Truthful Incentives for Privacy Tradeoff: Mechanisms for Data Gathering in Community Sensing

Adish Singla ADISH.SINGLA@INF.ETHZ.CH
Andreas Krause KRAUSEA@ETHZ.CH
ETH Zurich, Universitätstrasse 6, 8092 Zürich, Switzerland

Abstract

Community sensing, fusing information from populations of privately-held sensors, presents a great opportunity to create efficient and cost-effective sensing applications. Yet, reasonable privacy concerns often limit the access to such data streams. How should systems valuate and negotiate access to private information, for example in return for monetary incentives? How should they optimally choose the participants from a large population of strategic users with privacy concerns, and compensate them for information shared?

In this paper, we address these questions and present a novel mechanism, SeqTGreedy, for budgeted recruitment of participants in community sensing. By exploiting a link between privacy tradeoffs in community sensing and adaptive submodularity, we prove that SeqTGreedy is budget feasible, incentive compatible (truthful) for participants and achieves near-optimal utility for a large class of sensing applications. We demonstrate the effectiveness of our approach in a case study of air quality monitoring, using data collected from the Mechanical Turk platform. Compared to the state of the art, our approach achieves up to 10% increase in acquired utility (for a given budget) and up to 30% reduction in cost (to achieve a desired level of utility).

1. Introduction

Community sensing is a new paradigm for creating efficient and cost-effective sensing applications by harnessing the data of large populations of sensors. For example, the accelerometer data from smartphone users could be used for earthquake detection and fine grained analysis of earthquake phenomena. Velocity data from GPS devices (in smartphones or automobiles) could be used to provide real-time traffic maps. However, accessing this stream of private sensor data raises reasonable concerns about privacy of the individual users. For example, mobility patterns and the house or office locations of a user could possibly be inferred from their GPS tracks (Krumm, 2007). Beyond concerns about sharing sensitive information, there are general anxieties among users about sharing data from their private devices. These concerns limit the practical applicability of deploying such applications.

Applications of community sensing are numerous. Several case studies have demonstrated the principal feasibility and usefulness of community sensing. A number of research and commercial prototypes are build, often relying on special studies to recruit volunteers (Zheng et al., 2010) or by special contracts with service providers to get data on an anonymous basis (Wunnava et al., 2007). The SenseWeb system (Kansal et al., 2007) has been developed as an infrastructure for sharing sensing data to enable various applications. Methods have been developed to characterize traffic (Yoon et al., 2007), perform forecasts about future traffic situations (Horvitz et al., 2005) or predict a driver's trajectory (Krumm & Horvitz, 2006). Cell tower signals obtained from the service providers are leveraged for travel time estimation on roadways (Wunnava et al., 2007). Additionally, captured images and video clips from smartphones have been used to link places with various categories (Chon et al., 2012). Clayton et al. (2012) describes the design of a *Community Seismic Network (CSN)* to study seismic phenomena from dense network of low cost sensors hosted by volunteers from the community. Aberer et al. (2010) envisions a community driven sensing infrastruture with air quality monitoring as application.

ICML Workshop: Machine Learning Meets Crowdsourcing, Atlanta, Georgia, USA, 2013.

Privacy concerns in community sensing are expected and reasonable (Lieb, 2007; Wunnava et al., 2007; Olson et al., 2005). Irrespective of the models of privacy we consider (Sweeney, 2002; Dwork, 2006; Machanavajjhala et al., 2006), the key concern is about identifiability as users become members of increasingly smaller groups of people sharing the same characteristics inferred from data. Beyond general anxieties about the sharing of location and mobility data, studies have demonstrated that, even with significant attempts at obfuscation, home and work locations can be inferred from GPS tracks (Krumm, 2007).

Incentives to participants for privacy tradeoff. Olson et al. (2005) show that people's willingness to share information depends greatly on the type of information being shared, with whom the information is shared, and how it is going to be used. They are willing to trade off privacy if compensated in terms of their utility gain (Krause & Horvitz, 2008). In other words, they act as strategic agents who are willing to negotiate access to certain private information in return of, *e.g.*, monetary or other form of incentives. Empowering users to opt into such negotiations is the key idea that we explore in this paper.

1.1. Overview of our approach

Our goal is to design policies for selecting (and compensating) the participants which provide near-optimal utility for the sensing application under strict budget constraints. As basis for selection, the community sensing system receives obfuscated estimates of the private attributes – for concreteness we focus on sensor location. The users also declare a bid or cost as the desired monetary incentive for participation and hence privacy tradeoff. After receiving the bids, the mechanism sequentially selects a participant, commits to make her the payment, receives the actual private information, selects the next participant and so on. At the end, all selected participants are provided the agreed payment. We model the participants as strategic agents who aim to maximize their profit, by possibly misreporting their private costs. As a consequence, we require the mechanism to be truthful. In order to capture a large class of sensing applications, we only require the utility function to be submodular (Nemhauser et al., 1978; Krause & Guestrin, 2007).

To design our mechanism, we first reduce the sequential negotiation of the privacy tradeoff to problem of adaptive submodular maximization (Asadpour et al., 2008; Golovin & Krause, 2011; Guillory & Bilmes, 2010). Then, we extend recent results of truthful budget feasible mechanisms for submodular functions (Singer, 2010; Chen et al., 2011; Singer, 2012) to the adaptive setting.

Our main contributions are:

- An integrated approach to community sensing by incentivizing users for sharing private information.
- A novel mechanism, SeqTGreedy, for budgeted recruitment of strategic participants, which achieves near-optimal utility for the community sensing application. This mechanism is of independent theoretical interest and also for other applications, *e.g.*, viral marketing.
- Evaluation of our approach on a realistic case study of air quality monitoring based on data obtained through Mechanical Turk (MTurk, 2005).

1.2. Related Work

Himmel et al. (2005) propose to provide users with rewards such as free minutes to motivate them to accept mobile advertisements. Hui et al. (2011) develop MobiAd, a system for targeted mobile advertisements, by utilizing the rich set of information available on the phone and suggesting the service providers to give discounts to the users, in order to incentivitze use of the system. Liu et al. (2008) propose a game theoretic model of privacy for social networking based mobile applications and presents a tit-for-tat mechanism by which users take decisions about their exposed location obfuscation for increasing personal or social utility. Chorppath & Alpcan (2012) study a privacy game in mobile commerce, where users choose the degree of granularity at which to report their location and the service providers offer them monetary incentives under budget constraints. The best users' response and the optimal strategy for the company are derived by analyzing the *Nash equilibrium (NE)* of the underlying privacy game. This is very different from our setting as we focus on algorithmic aspects in choosing the best set of users for participation in community sensing.

2. Problem Statement

We now formalize the problem addressed in this paper.

Sensing phenomena. We focus on community sensing applications with the goal to monitor some spatial phenomenon, such as air quality or traffic. We discretize the environment as a finite set of locations $\mathcal{V}$, where each $v \in \mathcal{V}$ could, *e.g.*, denote a zip code or more fine grained street addresses, depending on the application. We quantify the utility $f(\mathcal{A})$ of obtaining measurements from a set of locations $\mathcal{A}$ using a set function $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$. Formally, we only require that f is *nonnegative*, *monotone* (i.e., whenever $\mathcal{A} \subseteq \mathcal{A}' \subseteq \mathcal{V}$ it holds that $f(\mathcal{A}) \leq f(\mathcal{A}')$) and *submodular*. Submodularity is an intuitive notion of diminishing returns, stating that, for any sets $\mathcal{A} \subseteq \mathcal{A}' \subseteq \mathcal{V}$, and any fixed location $a \notin \mathcal{A}'$ it holds that $f(\mathcal{A} \cup \{a\}) - f(\mathcal{A}) \geq f(\mathcal{A}' \cup \{a\}) - f(\mathcal{A}')$.

As a simple, concrete example, we may derive some non-negative value d_a for observing each location $a \in \mathcal{A}$, and may define $f(\mathcal{A}) = \sum_{a \in \mathcal{A}} d_a$. These conditions are rather general, satisfied by many sensing utility functions and f can capture much more complex notions, such as reduction of predictive uncertainty in a probabilistic model (Krause & Guestrin, 2007).

Sensing profile of users. We consider a community of N users, denoted by set $\mathcal{W}$, owning some sensing device such as a smartphone. Each user can make observations at a set of locations depending on her geolocation or mobility as well as the type of device used. We model this through a collection of *sensing profiles* $\mathcal{O} \subseteq 2^{\mathcal{V}}$ whereby we associate each user $w \in \mathcal{W}$ with a profile $y_w \in \mathcal{O}$, specifying the set of locations covered by her. We denote a given set of users $\mathcal{S} \subseteq \mathcal{W}$ jointly with their sensing profiles as $\mathbf{y}_{\mathcal{S}} \subseteq \mathcal{W} \times \mathcal{O}$. The goal is to select set of users $\mathcal{S}$ (also called *participants*) so as to maximize the utility of the sensing application given by $g(\mathbf{y}_{\mathcal{S}}) = f(\mathcal{A})$ where $\mathcal{A} = \bigcup_{s \in \mathcal{S}} y_s$. We assume that each users' maximal contribution to the utility is bounded by f_{max}.

Privacy profile of users. In order to protect privacy, we consider the setting where the exact sensing profiles y_w of the users (containing, *e.g.*, tracks of locations visited) are not known to the sensing system. Instead, y_w is only shared after obfuscation with a random perturbation intended to reduce the risk of identifiability (Sweeney, 2002; Dwork, 2006). The system's highly uncertain belief about the sensing profile of user w can therefore be represented as a random variable (also called *privacy profile*) Y_w with y_w being its realization. We use $\mathbf{Y}_{\mathcal{W}} = [Y_1, \ldots, Y_N]$ to refer to the collection of all (independent) variables associated with population $\mathcal{W}$ and assume that $\mathbf{Y}_{\mathcal{W}}$ is distributed according to a factorial joint distribution $P(\mathbf{Y}_{\mathcal{W}})$. The sensing profile y_w is revealed to the application only after it commits to provide the desired incentives to the user w. Adding this privacy component, the goal of the application is to select a (*random*) set of users $\mathcal{S}$ to maximize $\mathbb{E}_{\mathbf{Y}_{\mathcal{W}}}[g(\mathbf{y}_{\mathcal{S}})]$, *i.e.*, the expected utility, where the expectation is over realizations of $\mathbf{Y}_{\mathcal{W}}$ *w.r.t.* $P(\mathbf{Y}_{\mathcal{W}})$.

Incentives structure for privacy tradeoff. We assume that users are willing to share certain non-sensitive private information in return of monetary incentives. Each user w has a private cost $c_w \in \mathbb{R}_{\geq 0}$ that she experiences for her privacy tradeoff. Instead of revealing c_w, she only reveals a *bid* $b_w \in \mathbb{R}_{\geq 0}$. We are interested in *truthful* mechanisms, where it is a dominant strategy for a user to report $b_w = c_w$, *i.e.*, users cannot increase their profit (in *expectation*) by lying about their true cost. We assume that costs have known bounded support, *i.e.*, $c_w \in [c_{min}, c_{max}]$.

Optimization problem. Given a strict budget constraint $\mathcal{B}$, the goal of the sensing application is to design a mechanism $\mathcal{M}$, which implements an allocation policy to select participants $\mathcal{S}$ and a payment scheme to make *truthful* payments θ_s to each of the participants, with the goal of maximizing the expected utility. Instead of committing to a fixed set of participants $\mathcal{S}$ in advance (*non-adaptive* policy), we are interested in mechanisms that implement an *adaptive* policy taking into account the observations made so far (revealed sensing profiles of participants already selected) when choosing the next user. Formally, the goal of the mechanism is to adaptively select participants $\mathcal{S}^*$ along with the payments $\theta_{\mathcal{S}^*}$, such that

$$\mathcal{S}^* = \underset{\mathcal{S} \subseteq \mathcal{W}}{\arg\max}\, \mathbb{E}_{\mathbf{Y}_{\mathcal{W}}}[g(\mathbf{y}_{\mathcal{S}})] \text{ subject to } \sum_{s \in \mathcal{S}} \theta_s \leq \mathcal{B} \quad (1)$$

Here, the set of participants $\mathcal{S}$ selected and the payments $\theta_{\mathcal{S}}$ may depend on the realization of $\mathbf{Y}_{\mathcal{W}}$ as well. We formally introduce adaptive policies in Section 4.

3. Existing Mechanisms

We first review existing mechanisms that fall short of either privacy-preservation, adaptivity or truthfulness. In Section 4, we then build on these and present our main contribution: a privacy-respecting, truthful and adaptive mechanism.

3.1. Without privacy

Consider first an unrealistic setting, where the system has full information about the users' exact sensing profiles and their true costs. In such a setting, Problem 1 reduces to that of budgeted maximization of a monotone non-negative submodular function with non-uniform costs, studied by Sviridenko (2004). A simple algorithm combining partial enumeration with greedy selection guarantees a utility of at least $(1 - 1/e)$ times that obtained by optimal selection OPT. This result is tight under reasonable complexity assumptions (Feige, 1998). We denote this setting and mechanism as GREEDY. Note that each participant is paid its true cost in this untruthful setting. Now, consider the non-private setting with unknown true costs. This problem reduces to designing truthful budget feasible mechanism for monotone submodular set functions, as studied in (Singer, 2010; Chen et al., 2011; Singer, 2012), where a constant factor 7.91 approximation compared to OPT can be achieved. We use TGREEDY to denote this setting and mechanism. TGREEDY executes a greedy allocation on a reduced budget with carefully chosen stopping criteria (for ensuring budget feasibility), in order to select a set of participants and then computes the truthful payments to be made to them.

Table 1: Different information settings and mechanisms.

	Untruthful	**Truthful**
Privacy Off	GREEDY	TGREEDY
Privacy On (Non-Adaptive)	CONSTGREEDY	CONSTTGREEDY
Privacy On (Adaptive)	SEQGREEDY	**SeqTGreedy**

3.2. Non-adaptive with privacy

In our case, where privacy is preserved through random obfuscation, one must deal with the stochasticity caused by the uncertainty about users' sensing profiles. Here, the objective

$$G(\mathcal{S}) \equiv \mathbb{E}_{\mathbf{Y}_\mathcal{W}}[g(\mathbf{y}_\mathcal{S})] = \sum_{\mathbf{y}_\mathcal{W}} P(\mathbf{Y}_\mathcal{W} = \mathbf{y}_\mathcal{W}) f\left(\bigcup_{s \in \mathcal{S}} y_s\right)$$

in (1) can be seen as an expectation over multiple submodular set functions, one for each realisation of $\mathbf{Y}_\mathcal{W}$. However, as submodularity is preserved under expectations, the set function $G(\mathcal{S})$ is submodular as well. One can therefore still apply the mechanisms GREEDY and TGREEDY in order to obtain near-optimal *non-adaptive* solutions (*i.e.*, the set of participants is fixed in advance) to Problem (1). We denote these non-adaptive (constant) mechanisms applied to our privacy preserving setting as CONSTGREEDY and CONSTTGREEDY. .

3.3. Untruthful, adaptive with privacy

Instead of non-adaptively committing to the set $\mathcal{S}$ of users selected a priori, one can hope to obtain increased utility through adaptive (sequential) selection, *i.e.*, by taking into account the observations from the users selected so far when choosing the next user. Without assumptions, computing such an optimal policy for Problem (1) is intractable. Fortunately, as long as the sensing quality function f is monotone and submodular (see Section 2), Problem (1) satisfies a natural condition called *adaptive submodularity* (Golovin & Krause, 2011). This condition generalizes the classical notion of submodularity to sequential decision problems as faced here.

Adaptive submodularity requires, in our setting, that the expected benefit of any fixed user $w \in \mathcal{W}$ given a set of observations (*i.e.*, set of users and observed sensing profiles) can never increase as we make more observations. Formally, consider the *conditional expected marginal gain* of adding a user $w \in \mathcal{W} \setminus \mathcal{S}$ to an existing set of observations $\mathbf{y}_\mathcal{S} \subseteq \mathcal{W} \times \mathcal{O}$:

$$\begin{aligned}\Delta_g(w|\mathbf{y}_\mathcal{S}) =& \mathbb{E}_{Y_w}[g(\mathbf{y}_\mathcal{S} \cup \{(w, y_w)\}) - g(\mathbf{y}_\mathcal{S})|\mathbf{y}_\mathcal{S}] \\ =& \sum_{y \in \mathcal{O}} P(Y_w = y|\mathbf{y}_\mathcal{S}) \cdot [g(\mathbf{y}_\mathcal{S} \cup \{(w, y)\}) - g(\mathbf{y}_\mathcal{S})]\end{aligned}$$

Function g with distribution $P(\mathbf{Y}_\mathcal{W})$ is adaptive submodular, if $\Delta_g(w|\mathbf{y}_\mathcal{S}) \geq \Delta_g(w|\mathbf{y}_{\mathcal{S}'})$ for $\mathbf{y}_\mathcal{S} \subseteq \mathbf{y}_{\mathcal{S}'}$. Thus, the gain of a user w, in expectation over its unknown sensing profile, can never increase as we select and observe more participants.

Proposition 1. *Suppose f is monotone and submodular. Then the objective g and distribution P used in Problem 1 are adaptive submodular.*

Above Proposition follows from Theorem 6.1 of Golovin & Krause (2011), assuming distribution P is factorial (i.e., the random obfuscation is independent between users). Given this problem structure, for the simpler, untruthful setting (i.e., known true costs), we can thus use the sequential greedy policy for the stochastic submodular maximization studied by Golovin & Krause (2011). This approach is denoted by SEQGREEDY and obtains a utility of at least $(1 - {}^1\!/e)$ times that of optimal sequential policy SEQOPT.

Table 1 summarizes the settings and mechanisms considered so far. They all fall short of at least one of the desired characteristics of privacy-preservation, truthfulness or adaptivity. In next section, we present our main contribution – SEQTGREEDY, an adaptive mechanism for the realistic setting of privacy-sensitive and strategic agents.

4. Our main mechanism: SeqTGreedy

We now describe our mechanism $\mathcal{M} = (\boldsymbol{\pi}_\mathcal{M}, \boldsymbol{\theta}_\mathcal{M})$, with allocation policy $\boldsymbol{\pi}_\mathcal{M}$ and payment scheme $\boldsymbol{\theta}_\mathcal{M}$. $\mathcal{M}$ first obtains the bids $B_\mathcal{W}$ and privacy profiles $P(\mathbf{Y}_\mathcal{W})$ from all users, runs the allocation policy $\boldsymbol{\pi}_\mathcal{M}$ to adaptively select participants $\mathcal{S}$ and makes observations $\mathbf{y}_\mathcal{S}$ during selection. At the end, it computes payments $\theta_\mathcal{S}$ using scheme $\boldsymbol{\theta}_\mathcal{M}$. The allocation policy $\boldsymbol{\pi}_\mathcal{M}$ can be thought of as a decision tree. Formally, a policy $\pi : 2^{\mathcal{W} \times \mathcal{O}} \to \mathcal{W}$ is a partial mapping from observations $\mathbf{y}_\mathcal{S}$ made so far to the next user $w \in \mathcal{W} \setminus \mathcal{S}$ to be picked, denoted by $\pi(\mathbf{y}_\mathcal{S}) = w$. We seek policies that are provably competitive with the optimal (intractable) sequential policy SEQOPT. $\boldsymbol{\theta}_\mathcal{M}$ computes payments which are truthful in expectation (a user cannot increase her total expected profit by lying about her true cost, for a fixed set of bids of other users) and individually rational ($\theta_s \geq b_s$). For budget feasibility, the allocation policy needs to ensure that the budget $\mathcal{B}$ is sufficient to make the payments $\theta_\mathcal{S}$ to all selected participants. Next, we describe in detail the allocation policy and payment scheme of SEQTGREEDY with these desirable properties.

Policy 1 Allocation policy of SeqTGreedy

Parameters: $\mathcal{B}$; $\mathcal{W}$; $\mathbf{Y}_{\mathcal{W}}$; $B_{\mathcal{W}}$;
Outputs: *participants*: $\mathcal{S}=\emptyset$; *observations*: $\mathbf{y}_{\mathcal{S}}=\emptyset$; *marginals*: $\mathbf{\Delta}_{\mathcal{S}}=\emptyset$;
Variables: *remaining users* $\mathcal{W}'=\mathcal{W}$; *set* $\alpha_{\geq 1}=2$;
while $\mathcal{W}'\neq\emptyset$ **do**
  $w^*=\arg\max_{w\in\mathcal{W}'}\frac{\Delta_g(w|\mathbf{y}_{\mathcal{S}})}{b_w}$;
  $\Delta_{w^*}=\Delta_g(w^*|\mathbf{y}_{\mathcal{S}})$;
  if $B_{\mathcal{S}}+b^*_w\leq\mathcal{B}$ **then**
    if $b^*_w\leq\frac{\mathcal{B}}{\alpha}\cdot\frac{\Delta_{w^*}}{\left((\sum_{s\in\mathcal{S}}\Delta_s)+\Delta_{w^*}\right)}$ **then**
      $\mathcal{S}=\mathcal{S}\cup w^*$; $\mathbf{\Delta}_{\mathcal{S}}=\mathbf{\Delta}_{\mathcal{S}}\cup\Delta_{w^*}$
      Observe y_{w^*} ; $\mathbf{y}_{\mathcal{S}}=\mathbf{y}_{\mathcal{S}}\cup(w^*,y_{w^*})$
      $\mathcal{W}'=\mathcal{W}'\setminus w^*$
    else
      $\mathcal{W}'=\emptyset$
    end if
  else
    $\mathcal{W}'=\mathcal{W}'\setminus w^*$
  end if
end while
return $\mathcal{S}$; $\mathbf{y}_{\mathcal{S}}$; $\mathbf{\Delta}_{\mathcal{S}}$;

4.1. Allocation policy of SeqTGreedy

Policy 1 presents the allocation policy of SeqTGreedy. The main ingredient of the policy is to greedily pick the next user that maximizes the expected marginal gain $\Delta_g(w|\mathbf{y}_{\mathcal{S}})$ per unit cost. The policy uses additional stopping criteria to enforce budget feasibility, similar to TGreedy (Chen et al., 2011). Firstly, it runs on a reduced budget $\mathcal{B}/\alpha$. Secondly, it uses a proportional share rule ensuring that the expected marginal gain per unit cost for the next potential participant is at least equal to or greater than the expected utility of the new set of participants divided by the budget. We shall prove in the Section 4.3 that $\alpha=2$ achieves the desired properties.

4.2. Payment characterization of SeqTGreedy

The payment scheme is based on the characterization of threshold payments used by TGreedy (Singer, 2010). However, a major difficulty arises from the fact that the computation of payments for a participant depends also on the unallocated users, whose sensing profiles are not known to the mechanism. Let $\mathcal{S}$ denote the set of participants allocated by $\boldsymbol{\pi}_{\mathcal{M}}$ along with making observations $\mathbf{y}_{\mathcal{S}}$. Let us consider the set of all possible realizations of $\mathbf{Y}_{\mathcal{W}}=\mathbf{y}_{\mathcal{W}}\subseteq\mathcal{W}\times\mathcal{O}$ consistent with $\mathbf{y}_{\mathcal{S}}$, i.e., $\mathbf{y}_{\mathcal{S}}\subseteq\mathbf{y}_{\mathcal{W}}$. We denote this set by $\mathbf{Z}_{\mathcal{W},\mathcal{S}}=[\mathbf{y}^1,\mathbf{y}^2\ldots\mathbf{y}^r\ldots\mathbf{y}^Z]$, where $Z=|\mathbf{Z}_{\mathcal{W},\mathcal{S}}|$. We first discuss how to compute the payment for each one of these possible realizations $\mathbf{y}^r\in\mathbf{Z}_{\mathcal{W},\mathcal{S}}$, denoted by $\theta^d_s(\mathbf{y}^r)$ (where d indicates here an association with the deterministic setting of knowing the exact sensing profiles of all users $w\in\mathcal{W}$). These payments for specific realizations are then combined together to compute the final payment to each participant.

Payment θ^d_s for a given $\mathbf{y}_{\mathcal{W}}$. Consider the case where the variables $\mathbf{Y}_{\mathcal{W}}$ are in state $\mathbf{y}_{\mathcal{W}}\in\mathbf{Z}_{\mathcal{W},\mathcal{S}}$ and let $\mathcal{S}$ be the set of participants allocated by the policy. We use the well-known characterization of Myerson (1981) of truthful payments in single-parameter domains. It states that a mechanism is truthful when 1) allocation rule is monotone (*i.e.*, an already allocated user cannot be unallocated by lowering her bid, for a fixed set of bids of others) and 2) allocated users are paid threshold payments (*i.e.*, the highest bid they can declare before being removed from the allocated set). Monotonicity follows naturally from the greedy allocation policy which sorts users based on expected marginal gain per unit cost. To compute threshold payments, we need to consider a maximum of all the possible bids that a user can declare and still get allocated, as explained next.

Let us renumber the users in the order $\mathcal{S}=\{1,\ldots,i,\ldots,k\}$ in which they were allocated by mechanism $\mathcal{M}$ and let us analyze the payment for participant $s=i$. Consider running the policy on an alternate set $\mathcal{W}'=\mathcal{W}\setminus\{i\}$ and let $\mathcal{S}'=\{1,\ldots,j,\ldots,k'\}$ be the allocated set (users renumbered again based on order of allocation when running the policy on $\mathcal{W}'$). $\mathbf{\Delta}_{\mathcal{S}}$ and $\mathbf{\Delta'}_{\mathcal{S}'}$ are the marginal contributions of the participants in the above two runs of the policy. We define $\Delta_{i(j)}$ to be the marginal contribution of i (from $\mathcal{S}$) if it has to replace the position of j (in set $\mathcal{S}'$). Now, consider the bid that i can declare to replace j in $\mathcal{S}'$ by making a marginal contribution per cost higher than j, given by $b_{i(j)}=\frac{\Delta_{i(j)}\cdot b_j}{\Delta'_j}$. Additionally, the bid that i can declare must satisfy the proportional share rule, denoted by $\rho_{i(j)}=\frac{\mathcal{B}}{\alpha}\cdot\Delta_{i(j)}/((\sum_{s'\in[j-1]}\Delta'_{s'})+\Delta_{i(j)})$. By taking the minimum of these two values, we get $\theta^d_{i(j)}=\min(b_{i(j)},\rho_{i(j)})$ as the bid that i can declare to replace j in $\mathcal{S}'$. The threshold payment for participant $s=i$ is given by $\theta^d_i=\max_{j\in[k'+1]}\theta^d_{i(j)}$.

Computing the final payment θ_s. For each $\mathbf{y}^r\in\mathbf{Z}_{\mathcal{W},\mathcal{S}}$, compute $\theta^{d,r}_i=\theta^d_i(\mathbf{y}^r)$. The final payment made to participant s is given by $\theta_s=\sum_{\mathbf{y}^r\in\mathbf{Z}_{\mathcal{W},\mathcal{S}}}P(\mathbf{Y}_{\mathcal{W}}=\mathbf{y}^r|\mathbf{y}_{\mathcal{S}})\cdot\theta^{d,r}_s$. Note that the set $\mathbf{Z}_{\mathcal{W},\mathcal{S}}$ could be exponentially large, and hence computing the exact θ_s may be intractable. However, one can use sampling to get estimates of θ_s in polynomial time (using Hoeffding's inequality to bound sample complexity) and thus implement an approximately truthful payment scheme to any desired accuracy. Further, note that the approximation guarantees of $\mathcal{M}$ do not require computation of the payments at all, and only require execution of the allocation policy, which runs in polynomial time.

4.3. Analysis of SeqTGreedy

We now analyze the mechanism and prove its desirable properties. The proofs of all theorems are presented in the extended version of this paper. We only sketch them here.

Theorem 1. SeqTGreedy *is truthful in expectation, i.e., no user can increase her profit in expectation by lying about her true cost, for a fixed set of bids of other users.*

Firstly, truthfulness of payments $\theta_s^{d,r}$ is proved for a considered realization $\mathbf{y}^r$. This is done by showing the monotonicity property of the greedy allocation policy and proving the threshold nature of the payment $\theta_s^{d,r}$. Truthfulness of the actual payment θ_s follows from the fact that it is a linear combination of individually truthful payments $\theta_s^{d,r}$.

Theorem 2. *Payments made by* SeqTGreedy *are individually rational, i.e.* $\theta_s \geq b_s$.

This is proved by showing a lower bound of b_s on each of the payments $\theta_s^{d,r}$ used to compute the final payment θ_s.

Theorem 3. *For* $\alpha = 2$, SeqTGreedy *is budget feasible, i.e.,* $\theta_{\mathcal{S}} \leq \mathcal{B}$.

We first prove an upper bound on the payments made to any participant by $\mathcal{B} \cdot {}^{\Delta_s}/_{(\sum_{s \in \mathcal{S}} \Delta_s)}$ by adapting the proof from (Chen et al., 2011). Surprisingly, these payment bounds hold irrespective of the payment scheme used by the mechanism. Summing over these payments ensures budget feasibility.

Theorem 4. *For* $\alpha = 2$, SeqTGreedy *achieves a utility within a factor of 4.75 of that obtained by the optimal policy* SeqOpt *(with full knowledge of true costs), when the utility contribution of each participant is small compared to the overall utility achieved by the mechanism.*

We show that, because of the diminishing returns property of the utility function, the stopping criteria used by the mechanism based on proportional share and using only an α proportion of the budget still allows the allocation of sufficiently many participants to achieve a competitive amount of utility. We also use the fact that in our settings, each user can only contribute a maximal of f_{max} utility to the application, which, for a large-scale application, is small compared to the utility achieved by the mechanism under a given budget.

5. Experimental Evaluation

In this section, we carry out extensive experiments to understand the practical performance of our mechanism on a realistic community sensing case study.

Benchmarks. We compare against the following benchmarks and state-of-the-art mechanisms.

- SeqGreedy (unrealistically) assumes access to the true costs of the users, thus measuring the loss incurred by SeqTGreedy for enforcing truthfulness and serving as upper bound benchmark on untruthful mechanisms.
- Random allocates users randomly until the budget is exhausted and pays each participant its true cost. This represents a lower bound benchmark on untruthful mechanisms.
- ConstTGreedy is the non-adaptive variant of SeqTGreedy and the state-of-the-art truthful mechanism.
- TGreedy (unrealistically) assumes access to the exact sensing profiles of the users and hence provides insights in measuring the loss incurred due to privacy protection.

Metrics and experiments. The primary metric we measure is the utility acquired by the application. We also measure budget required to achieve a specified utility. To this end, we conduct experiments by varying the given budget and then varying the specified utility, for a fixed obfuscation level. To further understand the impact of random obfuscation, we then vary the level of obfuscation and measure i) *% Gain from adaptivity* (SeqTGreedy vs. ConstTGreedy), ii) *% Loss from truthfulness* (SeqTGreedy vs. SeqGreedy), and iii) *% Loss from privacy* (SeqTGreedy vs. TGreedy).

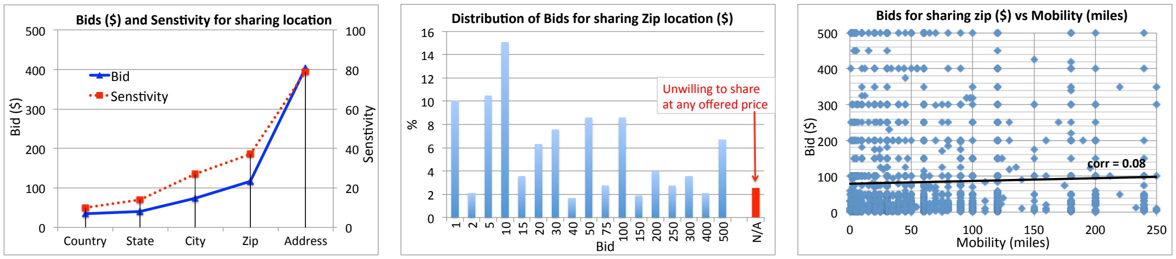
5.1. Experimental setup and data sets.

We now describe our setup and data collection from Mechanical Turk (henceforth MTurk).

Community sensing application. Suppose we wish to monitor air quality using mobile sensors (Aberer et al., 2010). We consider a granularity level of zip codes and locations $\mathcal{V}$ correspond to the zip codes of state Nevada, USA. We obtained information related to latitude, longitude, city and county of these zips from publicly available data [1]. This represents a total of 220 zip codes located in 98 cities and 17 counties. In order to encourage spatial coverage, we choose our objective f such that one unit utility is obtained for every zip code location observed by the selected participants. To simulate a realistic population of the N users, we also obtained the population statistics for these zip codes [2].

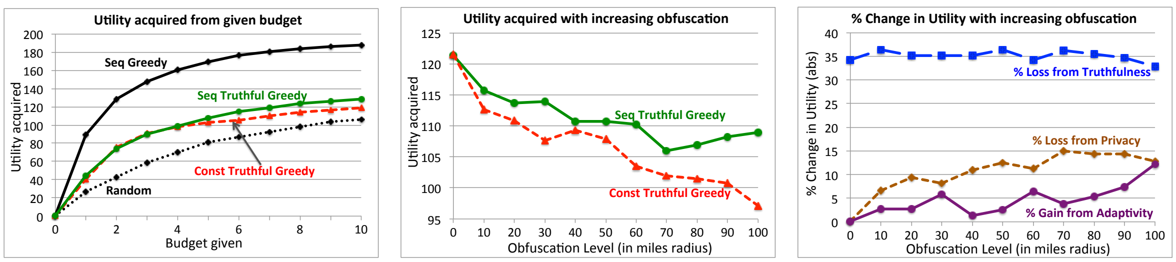
MTurk study and user attributes. We posted a Human Intelligent Task (HIT) on MTurk in form of a

[1] http://www.populardata.com/downloads.html
[2] http://mcdc2.missouri.edu/

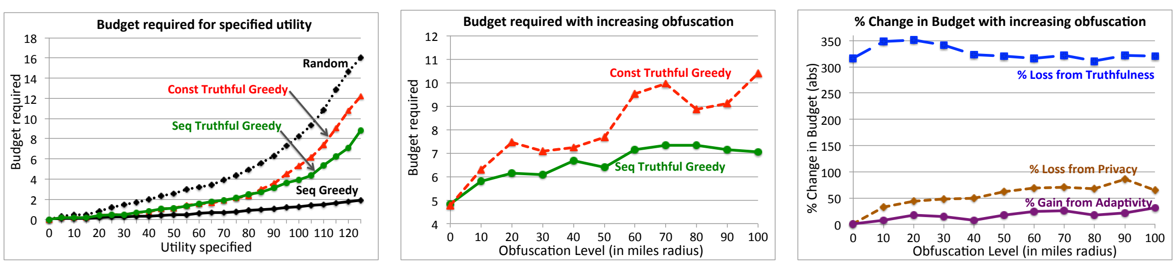


(a) Bids ($) and Senstivity (b) Distributions of Zip-Bids (c) Corr. of Zip-Bids vs Mobility

Figure 1: (**a**) Bids ($) and sensitivity ([1-100]) for different levels of privacy tradeoff; (**b**) Distribution of bids ($) for sharing zip code location; (**c**) Correlation of bids ($) for sharing zip with mobility (day distance in miles)



(a) Utility: Varying budget ($) (b) Utility: Varying obfuscation (c) % Δ Utility: Varying obfuscation



(d) Budget ($): Varying utility (e) Budget ($): Varying obfuscation (f) % Δ Budget: Varying obfuscation

Figure 2: In (**a**) and (**d**), for a fixed obfuscation level of 40 miles radius, budget given and desired utility are varied. In (**b**), (**c**), (**e**) and (**f**)) the obfuscation level is varied. (**b**) and (**c**) measure utility acquired for a given budget of 5$ and show up to *10% adaptivity gain*. (**e**) and (**f**) measure the budget required (in $) to achieve a utility of 120 and shows up to *30% adaptivity gain*.

survey, where workers were told about an option to participate in a community sensing application. Workers were asked to express their sensitivity (on scale of [1-100]), as well as the payment bids (in range of [1-500] $) they desire to receive about exposing their location at the granularity of home address, zip, city, state or country respectively. Additionally, workers were asked about their daily mobility to gather data for defining the sensing radius of the users in our experiments. A total of 650 workers participated in our HIT, restricted to workers from the USA with more than 90% approval rate and were paid a fixed amount of 0.25$ each. Fig 1(a) shows the mean bids and expressed sensitivity for different levels of obfuscation. Fig 1(b) shows the distribution of bids for exposing zip level location information. A mean daily mobility of 18 miles was reported. Fig 1(c) shows no correlation between their daily mobility (related to user's sensing radius and hence utility) and bids for exposing zip information (related to user's bid).

Parameter choices. We consider a population of size $N = 500$. We used the distributions of daily mobility to define the sensing radius of the users and distribution of expressed zip level bids as the bids in our experiments. We used $c_{min} = 0.01$ and $c_{max} = 1$ by scaling the bids in this range. We set the maximum possible utility obtained from each user to $f_{max} = 15$ by limiting the maximal number of observable zip code locations of each user to 15, which are randomly sampled from the locations covered by the user's sensing radius.

5.2. Results

Varying the given budget and specified utility. For a fixed obfuscation level of 40 miles radius, Fig 2(a) and 2(d) show the effect of varying the given budget and desired utility respectively. Fig 2(a) illustrates the bounded approximation of our mechanism SeqTGreedy w.r.t. SeqGreedy and up to 10% improvement over ConstTGreedy in terms of acquired utility. In Fig 2(d), we can see that the budget required to achieve a specified utility by our mechanism is unbounded w.r.t. SeqGreedy and we get up to 30% reduction in budget required by using the adaptive mechanism.

Utility acquired at different obfuscation levels. In Figs. 2(b) and 2(c), the acquired utility is measured for a given budget of 5$ by varying the obfuscation level. We can see that adaptivity helps acquire up to 10% higher utility and this adaptivity gain increases with higher obfuscation (more privacy). The loss from truthfulness is bounded (by 35%), agreeing with our approximation guarantees. The loss from the lack of private information grows initially. However, eventually we see that the gain from adaptivity helps to overcome the loss we incur due to privacy protection.

Budget required at different obfuscation levels. In Figs. 2(e) and 2(f), the required budget is computed for a desired utility value of 120 by varying the obfuscation level. We can see an increasing adaptivity gain, up to a total of 30% reduction in required budget. As the privacy level increases, the incurred loss from privacy (in terms of increased budget requirement) is nearly compensated by the adaptivity gain.

6. Conclusions

We presented a principled approach for negotiating access to private information in community sensing. By using insights from mechanism design and adaptive submodular optimization, we designed the first adaptive, truthful and budget feasible mechanism guaranteed to recruit a near-optimal subset of participants in community sensing. We demonstrated the feasibility and efficiency of our approach in a realistic case study. We believe that this integrated approach connecting privacy, utility and incentives provides an important step towards developing practical, yet theoretically well-founded techniques for community sensing.

References

Aberer, K., Sathe, S., Chakraborty, D., Martinoli, A., Barrenetxea, G., Faltings, B., and Thiele, L. Opensense: Open community driven sensing of environment. *IWGS*, 2010.

Asadpour, A., Nazerzadeh, H., and Saberi, A. Maximizing stochastic monotone submodular functions, 2008.

Chen, N., Gravin, N., and Lu, P. On the approximability of budget feasible mechanisms. In *SODA*, 2011.

Chon, Y., Lane, N. D., Li, F., Cha, H., and Zhao, F. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *Ubicomp*, 2012.

Chorppath, A. K. and Alpcan, T. Trading privacy with incentives in mobile commerce: A game theoretic approach. *Pervasive and Mobile Computing*, 2012.

Clayton, R., Heaton, T., Chandy, M., Krause, A., Kohler, M., Bunn, J., Olson, M., Faulkner, M., Cheng, M., Strand, L., Chandy, R., Obenshain, D., Liu, A., Aivazis, M., and Guy, R. Community seismic network. *Annals of Geophysics*, 54(6):738–747, 2012.

Dwork, C. Differential privacy. In *ICALP*, volume 4052, pp. 1–12, 2006.

Feige, U. A threshold of ln n for approximating set cover. *Journal of the ACM*, 45:314–318, 1998.

Golovin, D. and Krause, A. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research (JAIR)*, 42:427–486, 2011.

Guillory, A. and Bilmes, J. Interactive submodular set cover. In *ICML*, 2010.

Himmel, M., Rodriguez, H., Smith, N., and Spinac, C. Method and system for schedule based advertising on a mobile phone, 2005.

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

Hui, P., Henderson, T., Brown, I., and Haddadi, H. Targeted advertising on the handset : privacy and security challenges. In *Pervasive Advertising*, HCI'11. Springer, 2011.

Kansal, A., Nath, S., Liu, J., and Zhao, F. Senseweb: An infrastructure for shared sensing. *IEEE Multimedia*, 14(4), 2007.

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

Krause, A. and Horvitz, E. A utility-theoretic approach to privacy and personalization. In *AAAI*, 2008.

Krumm, J. Inference attacks on location tracks. In *PERVASIVE*, pp. 127–143, 2007.

Krumm, J. and Horvitz, E. Predestination: Inferring destinations from partial trajectories. In *Ubicomp*, pp. 243–260, 2006.

Lieb, D. A. MoDOT tracking cell phone signals to monitor traffic speed, congestion, Sept 2007.

Liu, H., Krishnamachari, B., and Annavaram, M. Game theoretic approach to location sharing with privacy in a community based mobile safety application. In *MSWiM*, pp. 229–238, 2008.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. In *ICDE*, 2006.

MTurk. Mechanical turk platform, 2005.

Myerson, R. Optimal auction design. *Mathematics of Operations Research*, 6(1), 1981.

Nemhauser, G.L., Wolsey, L.A., and Fisher, M. An analysis of the approximations for maximizing submodular set functions. *Math. Prog.*, 14:265–294, 1978.

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

Singer, Y. Budget feasible mechanisms. In *FOCS*, pp. 765–774, 2010.

Singer, Y. How to win friends and influence people, truthfully: Influence maximization mechanisms for social networks. In *WSDM*, 2012.

Sviridenko, M. A note on maximizing a submodular set function subject to knapsack constraint. *Operations Research Letters*, v.(32):41–43, 2004.

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

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

Yoon, J., Noble, B., and Liu, M. Surface street traffic estimation. In *MobiSys*, pp. 220–232, 2007.

Zheng, Y., Xie, X., and Ma, W.-Y. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Engineering Bulletin*, pp. 32–40, 2010.