User Recruitment for Mobile Crowdsensing over Opportunistic Networks

Merkouris Karaliopoulos*‡, Orestis Telelis† and Iordanis Koutsopoulos*‡
*Information Technologies Institute, Center for Research and Technology Hellas, Volos, Greece
†Department of Digital Systems, University of Piraeus, Piraeus, Greece
‡Department of Informatics, Athens University of Economics and Business, Athens, Greece

***Abstract*—We look into the realization of mobile crowdsensing campaigns that draw on the opportunistic networking paradigm, as practised in delay-tolerant networks but also in the emerging device-to-device communication mode in cellular networks. In particular, we ask how mobile users can be optimally selected in order to generate the required space-time paths across the network for collecting data from a set of fixed locations. The users hold different roles in these paths, from collecting data with their sensing-enabled devices to relaying them across the network and uploading them to data collection points with Internet connectivity. We first consider scenarios with deterministic node mobility and formulate the selection of users as a minimum-cost set cover problem with a submodular objective function. We then generalize to more realistic settings with uncertainty about the user mobility. A methodology is devised for translating the statistics of individual user mobility to statistics of space-time path formation and feeding them to the set cover problem formulation. We describe practical greedy heuristics for the resulting NP-hard problems and compute their approximation ratios. Our experimentation with real mobility datasets (a) illustrates the multiple tradeoffs between the campaign cost and duration, the bound on the hopcount of space-time paths, and the number of collection points; and (b) provides evidence that in realistic problem instances the heuristics perform much better than what their pessimistic worst-case bounds suggest.**

I. Introduction

Mobile crowdsensing has emerged over the last decade as a powerful mechanism for generating collective knowledge about a phenomenon or condition of interest through contributions of sensor data by individuals [7]. These data may be measurement samples, text and even photographs or video clips and are typically generated by mobile devices with sensing capabilities such as smartphones. The aggregation and processing of these data gives rise to diverse services ranging from traffic jam prediction and parking space management to environmental monitoring and social journalism.

Mobile crowdsensing implementations usually involve three main actors: the end-users that contribute the sensor data (*data providers*), the Service Provider (SP) processing the collected data to generate a service out of them, and the end-users that subscribe to this service (*data consumers*), most often through a mobile application running on their devices. A common challenge for most of these implementations is to identify those end-users who can contribute most value to the service and motivate their participation. In our work, the motivation part is served by monetary incentives. The sensor data providers announce the fees they charge for contributing to the data collection campaign and it is then up to the service provider to recruit those who bear the highest value for money for the its service.

Interestingly, the majority of the current literature assumes that the end-users use the cellular network resources for transferring data to the SP as soon as these are generated by their devices' sensors. We argue that there are at least three good reasons not to do so –and instead prefer alternative transport alternatives for the sensor data. Firstly, this practice implies a significant cost in terms of battery and data subscription plan consumption; if this cost is rationally reflected into the fees the users claim, it raises considerably the cost of the crowdsensing campaign. Secondly, and depending on the type of the collected data (*e.g.*, the quality photos sought in [11]), crowdsensing campaigns generate additional workload for the cellular network. This workload is anything but negligible considering the projected scale of these campaigns by the time crowdsensing reaches maturity and that many of these campaigns will be taking place during the network busy hours. Thirdly, on a more constructive note, there are alternative transport paradigms for sensor data that can alleviate these concerns. Opportunistic networking is viewed as a promising complement to the cellular networks in different respects, *e.g.*, for offloading delay-tolerant traffic load from them [10]. On the other hand, the device-to-device (D2D) communications mode (*e.g.*, [6]) essentially introduces a multihop opportunistic layer straight into the cellular network architecture, through which mobile devices can communicate directly, without the intervention of base stations.

The realization of crowdsensing over an opportunistic networking transport layer, where the assumption of continuous connectivity to the Internet breaks, radically transforms the possible roles and value of end-users. Hence, a user who senses data from various locations may not necessarily have a way to transfer them by her own means *only* to the SP; whereas users who do not sense any data themselves, may bear high value for a crowdsensing campaign as relays of data of other users thanks to their encounters with them. Our paper looks into the problem of sensor selection in this fundamentally different setting. We formulate the underlying optimization problem, analyze its theoretical properties, and propose practical algorithms for solving it.

A. Related work

Mobile crowdsensing and the closely-related paradigms such as participatory sensing and human/people-centric sensing have motivated much research work over the last decade. For recent surveys of this work and the remaining open challenges, the interested reader is referred to [7] and [14].

On the contrary, much sparser is the literature on the realization of crowdsensing over opportunistic networks. We are aware of three studies under this thread, *i.e.*, [15] [16] [17]. Common to all of them is the assumption that the sensed data are generated randomly or periodically by the mobile users, as the case is with health- or fitness-monitoring applications; none of them considers location-dependent data raising coverage concerns. The objective then, as with the original routing and data dissemination scenarios in opportunistic networks, is to deliver the maximum amount of these data to the sink(s), as fast as possible and with minimum replication overhead.

More specifically, in [16] the mobile sensing nodes are distinguished into those acting exclusively as data providers and those high-end ones also serving as sinks (*e.g.*, upload information to the Internet). The paper proposes two heuristic schemes for the delivery of sensed data to the sinks, which are shown to be performing comparably in terms of message delivery probability and delay, and complementarily with respect to the management of the nodes' storage space and message overhead. In [17] the setting is similar only now the nodes decide strategically upon encounters with other nodes whether they will contribute to the forwarding of a particular data item towards the single network sink or not. The authors assume that the sink rewards with credit only nodes that deliver messages to it; hence, they can approach the pairwise encounters as instances of two-person cooperative games, during which the two nodes need to identify which set of messages are worth exchanging with each other. They apply the Nash bargaining Theorem to obtain the optimal solution to these games and propose a greedy algorithm that iteratively selects the pair of files maximizing the utility of the exchange, within the constraints set by the battery resources of the devices. Finally, the work in [15] lies closer to the participatory sensing framework in [12], by including an explicit sensor recruitment phase, initiated well in advance of the data collection phase. Contrary to the other two studies, the sensed data are location-dependent but this dependence is accounted for indirectly via the recruitment phase: the nodes selected for the sensing task are those visiting more frequently and for longer time intervals the locations of interest, as this is inferred from historical data about the node mobility patterns.

B. Our contributions

We address the problem of user/sensor selection in the context of mobile crowdsensing campaigns that draw on the opportunistic networking paradigm. The campaign aims at the collection of location-based data, while users, depending on their mobility patterns, may undertake different roles, *i.e.*, sense or relay the original sensor data. To the contributions of our paper count the following:

- We formulate the problem under both deterministic and stochastic user mobility as instances of the minimum cost set cover problem with submodular objective functions (Sections III-A and IV-A).
- We describe a method for computing the probability with which different space-time paths are realized across the opportunistic network out of data about the nodes' mobility patterns in the past (Sections IV-B and IV-C).
- Since the problem is NP-hard, we propose practical greedy heuristics and derive the approximation ratios they achieve (Sections III-B and IV-E).
- We evaluate the overall approach over experimental node mobility datasets to demonstrate the multiple tradeoffs between the campaign parameters and provide evidence that the performance of the greedy heuristic in real problem instances is much better than what their worst-case analysis implies (Section V).

As in [15]- [17], our work considers the general mobile crowdsensing paradigm rather than a particular application and draws on contact traces to evaluate the performance of the proposed algorithms. However, contrary to [16] and [17], in our case there are certain locations that have to be sensed/covered. Our recruitment process, unlike the one in [15], explicitly considers the coverage of and the fees charged by end-users in the recruitment process. Moreover, the payments are monetary and made to all nodes contributing to the delivery of the sensor data to the sinks, as either sensor nodes or relays. Consequently, relevant performance metrics to our work are the campaign's coverage and cost rather than the message-related ones (delivery, delay, overhead) used for assessing more typical DTN applications.

II. System model and problem formulation

A. System model

The two main actors in the model we consider are the mobile end-users and an entity that organizes the mobile crowdsensing campaign, hereafter called the campaign organizer (CO)[1]. The CO is interested in collecting data from a set of Points of Interest (PoIs) within a time interval $[t_1, t_2]$; without loss of generality, assume $t_1 = 0$ and $t_2 = T_c$. These data can be collected by mobile users carrying sensing devices as long as they accept to participate in the crowdsensing campaign. Practically, this may imply downloading and running a custom application on their smartphones.

Let $\mathcal{L} = \{l_1, l_2, \ldots, l_L\}$, be the set of PoIs and $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ the set of mobile users who could potentially be engaged in the crowdsensing task. Practically, $\mathcal{U}$ could coincide with the full set of mobile users traversing the area of interest or a subset of those fulfilling some criteria such as the possession of a sensing-capable smartphone. The mobile users move in different manner so that each user u_k visits a subset of PoIs $S_k \subseteq \mathcal{L}$, $k = 1, 2, \ldots, N$ over time T_c. In parallel, u_k may encounter another user u_m participating

[1] In general, the CO entity does not need to coincide with the SP entity but may be a third entity to which the user selection task is delegated.

contact id	involved nodes		contact start time	contact end time
...	...	...	...	...
con_1	u_1	l_1	t_{1s}	t_{1e}
con_2	u_1	c_1	t_{2s}	t_{2e}
con_3	u_1	u_5	t_{3s}	t_{3e}
con_4	u_4	l_4	t_{4s}	t_{4e}
con_5	u_1	u_4	t_{5s}	t_{5e}
con_6	u_2	l_l	t_{6s}	t_{6e}
con_7	u_4	l_3	t_{7s}	t_{7e}
con_8	u_3	u_4	t_{8s}	t_{8e}
con_9	u_2	u_5	t_{9s}	t_{9e}
con_{10}	u_3	c_1	t_{10s}	t_{10e}
con_{11}	u_5	c_1	t_{11s}	t_{11e}
...	...	...	...	...

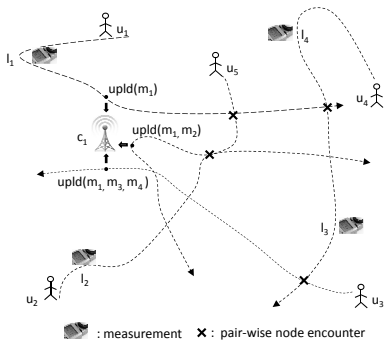


Fig. 1. Example of a crowdsensing campaign with a single collection point over an opportunistic network made up by the encounters on the left side. Data from PoI l_1 are uploaded by the same user who collected them. Data from all other points, l_2, l_3, and l_4 reach the collection point with the help of the intermediate users u_3 and u_5, who do not themselves collect any data.

in the crowdsensing campaign, and exchange with her data from any PoIs they have visited up to that time (essentially subsets of S_k and S_m, respectively). Eventually, the data from the different PoIs need to reach the server(s) of the CO (SP) at the fixed network and this is possible through a set $\mathcal{CP} = \{cp_1, cp_2, ..., cp_C\}$ of $C = |\mathcal{CP}|$ collection points such as WiFi APs scattered across the area of interest and accessible to the campaign participants free-of-charge.

Each user u charges a fee c_u for her participation in the campaign accounting for the extra consumption of her battery but also potential cognitive effort and time demanded by the data collection process. This is communicated to the CO in response to its campaign announcement, *e.g.*, over a crowdsourcing-type of interface (*e.g.*, [1]). The CO then seeks to identify and recruit those users who will collectively gather the data from all PoIs at the minimum possible cost. To formulate the problem the CO faces, consider the set of PoIs. Data from every PoI $l \in \mathcal{L}$ can reach the data collection points through a set of space-time paths (STPs) realized through the mobility of nodes and their time-ordered encounters within time T_c. Each STP involves the user serving as the sensing node and zero or more other users relaying the data till they reach anyone of the collection points. For example, in Fig. 1 $p_1 = \{l_1, u_1, c_1\}$, $p_2 = \{l_1, u_1, u_5, c_1\}$ and $p_3 = \{l_1, u_1, u_4, u_3, c_1\}$ are three different STPs, over which data from PoI l_1 can reach the collection point c_1. In all three cases, the data are collected by user u_1 but relayed by different nodes before being eventually uploaded to c_1.

B. Problem formulation - deterministic mobility

For the purposes of our modeling formulation, it is convenient to consider the sequences of users who *collectively* realize these STPs. To facilitate notation, we call them *data-collection paths* and abbreviate their subset that realizes STPs for a particular PoI l as DCP_l. In the earlier example of Fig. 1, the user sequences $\{u_1\}$, $\{u_1\ u_5\}$, and $\{u_1\ u_4\ u_3\}$ are data-collection paths in DCP_{l_1} corresponding to the STPs p_1, p_2, and p_3, respectively. Note that the DCP modeling construction already entails the chronological order of the involved encounters and abstracts away the actual location of the PoIs and the data collection points.

With this notation at hand, the optimization problem the CO seeks to solve can be written as:

$$\begin{aligned} \text{minimize} \quad & \sum_{u \in \mathcal{U}} x_u c_u \\ \text{s.t.} \quad & \sum_{p \in DCP_l} \prod_{u \in p} x_u \geq 1 \ \ \forall l \in \mathcal{L} \qquad (P1) \\ & x_u \in \{0, 1\} \end{aligned}$$

The set of L constraints in the second line of $(P1)$ require that the selection of users by the CO needs to be realizing *at least* one of the data-collection paths for each PoI.

C. Mobile crowdsensing over the cellular infrastructure: a plausible benchmark

The mobile crowdsensing implementation over multihop opportunistic networks generalizes its more common operation over the cellular network infrastructure. There, a mobile user directly uploads any data it collects from PoIs through the network radio access point (*e.g.*, 2/2.5G base station, 3G NodeB or LTE eNB) she is attached to. The data-collection paths are singletons and the respective STPs are the ordered pairs emerging from the DCPs when these are prefixed by the PoI id. Each user u_k is univocally linked to a set of PoIs, S_k, it can cover *independently* from other users and the user selection task simplifies to the familiar minimum-cost set cover problem with fixed user-specific cost values.

$$\begin{aligned} \text{minimize} \quad & \sum_{u \in \mathcal{U}} x_u c_u \\ \text{s.t.} \quad & \sum_{u: l \in S_u} x_u \geq 1 \ \ \forall l \in \mathcal{L} \qquad (P2) \\ & x_u \in \{0, 1\} \end{aligned}$$

It is straightforward to see that:

Proposition 1. *The optimal solution of (P2) serves as a tight lower bound to the optimal solution of (P1).*

Proof. Consider the optimal solution of (P2), $(P2)_{opt}$, consisting of one or more users who together cover all PoIs at minimum cost. When we replace the cellular network with an opportunistic one, the users that can upload data to the collection point(s) are, in the general case, only a subset (even the null one) of $(P2)_{opt}$. Hence, more users are needed to serve as data relays and/or uploaders and the optimal solution of (P1), $(P1)_{opt} \neq (P2)_{opt}$ with $|(P1)_{opt}| \geq |(P2)_{opt}|$ and $\sum_{u \in (P1)_{opt}} c_u \geq \sum_{u \in (P2)_{opt}} c_u$. Clearly, $(P1)_{opt}$ will coincide with $(P2)_{opt}$, if the set of users in $(P2)_{opt}$ suffices to generate valid STPs for all PoIs, without the need for additional users. An extreme scenario is that *all* users in $(P2)_{opt}$ encounter *themselves* the collection point(s) within the duration T_c of the crowdsensing campaign and after hitting the PoIs they cover as part of $(P2)_{opt}$. In more realistic scenarios, the two solutions also coincide when the time order of node pairwise encounters is such that a non-empty subset

$x \in (P2)_{opt}$ can upload to the collection points data from the residual $(P2)_{opt} \setminus x$ users, after encountering them and copying their data. Since these scenarios are generally feasible, the bound is tight. □

III. User recruitment through DCP selection: a greedy heuristic

A. Minimum-Cost Set Cover (Re-)Formulation

We can draw on the definition of data-collection paths to re-formulate the problem (P1) as a generalized *Set Covering* problem, involving only linear constraints (over integer variables) along with a submodular *cost function* over feasible solutions. In particular, define $\mathcal{P} = \cup_{l \in \mathcal{L}} DCP_l$ to be the set of all DCPs; each element $P \in \mathcal{P}$ constitutes a DCP for at least one PoI from $\mathcal{L}$. Under this definition, the CO seeks $\mathcal{Q} \subseteq \mathcal{P}$ of *minimum total cost*:

$$Z(\mathcal{Q}) = \sum_{u \in (\cup_{P \in \mathcal{Q}} P)} c_u, \tag{1}$$

such that, for every $l \in \mathcal{L}$, there exists at least one $P \in \mathcal{Q} \cap DCP_l$. Notice that this is indeed a minimum cost *Set Covering* formulation for the problem[2], wherein we choose data-collection paths to "cover" the PoIs. The most significant feature of this formulation is the objective function, $Z(\cdot)$ in (1); it is a *submodular* function over the space of feasible solutions. In particular, for any two subsets $\mathcal{Q}_1$, $\mathcal{Q}_2$ of $\mathcal{P}$, Z satisfies: $Z(\mathcal{Q}_1 \cup \mathcal{Q}_2) + Z(\mathcal{Q}_1 \cap \mathcal{Q}_2) = Z(\mathcal{Q}_1) + Z(\mathcal{Q}_2)$.

Introducing an additional binary variable, y_P for each $P \in \mathcal{P}$, the problem can be stated as an Integer Linear Program involving only linear constraints and a linear objective function, as follows:

$$\begin{aligned} \text{minimize} \quad & \sum_{u \in \mathcal{U}} c_u x_u & & \\ \text{s.t.} \quad & \sum_{P \in DCP_l} y_P \geq 1 & \forall l \in \mathcal{L} \qquad & (P3) \\ & x_u - y_P \geq 0 & \forall P \in \mathcal{P}, \forall u \in P & \\ & x_u, y_P \in \{0,1\} & \forall P \in \mathcal{P}, \forall u \in \mathcal{U} & \end{aligned}$$

The first set of constraints ensures that at least one data collection path is chosen for every PoI $l \in \mathcal{L}$. The second set of constraints forces the selection of user $u \in \mathcal{U}$ ($x_u = 1$), whenever a path P with $u \in P$ is chosen ($y_P = 1$).

B. A greedy heuristic for (P3)

(P3) comes under the category of NP-hard problems; hence the question arising is what can be said about its approximability and what could be an acceptable algorithm for (sub-optimally) tackling it. To begin with, for the generic Set Cover with a submodular cost function, the recent primal-dual algorithm of Koufogiannakis and Young [9] yields a Δ-approximation, where Δ corresponds to the maximum number of variables in each linear *"covering"* constraint (first set of constraints in (P3)). Whereas this approximation is particularly attractive for problem instances such as Vertex Cover, where $\Delta = 2$, the number of variables per constraint equation in (P3) equals the number of DCPs per PoI. This is highly variable and grows fast with the number of mobile users N and the hopcount of the respective STPs.

Therefore, a practically more promising alternative is provided by the following greedy heuristic.

Algorithm 1 Greedy heuristic for sensor selection under deterministic user mobility

1: $\mathcal{Q} \leftarrow \emptyset$; $U \leftarrow \emptyset$;
2: **while** $\exists l \in \mathcal{L} : DCP_l \cap \mathcal{Q} = \emptyset$ **do** :
3: $\quad P \leftarrow \arg \min_{P \in \mathcal{P} \setminus \mathcal{Q}} \Big[c(P \,|\, U)/|L(P)| \Big]$;
4: $\quad \mathcal{Q} \leftarrow \mathcal{Q} \cup \{P\}$; $U \leftarrow U \cup P$;
5: **return** $\mathcal{Q}$;

The heuristic is a straightforward adaptation of the well known greedy Set Covering heuristic by Chvátal [5], to the case of a submodular cost function. In each step the algorithm selects the DCP minimizing the ratio of the additional fees that have to be paid to new users involved in it, over the set of new PoIs it covers with respect to already selected DCPs. In describing the algorithm, we have defined $L(P) = \{ l \in \mathcal{L} \,|\, P \in DCP_l \}$, to be the set of PoIs covered by path $P \in \mathcal{P}$ and $c(P \,|\, U) = \sum_{u \in P \setminus U} c_u$ for any $U \subseteq \mathcal{U}$ and $P \in \mathcal{P}$ to be the excess cost related to users in P but not in U.

Proposition 2 shows that Algorithm 1 achieves a factor $|\mathcal{L}|$-approximation, *i.e.*, it is independent of the input set, $\mathcal{P}$, of data-collection paths and the cardinalities of $\{DCP_l\}$, $l \in \mathcal{L}$. This renders the greedy algorithm more robust than the primal-dual one of [9] in terms of worst-case performance, in a situation where we might need to generate $\mathcal{P}$ according to exogenous restrictions (time, processing power, memory). The actually generated family $\mathcal{P}$ affects the cost of the optimum solution to the problem, $c(\mathcal{Q}^*)$, but the algorithm's worst-case performance depends solely on $c(\mathcal{Q}^*)$ and $|\mathcal{L}|$.

Proposition 2. *The Greedy algorithm for minimum cost Set Cover with submodular cost function achieves factor $|\mathcal{L}|$-approximation of the optimum cost, where $|\mathcal{L}|$ denotes the number of ground elements (in our case, PoIs) to be covered.*

Proof. The proof amounts to showing that, in every step, the Greedy algorithm increases the current (partial) solution's cost by at most $c(Q^*)$, where Q^* denotes the optimum solution. Since the algorithm covers at least one more point from $\mathcal{L}$ at each step, the result follows.

For each step, s, let $P_s \in \mathcal{P}$ denote the path chosen at step s. Accordingly, define $\mathcal{Q}_s = \{P_1, P_2, \ldots, P_s\}$ to be the partial solution after step s, and $U_s = \cup_{s' \leq s} P_s$ and let $L_s \subseteq \mathcal{L}$ be the subset of PoIs that are covered after step s. Finally, set $\mathcal{Q}_0 = U_0 = L_0 = \emptyset$.

For $s = 1$, our argument holds, *i.e.*, $c(P_1 \,|\, \emptyset)/|L(P_1)| \leq c(\mathcal{Q}^*)$, because, otherwise, P_1 does not minimize the ratio of cost over number of covered points. At any step s, a

[2]It may also be considered as an – equivalent – *Hitting Set* Problem, wherein we seek a solution that *"hits"* every set STP_l, for every $l \in \mathcal{L}$.

sub-problem remains to be solved, concerning $|\mathcal{L} \setminus L_{s-1}|$ PoIs, where $|\mathcal{L} \setminus L_{s-1}| < |\mathcal{L} \setminus L_{s-2}|$. Moreover, each of the remaining paths $P \in \mathcal{P} \setminus \mathcal{Q}_{s-1}$ has cost $c(P \,|\, U_{s-1}) \leq c(P)$. This sub-problem has an optimum solution $\mathcal{Q}^*_{s-1}$, of total cost $c(\mathcal{Q}^*_{s-1} \,|\, U_{s-1})$ (we abuse slightly the notation $c(\cdot|\cdot)$ here). Then, $c(\mathcal{Q}^*_{s-1} \,|\, U_{s-1}) \leq c(\mathcal{Q}^*)$, because $\mathcal{Q}^*$ is also feasible for the sub-problem. The path P_s, chosen at step s, must satisfy $c(P_s \,|\, U_{s-1})/|L(P_s) \setminus L_{s-1}| \leq c(\mathcal{Q}^*_{s-1} \,|\, U_{s-1})$ for, otherwise, it does not minimize the ratio of cost over number of PoIs that it covers. Then, to conclude the proof, if $\mathcal{Q} = \{P_1, P_2, \ldots, P_{|Q|}\}$ is the solution output by the algorithm, we have:

$$\begin{aligned} c(\mathcal{Q}) &= \sum_{s=1}^{|\mathcal{Q}|} c(P_s \,|\, U_{s-1}) \\ &\leq \sum_{s=1}^{|\mathcal{Q}|} c(\mathcal{Q}^*_{s-1} \,|\, U_{s-1}) \cdot |L(P_s) \setminus L_{s-1}| \\ &\leq c(\mathcal{Q}^*) \times \sum_{s=1}^{|\mathcal{Q}|} |L(P_s) \setminus L_{s-1}| = |\mathcal{L}| \cdot c(\mathcal{Q}^*) \end{aligned}$$

where the latter equality stems from the fact that $|L(P_s) \setminus L_{s-1}| \geq 1$, for any step $s = 1, \ldots, |\mathcal{Q}|$. That is, the algorithm covers at least one new PoI in each step. The sum has to be totalling $|\mathcal{L}|$. □

A simple implementation of the Greedy algorithm described above requires $O\Big(|\mathcal{L}|^2(|\mathcal{U}| + |\mathcal{P}|)\Big)$ steps. The outer loop requires $O(|\mathcal{L}|)$ iterations in the worst-case; in each iteration we need to scan / update two matrices of $O(|\mathcal{L}| \cdot |\mathcal{U}|)$ and $O(|\mathcal{L}| \cdot |\mathcal{P}|)$ sizes, respectively.

IV. Accounting for the uncertainty of node mobility

A. Annotating DCPs with coverage probabilities

The analysis in section II-A is subject to the strong assumption that the user trajectories over the area of interest can be perfectly known/predicted. This is the case with only a limited set of opportunistic network instances such as those realized by track-based transportation system nodes.

On the contrary, in almost all envisaged realizations of mobile crowdsensing, there is no control over the way the end-user nodes move. Nevertheless, there is almost always structure in their mobility patterns and multi-timescale periodicities, induced by their daily routines and social activities (*e.g.*, trips from home to workplace and back and visits to friends, relatives and recreation places). The challenge for the recruitment phase of the campaign is then to take advantage of historical information about these mobility patterns so as to optimize its selection of users.

Technically, the space-time paths between PoIs and collection points are realized probabilistically rather than deterministically over the lifetime of the campaign so that a given DCP P covers a PoI, say l, with probability q_{Pl} that depends directly on whether the involved mobile nodes *actually* visit the PoI and encounter with each other. As a result, one plausible objective emerging for the recruitment process is the minimization of the payments that have to be made to the selected nodes to achieve a probabilistic guarantee about the coverage of the PoIs.

More formally, the optimization problem faced by the CO entity can be written as:

$$\begin{aligned} \text{minimize} \quad & \sum_{u \in \mathcal{U}} c_u x_u \\ \text{s.t.} \quad & \sum_{P \in DCP_l} y_P \cdot q_{Pl} \geq 1 \quad \forall l \in \mathcal{L} \qquad (P4) \\ & x_u - y_P \geq 0 \qquad \forall P \in \mathcal{P}, \forall u \in P \\ & x_u, y_P \in \{0, 1\} \qquad \forall P \in \mathcal{P}, \forall u \in \mathcal{U} \end{aligned}$$

Compared to (P3), the first set of constraints now demand that the *expected* number of paths covering each PoI be at least one. Practically, the *coverage probabilities* $\{q_{Pl}\}$ can be computed with the help of information about the nodes' mobility in the past. Such information may come from logs of the GPS or other positioning systems exported by the application itself or from other sources such as check-ins in online (location-based) social networking sites [4].

B. Representation of node mobility data

The first step in this direction is a concise representation of the nodes' mobility profiles. To this end we partition the two main dimensions of this mobility, *i.e.*, time and space, into *T* intervals and *S* blocks, respectively, and represent each user with an SxT row-stochastic user location probability matrix f_u; each element $f_u(s,t)$ represents the time-varying probability that the user u lies in space block s during time interval t.

The number of space blocks is determined by the size of the area of interest and the radio coverage of the PoIs and sinks, *i.e.*, the distance over which measurement data can be collected from (resp. uploaded to) them. The assumption is that no more than one PoI or sink lies within a given space block so that $S > L$. The number of columns T is computed as $T = T_c/t_s$, where t_s equals the time step over which the historical data are aggregated and processed to derive the time-dependent probability distribution of the node's locations $f_u(s,t), 1 \leq s \leq S$. The value of t_s is either chosen a priori or indirectly induced by the frequency of the location reports/measurements, as exported by GPS or other positioning technologies. In general, smaller t_s values result in higher precision but also higher storage and processing requirements for the node mobility representation.

C. Computation of STPs and their formation probabilities

This step involves the derivation of space-time paths of the form (PoI_l, U_P, CP_k), where $U_P \subseteq \mathcal{U}$; namely, space-time paths that originate from some PoI, end at a collection point and are realized over a subset of user nodes. This computation is carried out iteratively over successive time intervals drawing on the user location probability matrices $\{f_u\}$ and builds a

list L_{dcp} of all space-time paths that can emerge with non-zero probability. At any point in time, L_{dcp} lists a number of "open" paths of the form (PoI_l, U_P), a subset of which will eventually "close" by the end of the campaign with a suffix node corresponding to a collection point; these are the space-time paths of interest. Besides the derivation of the paths as such, this processing step continuously updates their cumulative formation probabilities.

Technically, the following four processing steps are repeated upon each time interval.

1) Search for visits of user-nodes at space blocks hosting PoIs: Such visits imply that data from a PoI, say l, can be collected from a user, say u, at time interval t and correspond to a non-zero value of $f_u(l,t)$. Two possibilities exist:

- This is the first time u visits l: then a new DCP entry $[(PoI_l, u), f_u(l,t)]$ is added to the list of possible DCPs.
- There is already an entry $[(PoI_l, u], f_P)$ in the list due to a visit of u at space block l in the past: then the second field of the entry (the formation probability of the DCP) is updated to $1-(1-f_P)(1-f_u(l,t))$. This value equals the probability that u has run across PoI_l up to and including the interval t.

2) Search for encounters giving rise to new possible paths: For each entry $(P, f_P) \in L_{dcp}$, we are considering the most recently added user-node v and the set of positive values $S_P(v,t) \subseteq S$ in its location probability matrix column $f_v(:,t)$; this set corresponds to locations (space blocks) that user v visits at time interval t with positive probability. We then iterate over users $z \in U \setminus v$, and insert a new entry (P', f'_P) in the L_{dcp} when $Z = S_P(v,t) \cap S_P(z,t) \neq \emptyset$. In the new entry, $P' = (P, z)$ and $f'_P = f_P \prod_{m \in Z} f_v(m,t) f_z(m,t)$. This is the probability that in time interval t, user node v encounters node z in any of their common possible locations over this time interval.

3) Search for encounters that increase the cumulative formation probability of existing paths: The computations are slightly more involved than in step 2 above. This time, we need to: (a) consider the formation probabilities over time of the subpath that arises when the last node that was appended to the path P is removed; (b) subtract from them the part of these probabilities that has already been factored in the computation of the current formation probability value of P; (c) inflate the remainder of this probability by the probability of an encounter between the ultimate and the penultimate nodes in P over the current time interval.

4) Search for visits to locations hosting sink nodes: These visits enable uploading the measurements to the collection points, essentially closing a path in list L_{dcp}. This time, for each entry $(P, f_P) \in L_{dcp}$, the check is over the positive values in the column $f_v(:,t)$. If there is a non-zero value p' in any space block hosting a collection point CP, then (P, CP) is promoted to a candidate STP with formation likelihood $f_P \cdot p'$.

On the computational front, the first task, which seeks for visits of user-nodes at space blocks hosting PoIs, requires reading L matrix entries per user per time unit, for an overall complexity of $O(TNL)$. The second task involves reading $O(L)$ matrix entries in $O(N)$ matrices R per existing possible DCP per time unit. Since their worst-case count is $O(NL)$, the overall complexity is $O(TN^2L^2)$. Finally, the last task requires $O(S)$ readings per user in the matrix R_u, namely $O(NS)$ overall. The processing complexity of this stage is reduced when the hopcount of candidate DCPs is bounded. Even when the opportunistic protocol does not set a hard bound on the hopcount of STPs (hence, DCPs), the CO may choose to filter out STPs with hopcount higher than some threshold. As the hopcount of permitted DCPs shrinks, fewer matrices need to be read and fewer checks for redundant paths have to be carried out during this processing step.

D. Extraction and merging of data-collection paths

The outcome of the earlier step is a list of STPs of the form $[PoI_l, U_P, CP_k]$ and their formation probabilities. The data-collection path of the STP corresponds to the user-node set U_P realizing the STP. The important remark is that for the purposes of the formulations in (P3) and (P4), all STPs involving PoI_l and *any permutation* of U_P together with any CP $\in \mathcal{CP}$, map to the same DCP, *i.e.*, a single member of DCP_l. Hence, if π is the set of permutations of U_P, the probability q_{Pl} that U_P will cover PoI_l is

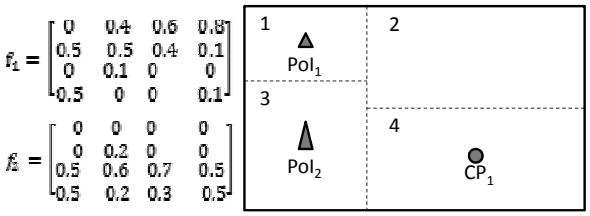
$$q_{Pl} = 1 - \prod_{dcp \in \pi, cp \in \mathcal{CP}} (1 - f_{(PoI_l, dcp, cp)})$$

At the end of this task, for each PoI l we have a set of (P, q_{Pl}) entries, which could directly be used in the problem formulation in (P4). Note that the same DCP, *i.e.*, the same set of users when collectively recruited, may be covering two or more PoIs with different coverage probabilities.

Example: Figure 2a exemplifies the computation of STPs and extraction of DCPs for a toy example with two users u_1 and u_2 moving in a four-block area with two PoIs and a single collection point. The campaign duration T_c is split into four time intervals, the user location probabilities in each one of those being given by matrices f_i, $i = 1, 2$. Then Table 2b lists all the STPs that emerge over T_c along with the evolution of their *cumulative* formation probabilities. For instance, data from PoI_1 can be collected by u_1 at time intervals 2,3 and 4, so that the coverage probability of PoI_1 by u_1 over the duration of the campaign equals 0.952. Ten different STPs originate from the two PoIs but only four of them yield data-collection paths. User u_1 covers both PoIs, with probabilities 0.076 and 0.01, respectively, whereas u_2 covers only PoI_2 but with a significantly higher probability (0.625). In this example, even if both users are recruited, the expected number of DCPs is < 1 for the two PoIs.

E. A greedy heuristic for (P4)

Despite the introduction of the coverage probabilities $\{q_{Pl}\}$, (P4) remains a monotone covering problem with a submodular objective function so that the findings in [9] are still valid. Yet the new worst-case approximation ratio Δ' is looser than



a. Area and user location probability matrices f_u

STPs	Time blocks j				MCPs
	j = 1	j=2	j = 3	j = 4	
PoI_2-u_2	$A_2(3,1) = 0.5$	0.8	0.94	0.97	-
PoI_1-u_1		$A_1(1,2) = 0.4$	0.76	0.952	-
PoI_2-u_1		$A_1(3,2) = 0.1$	0.1	0.1	-
PoI_2-u_2-cp_1		0.1	0.31	0.625	$\{u_2\}$
PoI_2-u_2-u_1		0.08	0.08	0.123	-
PoI_2-u_1-cp_1				0.01	$\{u_{11}$
PoI_1-u_1-cp_1				0.076	$\{u_{11}$
PoI_2-u_2-u_1-cp_1				0.08	$\{u_1\, u_2\}$
PoI_1-u_1-u_2				0.038	-
PoI_2-u_1-u_2				0.005	-

b. STPs, their cumulative formation probabilities, and related DCPs

Fig. 2. Simple example for the extraction of DCPs and the computation of their coverage probabilities: L=2, N=2, $|\mathcal{CP}| = 1$, T=4 and S=4.

its counterpart for (P3) since the uncertainty about the node mobility results in at least as many DCPs per PoI as in (P3), *i.e.*, $\Delta' \geq \Delta$.

Likewise, it is straightforward to adapt the greedy algorithm in Section III-B for addressing (P4): Compared to its counterpart for (P3), Algorithm 2 features one additional variable Cov_l per PoI. The variables log the expected number of users covering each PoI as the algorithm runs and serve two purposes. First, they yield the stopping condition for the algorithm iterations; secondly, they shape the denominator of the ratio that assesses the utility of each candidate DCP and points to the next selection. Proposition 3 address the capacity of Algorithm 2 to approximate the optimal solution of (P4)

Algorithm 2 Greedy heuristic for sensor selection under stochastic user mobility

1: $\mathcal{Q} \leftarrow \emptyset;\ U \leftarrow \emptyset;\ Cov_l = 0\ \forall l \in \mathcal{L}$
2: **while** $\exists l \in \mathcal{L} : Cov_l < 1$ **do** :
3: $\quad P \leftarrow \arg\min_{P \in \mathcal{P} \setminus \mathcal{Q}} \Big[c(P \,|\, U) / \sum_{l: Cov_l < 1} q_{Pl} \Big];$
4: $\quad \mathcal{Q} \leftarrow \mathcal{Q} \cup \{P\};\ U \leftarrow U \cup P;$
5: $\quad Cov_l \leftarrow Cov_l + q_{Pl}\ \forall l \in L(P)$
6: **return** $\mathcal{Q}$;

Proposition 3. *The Greedy algorithm for sensor selection under stochastic user mobility achieves factor $|\mathcal{L}|/\rho$-approximation of the optimum cost, where $|\mathcal{L}|$ is the number of PoIs to be covered and $\rho = \min_{P,l:q_{Pl}>0} q_{Pl}$ is the minimum (non-zero) coverage probability over all (DCP, PoI) pairs.*

The proof is given in the extended version of this paper [2].

The worst-case complexity of Algorithm 2 is worse than that of Algorithm 1 as well. The outer loop in this case may require $O(|\mathcal{L}| \cdot |\mathcal{P}|)$ iterations, as in every such iteration, the coverage of any PoI may increase by only ρ. Strictly speaking, only $O(\rho^{-1} \cdot |\mathcal{L}|)$ iterations are needed, given that each PoI is fully covered by at most ρ paths. Generally, however, $\rho^{-1} = O(|\mathcal{P}|)$. In every iteration we need to scan / update two matrices of $O(|\mathcal{L}| \cdot |\mathcal{U}|)$ and $O(|\mathcal{L}| \cdot |\mathcal{P}|)$ sizes; this yields a total complexity of order $O\Big(|\mathcal{P}| \cdot |\mathcal{L}|^2 \cdot (|\mathcal{U}| + |\mathcal{P}|)\Big)$.

V. Evaluation of the recruitment algorithm

The aim of this section is to evaluate the performance of the greedy heuristics in realistic problem settings. Although the approximation ratios in III-B and IV-E are quite discouraging, it has been frequently observed in empirical studies of greedy algorithms that they tend to perform much better on specific benchmarks than suggested by their worst-case performance guarantee (see, for example, [8])

A. Performance determinants

The cost of the crowdsensing campaign is determined by: user-related factors that do not lie, at least originally, under the control of the CO such as the mobility patterns of users and the fees they charge for their contributions; the number of available collection points C for uploading the collected data; the duration T_c of the crowdsensing campaign; and the upper bound H_b on the hopcount of the STPs (resp. DCPs) that cover the PoIs. The campaign duration is almost entirely within the discretion of the campaign designer. Higher T_c values imply additional flexibility in the selection of end-user sensors, within the delay tolerance constraints of the application at hand each time.

B. Datasets and methodology

Methodologically, the evaluation is carried out over experimental datasets listing sequences of node encounters, *i.e.*, the kind of files that are used extensively for the study of opportunistic forwarding protocols. In particular, two mobility traces among those made publicly available by the Haggle project [13], have been used to emulate the way nodes encounter with each other and hit the static PoIs and collection points. The two traces combine mobile and static nodes and log pairwise encounters between both node types. The first trace, hereafter called the *Content* trace, was collected over an interval of approximately two months in the city of Cambridge, UK. It involves 36 mobile iMotes carried by students of the Cambridge University and 18 fixed nodes located at various places around the city such as pubs, shop windows, a supermarket and points at the commercial city center. The second trace, which will be referred to as *Inf06*, was collected within the dramatically smaller spatial and temporal coordinates of the Infocom '06 conference venue. It features 98 nodes; 78 of them are iMotes carried by conference participants and the remaining 20 are fixed nodes situated at various places in the conference hotel such as conference rooms, the bar, the concierge and the hotel elevators. Both traces have been preprocessed to cater for contact log asymmetries; more details about the traces are provided in [13].

For our evaluation purposes, the set of mobile nodes is mapped to the user set $\mathcal{U}$, whereas (sub)sets of the static nodes

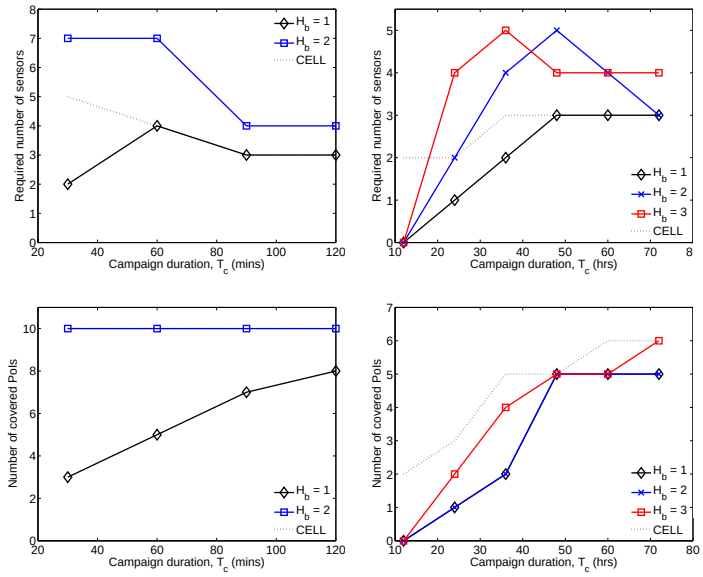


Fig. 3. Campaign cost (top) and covered PoIs (bottom) under the Greedy algorithm for the Inf06 (10 PoIs, left) and the Content (6 PoIs, right) traces: $c_u = 1\ \forall u \in \mathcal{U}$, C=1, variable bounds on STP hopcount.

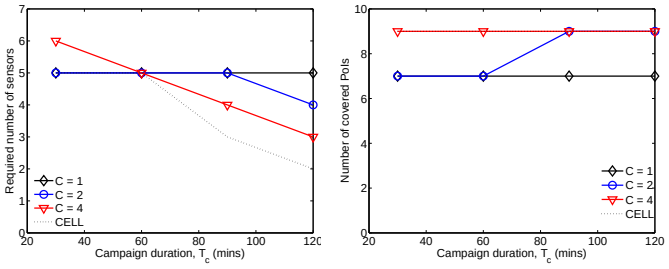


Fig. 4. Campaign cost and covered PoIs under the Greedy algorithm for the Inf06 (10 PoIs) trace: $c_u = 1\ \forall u \in \mathcal{U}$, $H_b = 2$, variable number of collection points.

are mapped to the PoI and collection point sets, $\mathcal{L}$ and $\mathcal{CP}$, respectively. We vary the campaign duration by extracting and working with varying-length parts of the traces. The format of the used datasets matches scenarios with deterministic knowledge about the mobility of user nodes. This lets us separate the performance assessment of the greedy user selection heuristic from the precision of inferring/computing space-time paths out of historical data about the users' mobility patterns (ref. Sections IV-B and IV-C); the latter is an independent issue with its own long research thread (for example, see [3] [4]). In all plots we compare the solution furnished by the greedy heuristic for (P3) to the solution produced by the greedy heuristic for (P2),*i.e.*, the selection of users when the crowdsensing campaign is run over a cellular network, where end-users can upload data as soon as they collect them. OPP and CELL, where used, are legend abbreviations corresponding to the two solutions.

C. Results

1) Sensitivity to T_c and H_b: In these experiments, we vary the campaign duration T_c and the hopcount H_b of the realized paths. There is a single collection point and all users charge the

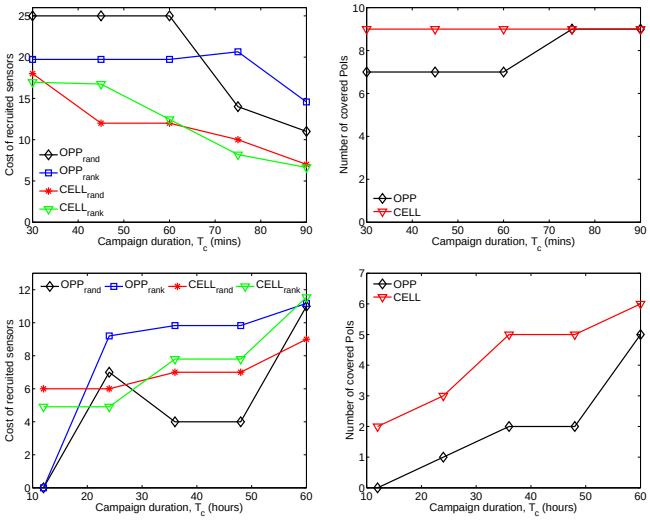


Fig. 5. Campaign cost and coverage under the Greedy algorithm for the Inf06 (top, 9 PoIs) and Content (bottom, 6 PoIs) traces: C = 2, H_b = 2, c_{min} = 1, c_{max}=7.

same unit fee for contributing to the campaign so that its cost coincides with the number of selected users. The campaign duration is in the order of hours (days) for the much denser (sparser) Inf06 (Content) dataset.

Overall, longer campaign durations enable the realization of more data-collection paths. When only singleton DCPs are allowed, nodes tend to cover more PoIs over time. The coverage increases and so does the cost as long as the additional PoIs are covered by "new" users who are not selected at smaller T_c values. As T_c grows further and more paths emerge letting the same user cover more than one PoI, the coverage improvement may come at reduced overall campaign cost, as shown in Fig. 4.a and Fig. 4.b. These trends get even clearer, in particular for the Content trace, as the bound on the paths' hopcount is relaxed. Now the rate at which new paths are realized as T_c grows is faster since the pairwise node encounters give rise to more two- and three-hop DCPs, some of which collectively cover several PoIs.

In all cases, the user selection made by the greedy heuristic costs only marginally more than what the campaign would cost if carried out over the cellular network infrastructure with immediate uploading of the collected data.

2) Varying the number of collection points: The increase of collection points stands as an alternative to lengthening the crowdsensing campaign duration or letting longer data-collection paths with respect to the achieved coverage of PoIs. This improves in all three ways and, of course, under their combined effect, as Fig. 4.c demonstrates.

The density of encounters in the Inf06 trace is such that, even with two-hop DCPs, the required number of users for covering the nine PoIs drops from six (if the campaign cannot last longer than half an hour) down to three (if the campaign organizer can wait for an interval of two hours). In all cases, these scores are almost as good as those achieved by the greedy algorithm for (P2).

3) Robustness to the distribution of the user fees: In the previous two sets of experiments, the fee charged by all users was the same. In this final set, we let users charge differently their potential participation in the campaign. In particular, we fix the range of charged fees to $[c_{min}\ c_{max}]$ and consider two alternatives for the way the user fees are distributed over this range. In the first one (*rand*), the user fees are randomly (*i.e.*, uniformly) distributed; in the second case (*rank*), we introduce positive correlation between the fee and the number of encounters a user gets involved in over the campaign lifetime. More specifically, we rank users in order of increasing number of encounters they have over this time. If $rank(u), u \in \mathcal{U}$ is the rank of user u in this respect, the fee she claims is

$$c_u = c_{min} + rank(u) \cdot (c_{max} - c_{min})/(N-1)$$

so that the most "social" user, the one who gets involved in most encounters, charges c_{max}, whereas the least social one charges c_{min}. This distribution models what might arise over time in these crowdsensing campaigns, namely attempts of users to relate their claims to the importance *they anticipate that they have* for the campaigns.

Looking at the experimentation results, the trend in the much denser Inf06 trace changes as the campaign spans longer time intervals (ref. Fig. 5). For campaigns up to 1-hr long, the algorithm selects the same user sets under both fee distribution alternatives. The higher campaign cost under the random fee distribution rather implies that the number of encounters is not the right cue for the users' importance/contribution to the campaign: namely, the users who actually help cover PoIs charge lower fees in the second case than in the random one, hence are among the less "social" in terms of number of encounters. On the contrary, as the campaign lasts longer, the algorithm exploits the diversity in the choice of paths and selects different paths/users in the two cases, in an attempt to retain the overall cost as small as possible.

In the much sparser Content trace, the number of encounters appears to be capturing better the relative importance of users for the campaign. Under *rank*, the algorithm cannot avoid choosing some of the more expensive users since these are the only ones who can cover certain PoIs. Hence, the campaign's cost is clearly higher when users consciously attempt to capitalize on their mobility and frequent encounters with other users. Only for longer campaigns (~ 3 hours), does the algorithm gain enough flexibility in terms of candidate DCPs to make choices of comparable cost under *rand*.

VI. Conclusions

We have looked into the problem of user/sensor selection in crowdsensing campaigns drawing on opportunistic networking techniques. we have focused on the coverage dimension of the crowdsensing campaign design problem since the opportunistic networking layer adds significant complexity and interest, of both theoretical and practical nature, to it. The underlying optimization problem has been formulated as a minimum-cost set cover problem with submodular cost function for scenarios of deterministic and stochastic user mobility. We have described practical greedy heuristics and derived their approximation ratios for the problem. Experimental evidence suggests that these heuristics perform far better than what their (worst-case) approximation ratios let hope for.

Acknowledgment

M. Karaliopoulos and I. Koutsopoulos acknowledge the support of the ERC08- RECITAL project, co-financed by Greece and the European Social Fund through the Education and Lifelong Learning Operational Program of the Greek National Strategic Reference Framework 2007-2013.

References

[1] Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome, 2014.

[2] Extended version of the paper with proofs. https://www.dropbox.com/l/nhgANbY7cQVweVZS1QmkBo, Jul 2014.

[3] D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Comput.*, 7(5):275–286, Oct. 2003.

[4] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proc. 17th ACM SIGKDD*, KDD '11, pages 1082–1090, 2011.

[5] V. Chvátal. A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

[6] D. Feng, L. Lu, Y. Yuan-Wu, G. Li, S. Li, and G. Feng. Device-to-device communications in cellular networks. *IEEE Communications Magazine*, 52(4):49–55, Apr. 2014.

[7] R. Ganti, F. Ye, and H. Lei. Mobile crowdsensing: current state and future challenges. *IEEE Communications Magazine*, 49(11):32–39, Nov. 2011.

[8] T. Grossman and A. Wool. Computational experience with approximation algorithms for the set covering problem. *European Journal of Operational Research*, 101(1):81 – 92, 1997.

[9] C. Koufogiannakis and N. E. Young. Greedy Δ-Approximation Algorithm for Covering with Arbitrary Constraints and Submodular Cost. *Algorithmica*, 66(1):113–152, 2013.

[10] Y. Li, M. Qian, D. Jin, P. Hui, Z. Wang, and S. Chen. Multiple mobile data offloading through disruption tolerant networks. *IEEE Trans. on Mobile Computing*, 13(7):1579–1596, Jul 2014.

[11] D. Quercia, R. Schifanella, and L. M. Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proc. 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 116–125, 2014.

[12] S. Reddy, D. Estrin, and M. Srivastava. Recruitment framework for participatory sensing data collections. In *Pervasive Computing*, volume 6030 of *LNCS*, pages 138–155. Springer Berlin Heidelberg, 2010.

[13] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. CRAWDAD data set cambridge/haggle (v. 2006-01-31). Downloaded from http://crawdad.org/cambridge/haggle/, Jan. 2006.

[14] M. Srivastava, T. Abdelzaher, and B. Szymanski. Human-centric sensing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1958):176–197, 2012.

[15] G. S. Tuncay, G. Benincasa, and A. Helmy. Participant recruitment and data collection framework for opportunistic sensing: A comparative analysis. In *Proc. 8th ACM MobiCom Workshop on Challenged Networks*, CHANTS '13, pages 25–30, 2013.

[16] Y. Wang and H. Wu. Delay/fault-tolerant mobile sensor network (dft-msn): A new paradigm for pervasive information gathering. *IEEE Trans. on Mobile Computing*, 6(9):1021–1034, September 2007.

[17] X. Xie, H. Chen, and H. Wu. Bargain-based stimulation mechanism for selfish mobile nodes in participatory sensing network. In *Proc. 6th IEEE SECON*, pages 72–80, June 2009.