cars and road segments in the spatiotemporal setting (0, 15 and 30 minutes timeslices).

(w,t), we effectively select a distribution $P((s,t) \mid (w,t))$ over road segments at the same time step t.

We modify our objective function $F(\mathcal{B})$ in order to ignore observations made in the future. To ensure this, we modify our objective $R(\mathcal{A})$ in the following way:

$$R(\mathcal{A}) = \sum_{(s,t)\in\mathcal{V}} \mathbb{E}\left[\mathcal{D}_s(\mathrm{Var}(\mathcal{X}_{s,t}) - \mathrm{Var}(\mathcal{X}_{s,t} \mid \mathcal{X}_{\mathcal{A}_{\leq t}}))\right]$$

where $\mathcal{A}_{\leq t} \subseteq \mathcal{A}$ refers to the subset of $\mathcal{A}$ corresponding to road segments observed before or at time step t. Note that the submodularity of $F(\mathcal{B})$ holds just as in the single time step variant.

Based on this modified objective, we now use the greedy algorithm to select a set of observations $\mathcal{B} = \{(w_1,t_1),\ldots,(w_k,t_k)\}$ across sensors and time. We first consider the restriction that we may query the same sensor only once. Figure 3(c) presents the results of this experiment. We can again observe that the optimized selection of sensors drastically improves over random sampling, reducing the error from .43 to .29 in only 10 observations, whereas random sampling requires 40 observations to achieve the same reduction in error. However, the discrepancy between selecting cars and selecting road segments gets larger in the spatiotemporal setting.

6. Discussion

Other applications: A great variety of community sensing applications are feasible, beyond traffic monitoring. Promising services include applications in fitness and recreation, where community sensors such as cellphone cameras can monitor the influence of construction and weather on roads, paths, and trails. Physiological sensors such as heart rate monitors can help model the current and forthcoming effort required on biking trails. In other applications, consumers may share imagery or audio of updated window displays, restaurant quality, or service provider interactions. Business headquarters may use community sensors to infer customer interests, urban moods, franchisee performance, and teen cultures. Real-time coverage of newsworthy events and citizen concerns can also be achieved through such systems.

For many applications the benefits of receiving shared data (*e.g.*, drivers 500 yards back on the freeway receive an alert, "Get ready to stop—there's been an accident ahead!") can provide sufficient incentive for sharing. For other applications, people may be provided with a variety of economically sound reimbursements for contributing data. Implementing means to comfortably share data lowers the cost and can thus change the cost-benefit equation.

Related work: Some applications using privately owned sensors are already being deployed in systems such as SensorBase.org, Weather Underground[7], SlamXR[8], and MotionBased.org. However, these systems purely rely on contributed data, with no facility of actively soliciting data, as considered in this paper. The SenseWeb system [9] has been developed as a platform for sharing sensor data, and could potentially be used as an infrastructure for integrating observations obtained using techniques described in this paper. In [7] spatial dynamic voting (SDV) is proposed to infer data demand, similar to our demand model. However, they do not present techniques for integrated modeling of phenomenon, demand, sensor availability and preferences as done in this paper.

The problem of selecting observations for monitoring spatial phenomena has been investigated extensively in geostatistics (*c.f.*, [3] for an overview), and more generally (Bayesian) experimental design (*c.f.*, [1]). Heuristics for actively selecting observations in GPs in order to achieve uniform variance reduction have been proposed by [26]. Sensor selection considering both the value of information together

[7] http://www.wunderground.com/
[8] http://www.msslam.com/slamxr/slamxr.htm

with the cost of acquiring the information in the context of sensor networks was first formalized in [32]. Submodularity has been used to analyze algorithms for placing a fixed set of sensors [12, 13] and coordinate mobile robots [27]. These approaches neither consider the case of highly uncertain sensor availability nor address privacy concerns, such as notions of selection noise, and user preferences about data access as done in this paper.

7. Conclusions

We presented an approach to collect in an active manner observations from privately held sensors by making data sharing more acceptable. We developed formal models that integrate sharing preferences with probabilistic models for determining the value of probing different sensors to predict a phenomenon, application demand for sensing accuracy, and sensor availability. We described an algorithm to select a near-optimal subset of observations, using the demand-weighted error reduction as a measure of context-specific value of information. We demonstrated the feasibility of our approach on a realistic traffic monitoring application. For this scenario, we estimated models for phenomena (road speeds), demand (route-planning requests), and sensor availability (likely presence of cars on road segments) from real traffic data and driving traces. Our results indicate that optimized selection of observations can significantly reduce the number of queries needed to achieve a specified level of accuracy. We believe that the community sensing methodology and potential extensions hold promise for unlocking streams of data from privately held sensors for a wide spectrum of spatiotemporal phenomena. We hope that the approach we have described will stimulate other efforts to employ distributed sensors in a utilitarian manner while observing preferences about privacy.

References

[1] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Stat. Science*, 10(3):273–304, Aug. 1995.

[2] P. Cheng, Z. Qiu, and B. Ran. Particle filter based traffic state estimation using cell phone network data. In *IEEE Intelligent Transportation Systems Conference*, 2006.

[3] N. A. C. Cressie. *Statistics for Spatial Data*. Wiley, 1991.

[4] A. Das and D. Kempe. Algorithms for subset selection in linear regression. In *STOC*, 2008.

[5] C. Dwork. Differential privacy. In *ICALP*, 2006.

[6] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey. An analysis of approximations for maximizing submodular set functions - ii. *Math. Prog. Study*, (8):73–87, 1978.

[7] K. Goldberg, D. Song, and A. Levandowski. Collaborative teleoperation using networked spatial dynamic voting. *Proceedings of the IEEE*, 91(3):430–439, 2003.

[8] E. Horvitz, J. Apacible, R. Sarin, and L. Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *UAI*, 2005.

[9] A. Kansal, S. Nath, J. Liu, and F. Zhao. Senseweb: An infrastructure for shared sensing. *IEEE Multimedia*, 14(4).

[10] R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *ICML*, 2002.

[11] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *UAI*, 2005.

[12] A. Krause and C. Guestrin. Near-optimal observation selection using submodular functions. In *AAAI Nectar*, 2007.

[13] A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg. Near-optimal sensor placements: Maximizing information while minimizing communication cost. In *IPSN*, 2006.

[14] A. Krause and E. Horvitz. A utility-theoretic approach to handling privacy in online personalization. Technical Report MSR-TR-2007-135, Microsoft Research, October 2007.

[15] A. Krause, E. Horvitz, A. Kansal, and F. Zhao. Foundations of community sensing. Technical Report MSR-TR-2007-136, Microsoft Research, October 2007.

[16] J. Krumm. Inference attacks on location tracks. In *Pervasive*, 2007.

[17] J. Krumm and E. Horvitz. Predestination: Inferring destinations from partial trajectories. In *UbiComp*, 2006.

[18] J. B. Kruskal and M. Wish. Multidimensional scaling. *Sage University series Quant. Appl. in Soc. Sci. 07-011*, 1978.

[19] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.

[20] D. A. Lieb. MoDOT tracking cell phone signals to monitor traffic speed, congestion. SEMissourian.com, Sept. '07.

[21] A. Meliou, A. Krause, C. Guestrin, and J. Hellerstein. Nonmyopic informative path planning in spatio-temporal models. In *AAAI*, 2007.

[22] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[23] D. J. Nott and W. T. M. Dunsmuir. Estimation of nonstationary spatial covariance structure. *Biometr.*, 89:819–829, '02.

[24] J. Olson, J. Grudin, and E. Horvitz. A study of preferences for sharing and privacy. In *CHI*, 2005.

[25] C. E. Rasmussen and C. K. I. Williams. *Gaussian Process for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, 2006.

[26] S. Seo, M. Wallat, T. Graepel, and K. Obermayer. Gaussian process regression: Active data selection and test point rejection. In *IJCNN*, volume 3, pages 241–246, 2000.

[27] A. Singh, A. Krause, C. Guestrin, W. Kaiser, and M. Batalin. Efficient planning of informative paths for multiple robots. In *IJCAI*, 2007.

[28] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[29] S. Wunnava, K. Yen, T. Babij, R. Zavaleta, R. Romero, and C. Archilla. Travel time estimation using cell phones (ttecp) for highways and roadways. Technical report, Florida Department of Transportation, April 2007.

[30] Y. Yim. California partners for advanced transit and highways (path). Technical Report UCB-ITS-PRR-2003-25, Institute of Transportation Studies, Berkeley, 2003.

[31] J. Yoon, B. Noble, and M. Liu. Surface street traffic estimation. In *MobiSys '07*, pages 220–232, New York, NY, USA, 2007. ACM.

[32] F. Zhao, J. Shin, and J. Reich. Information-driven dynamic sensor collaboration for tracking applications. *IEEE Signal Processing*, 19(2):61–72, 2002.