SPREAD, a Crowd Sensing Incentive Mechanism to Acquire Better Representative Samples

Luis G. Jaimes
Department of Electrical Engineering
University of South Florida
Tampa, Florida 33620
Email: ljaimes@mail.usf.edu

Idalides Vergara-Laurens
Department of Computer Science and Engineering
University of South Florida
Tampa, Florida 33620
Email: ijvergara@mail.usf.edu

Alireza Chakeri
Department of Computer Science and Engineering
University of South Florida
Tampa, Florida 33620
Email: ijvergara@mail.usf.edu

***Abstract*—Crowd sensing is an approach to collect many samples of a phenomena of interest by distributing the sampling across a large number of individuals. While any one individual may not provide sufficient samples, aggregating samples across many individuals may provide high-quality and high-coverage measurements of a phenomena. In this work, we propose an incentive assignment mechanism for crowd sensing variable phenomena (e.g., temperature) that balances the goal of maximizing coverage of the area of interest, while at the same time staying within a budget constraint. This algorithm not only takes into account the area covered by the participants' sensors, but also the spread of these sensors through a target area. This characteristic enables more representative sampling than existing methods, assuming the same budget. Compared to existing methods, this algorithm improves the spread of the set of acquired samples by more than 56 % percent, without sacrificing the number of samples purchased from human sensors.**

Keywords: Graph Set Cover; Incentive Mechanism for Crowd Sensing.

I. Introduction and Motivation

Smart phones are devices that, in addition to allowing us to communicate with each other, are becoming powerful computation tools. These devices have the potential to sense the environment around us with a fine level of temporal and spatial granularity, and quickly transmit sensed data back to the cloud for processing and sharing. Crowd sensing (CS) leverages these sensing capacities to collect many samples of a phenomena of interest from a large number of individuals. For CS to work well, samples must come from many individuals spread across an area of interest. Thus, a critical challenge in CS is motivating participation from individuals in under sampled areas.

Motivating participation takes many forms. Some people provide data because it makes them feel good to donate to a cause (altruism), while others provide information only in their self-interest. If the data collection application collects samples passively, CS participants do not need to be motivated at all. They only need to indicate they are willing to donate data once and the smart phone does the rest of the work.

Monetary rewards are also powerful motivators. Rewards can be static or dynamic. In the static case, the amount to pay to each participant is estimated in advance and does not change. In the dynamic case, the reward changes based on the reservation wage (i.e., the minimum amount of money a participant is willing to accept to do a task). A common approach is to combine monetary rewards with other types of incentives (e.g., intrinsic motives, social-based incentives, etc) in order to decrease the users reservation wage.

In this paper we combine the Reverse Auction Dynamic Price with Recruitment (RADP-VPC-RC) mechanism proposed by Lee and Hoh [6] with both the Greedy Set Cover and the Weighted Variance Maximization algorithms to create SPREAD, an incentive mechanism to obtain the lowest cost samples that are best distributed to cover the area of interest within a fixed budget.

We compare the spread (i.e., variance) and number of users of the sample set acquired by our system with both RADP-VPC-RC and the Greedy Incentive Algorithm (GIA) [3].

The rest of the paper is organized as follows. Section II includes a brief literature review. Section III describes the SPREAD algorithm, upon which we base our location-based incentive mechanism. Section IV presents a set of three experiments and results. Finally, Section V concludes the paper and provides brief directions for future research.

II. Related Work

Addressing the problem of area coverage in CS is a complex task for incentive mechanisms. Suppose that the variable of interest is temperature, and the goal is to estimate the temperature in a city. The logical decision would be to buy samples from users who are uniformly spread throughout the city. Unfortunately, two problems lead to poor coverage of an area. First, samples may have geographically unbalanced prices (i.e., cheaper samples in some regions and too expensive in others). Second, some regions may have a lack of participants, while an excess may be available in other regions. Both problems are evident in Figure 1. In the former case, the system buys just the cluster with the cheapest samples. In the latter case, only a subset of the area of the interest is well-sampled.

Some representative approaches to overcome these problems include the work of Reddy *et al.* [9] who propose the use of mobility profiles as part of the participant's selection criteria in the recruitment process. Using a similar approach, Falaki *et al.* [1] and Shilton *et al.* [10] propose increasing the participant's demographic diversity and social network affiliation. They propose to leverage the mobility patterns of different groups to increase the sensing coverage area. Kuznetsov and Paulos [5] propose the involvement of stakeholders such as students, parents, bicyclists, and homeless people. They say that each of these groups interacts with different public spaces and thus can provide more diverse coverage. Jaimes *et al.* [3] address the problem of price and user location imbalance by the combination of the Greedy Budgeted Maximum Coverage Algorithm (GBMCA) [4] and the Reverse Auction Dynamic Price(RADP-VPC) [6]. The result is a greedy algorithm that selects a representative subset of the users according to their location to maximize the area coverage while minimizing the cost. Finally, Mendez *et al.* [7] propose the use of density maps to capture the variability of the variable of interest in different regions as well as estimate the number of participants per region.

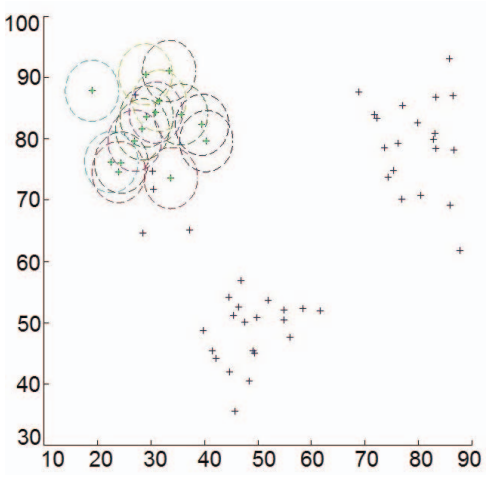

Fig. 1. Geographically imbalanced sample prices

A. The Reverse Auction Based Dynamic Price Scheme (RADP-VPC-RC)

Lee and Hoh [6] present the Reverse Auction Dynamic Price with Virtual Participation Credit and Recruitment (RADP-VPC-RC) an incentive mechanism inspired in micro-economic theory and motivated by the following scenario. There are *n* participants spread through the area of interest. They are willing to sell their sensing data to an auctioneer who in turn wants to acquire the *m* least expensive samples, within a time window (i.e,. rounds). The same process is repeated for *k* different rounds and is carried out through a reverse auction. The first component (RADP) is a recurrent reverse auction that dynamically reduces the reservation wage of each participant (i.e., the dynamic component). Once the auctioneer buys the *m* cheaper samples, the winners (i.e., those who managed to sell their samples) increase their bid's price by 10% expecting to increase their profits and losers decrease it by 20% in the hope of winning in the next round. In order to model the drop out, losers (i.e., those who did not manage to sell their samples) evaluate their return on investment (i.e., which is based on their reservation wage and sum of the cost that the user incurs in collecting the data) which is represented by equation 1. If it is below a certain threshold, then they withdraw from the system, otherwise they continue participating in the next round.

$$S_i^r = \frac{e_i^r + \beta_i}{p_i^r.t_i + \beta_i} \tag{1}$$

Here, S_i^r corresponds to the return of investment (ROI) of the participant i until round r. The terms e_i^r and β_i in the fraction numerator corresponds to the earned reward by users i until round r and to user tolerance period respectively. Finally, the product $p_i^r.t_i$ corresponds to the user's minimum expected reward. The first term of the product p_i^r indicates the number of participation instances of user i up to the current auction round r and second term t_i represents the user's reservation wage.

In order to increase the level of participation, users who lost in the previous round $r-1$ and continue participating are granted a Virtual Participation Credit (VPC). Each time the users lose, they accumulate their VPC by $v_i^r = v_i^{r-1} + \alpha$, where α corresponds to the amount of VPC granted in round r and v_i^r the amount earned up to the current round. Hence, losers increase their chances to win by using their virtual bid's price $b_i^{r*} = b_i^r - v_i^r$ rather than the real bid's price. Additionally, the RC mechanism allows the auctioneer to communicate the maximum price paid φ_k to a winner in the last round to the users who dropped out in previous rounds. This information allows them to re-evaluate their ROI and potentially return to the system in the next auction round. This expected ROI is evaluated as follows:

$$S_k^{r+1} = \frac{e_k^r + \varphi_k + \beta_k}{(p_k^r + 1).t_k + \beta_k} \tag{2}$$

This economic model (i.e., Lee and Hoh proposal) address two common problems of reverse auctions: Users dropping out and price's cost explosion. However, other problems such as coverage and data quality are not addressed. Coverage is somewhat addressed by the use of VPC and applying RADP-VPC-RC simultaneously in different regions and data quality is rarely considered. However, Jaimes *et al.* [3], Pham *et al.* [8], Sun *et al.* [11] address the problems of coverage and data quality respectively. In the coverage case, using a Budgeted Maximum Coverage approaches. In the data quality case, Pham *et al.* [8] uses a multi-objective Knapsack approach and Sun *et al.* [11] uses an energy management approach which involves quality-of-information (QoI) to assure data quality.

B. Greedy Incentive Algorithm (GIA)

Using reverse auctions as a general framework, Jaimes *et al.* [3] propose GIA, a greedy incentive algorithm that uses the Budgeted Maximum Coverage Problem (BMCP) to address the problem of area coverage for incentives in CS. Given a target area A, and a set of sensors $S = \{u_1, u_2, \ldots, u_n\}$ deployed in A, GIA uses a disk geometrical model to cover this area.

The algorithm assigns the elements covered by each $disk_i$ to each set S_i, associates the weight w_i of S_i as its cardinality, and its cost c_i as the cost of the sample provided by the sensor u_i located at the center of the $disk_i$. Let $G \subseteq S$ be a collection of sets, $w(G)$ and $c(G)$ denote the total number of elements covered by G and the total cost of the sets in G, respectively. Additionally, let W_i', $i = 1, \ldots, n$, denote the total number of elements covered by the set S_i but not covered by any set in G. Thus, the main idea is to choose the S_i that maximizes $\frac{W_i'}{c_i}$ in every iteration.

A GIA's main limitation is the direct relation between the sample set variance and budget. The algorithm prioritizes the number of acquired samples over the variance of the acquired sample set. GIA starts buying the cheapest sample and then uses the $\frac{W_i'}{c_i}$ criterion to continue the samples' acquisition process. This means that using the same budget GIA will purchase two cheap samples relatively close from the first acquired sample instead of buying the sample located farthest from the original one. On the other hand, SPREAD starts acquiring the sample that maximize the variance to the already acquired sample set. Thus, using SPREAD ensures obtaining a representative sample set from the target area since the very beginning even with budget constraints.

III. The SPREAD Algorithm

The main idea behind the SPREAD algorithm is to obtain the subset of samples that best represent the target area within a budget constraint. Our hypothesis is that a well spread, or spatially distributed subset of samples, is a better representation of the population than just an isolated cluster of samples. SPREAD can be used as a module of any incentive algorithm to acquire a set samples that maximize the area coverage at a minimum cost. In the context of this paper, SPREAD is used in combination with the Reverse Auction Dynamic Price with Virtual Participation Credit and Recruiting (RADP-VPC-RC) proposed by Lee and Hoh [6].

A. Geometric Coverage Model

In order to address the coverage problem, SPREAD uses the geometric disk model represented by Equation 3.

$$f(d(u_i, u_j)) = \begin{cases} 1 & \text{If } d(u_i, u_j) \leq R \\ 0 & \text{Otherwise} \end{cases} \tag{3}$$

Here, $d(u_i, u_j)$ is the Euclidean distance between sensors u_i and u_j and $R > 0$ is a parameter that is tuned according to the size of the target area. Indeed, this function defines a disk centered at sensor u_i with radius R. All sensors within such disk are considered the neighbors of u_i. Thereby, every sensor is drawn as a disk of radius R centered at its own location. Of course, each sensor is at least covered by the disk centered at the sensor itself and may have some neighbors.

B. The Algorithm

The algorithm for sample acquisitions works in two main stages. In the fist stage, the algorithm selects the set of samples that cover all the users at minimum cost. However, given the budget constraints, just a subset of that set can be acquired. In the second stage, a greedy algorithm buys the subset of samples that maximize the variance at minimum cost.

1) Creating the User Graph: A target area and a set of participants willing to sell their samples are represented by a rectangular grid and a set of location respectively. The following procedure describes the algorithm to obtain the candidates to be acquired in every round of the reverse auction.

Using the geometric model described before, a disk of radius R is drawn on every sample's location on the grid. Then, an edge from the disk center u_k to each u_j is drawn when $d(u_k, u_j) \leq R$, for $j = 1 \ldots n_{k-1}$, where n_k is the number of points within the disk.

The output of this procedure corresponds to the weighted graph $G = (V, E)$ as shown in Figure 2. Here, the set of vertices V represents the samples' locations, the vertex weights represent the sample's price, and the set of edges E represents the covering relation between the vertices. Thus, every vertex u_j connected by an edge to the vertex u_k is called the neighbor of u_k.

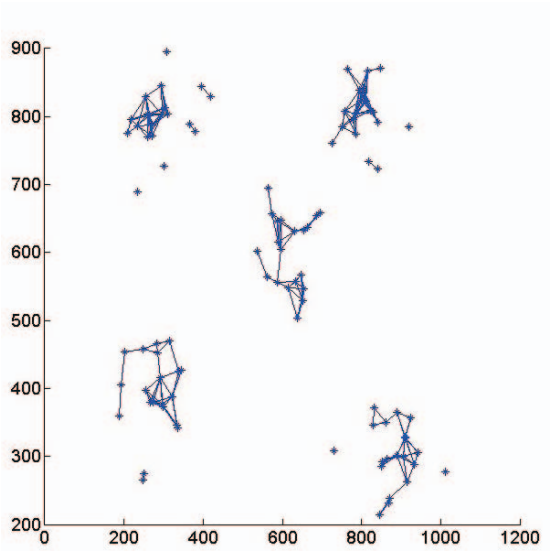

Fig. 2. Location Graph

2) Filtering the Candidates: The problem of covering all users at minimum cost can be reformulated in the following way: Given a graph $G = (V, E)$, and collection of sets $S_i = \{u_k\} \cup \{neighbors(u_k)\}$, $u_k \in V$ and $S_i \in V$, find a collection J of these sets S_i whose union equal V and such that $\sum_{i \in S} w_i$ is minimized.

We use Set Cover as modeling framework to solve the filtering problem. Since Set Cover is a NP-problem, we use Algorithm 1 which corresponds to a greedy version of Set Cover proposed by Gori to carry out a preliminary sample's selection. *et al.* [2]. The set of pre-selected samples in this stage corresponds to the set of vertices J located at the center of the disks as shown in Figure 3.

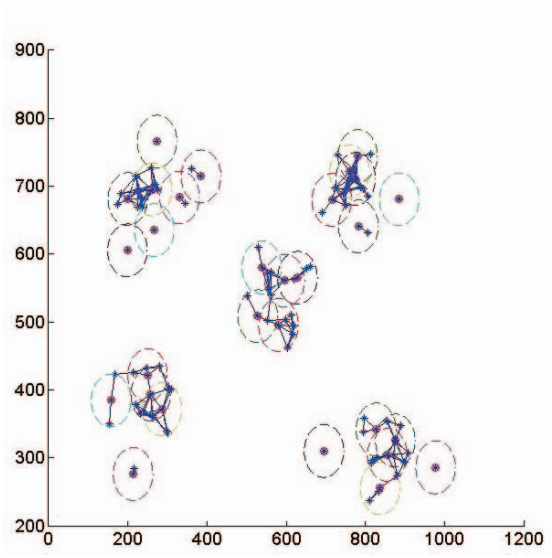

Fig. 3. Filtering the candidates using Algorithm 1

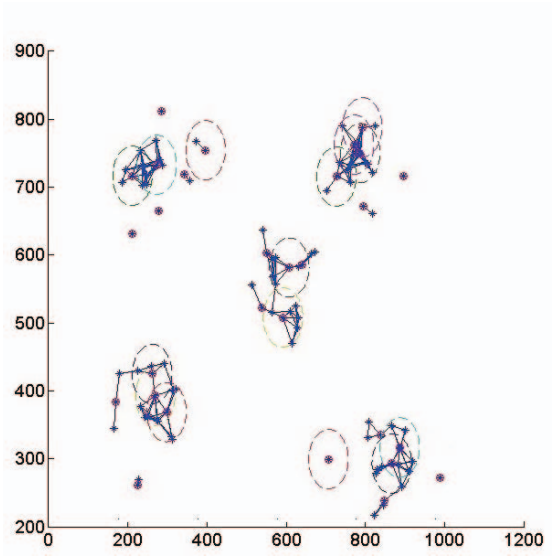

Fig. 4. Samples acquired using Algorithm 2

3) Acquiring the samples: The sample acquisition process is based on the internal distance between candidates. The main idea is to select the users that are better distributed over the area of interest. Therefore, we propose algorithm 2. The input for the algorithm 2 are the set of candidates V, selected at the previous section, and the available budget B. First, the algorithm selects the vertice u that minimizes the ratio between its *price* and *cardinality* in order to select the cheapest user with the largest number of neighbors. The node u is removed from V and it is included in the set of selected points V'. Next, the algorithm computes the distance between each remaining node v on V and the selected point w on V'; then the algorithm selects the node v that maximizes such distance times its cardinality $v_{cardinality}$. The selected node v is added to the set V' of selected points, and the price of v is discounted from the available budget. This process is repeated until the budget is consumed. Finally, the output of algorithm 2 is the set of selected points V'.

Algorithm 1: Filter the Candidates based on Greedy Set Covering

input : V Family of sets $S_1, \ldots, S_n (V := \bigcup_{k=1}^{n} S_i)$
output: $J \subseteq \{1, \ldots, n\}$, s.t $\bigcup_{j \in J} S_j = V$

begin
  $U \leftarrow V$
  $J \leftarrow \emptyset$
  while $U \neq \emptyset$ **do**
    select $i' \in \{1, \ldots, n\}, s.t. |S'_i \cap U|$ *is maximum*
    $U \leftarrow U \setminus S'_i$
    $J \leftarrow J \cup \{i'\}$
  end
  return J
end

Algorithm 2: Acquire samples based on maximization of the distance between points

input : V a collection of vertices with users' location and price
input : B available budget
output: $V' \subseteq V$, selected points

selectVertices(V,B)
begin
  $V' \leftarrow \emptyset$
  select $u \in V$ that minimize $\frac{u_{price}}{u_{cardinality}}$
  $V \leftarrow V \setminus u$
  while $Stop \neq 1 \wedge V \neq \emptyset$ **do**
    $V' \leftarrow V' \cup u$
    $B \leftarrow B - u_{price}$
    $Stop \leftarrow 1$
    $sDistance \leftarrow 0$
    for $v \in V$ **do**
      for $w \in V'$ **do**
        $nDistance \leftarrow nDistance + dist(v, w)$
      end
      if
      $nDistance * v_{cardinality} > sDistance \wedge v_{price} \leq B$
      then
        $u \leftarrow v$
        $Stop \leftarrow 0$
        $sDistance \leftarrow nDistance * v_{cardinality}$
      end
    end
    $V \leftarrow V \setminus u$
  end
  return V'
end

IV. Experiments and results

A. Experimental setup

In order to validate the proposed approach, we used a generic area of interest of 1000×1000 points. The number of deployed users is 100 whom are randomly distributed using a normal distribution with 5 clusters according to the parameters on table I. In addition, the *true valuation* parameter is generated following a normal distribution for each cluster according to the line *True valuation* on table I. Finally, in order to obtain statistical significance, the experiments are replicated 15 times.

B. Experiment 1

In this experiment we compare the variance of the samples acquired by RADP-VPC-RC and SPREAD under different budgets. Here, we increase the budget from 60 to 300 with step size of 20. As shown in Figure 5 for every budget value SPREAD acquires a set of samples with a variance of 23 times greater on average than the variance of the sample set acquired by RADP-VPC-RC. Thereby, even with a limited budget SPREAD is able to obtain a representative samples set of the target area.

Figure 6 shows a comparison between the number of samples acquired by RADP-VPC-RC and SPREAD. On average, RADP-VPC-RC acquires twice as many samples as SPREAD. However, the samples acquired by RADP-VPC-RC are physically located very close to each other, rendering pretty much the same information.

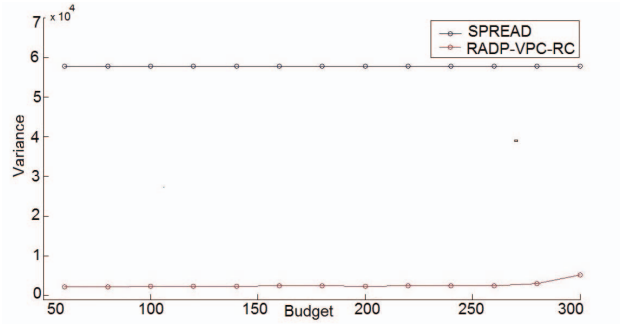

Fig. 5. Experiment 1, SPREAD vs RADP-VPC-RC, sample set variance under different budget

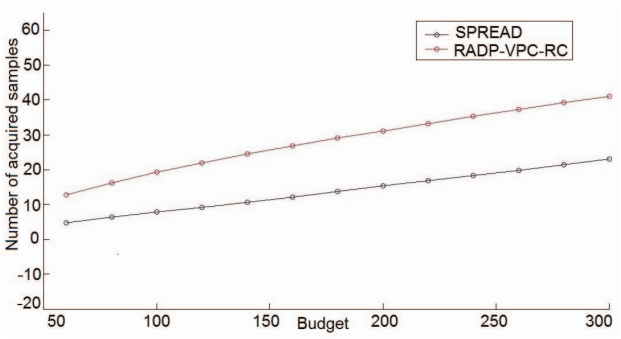

Fig. 6. Experiment 1, SPREAD vs RADP-VPC-RC, samples acquired under different budget

C. Experiment 2

The goal of this experiment is to test how well SPREAD scales in terms of the variance of sample sets as well as the number of acquired samples under different budget values. For this experiment, we increase the budget from 60 to 300 with step size of 20. Figure 7 shows that in the case of SPREAD, the sets of samples acquired by all the different budget values exhibited a high and constant variance. Besides, as shown in Figure 8 as the budget grows, the number of acquired samples also grows linearly. These two characteristics allow to SPREAD the acquisition of a representative sample set that reflect more realistically the nature of the variable of interest.

D. Experiment 3

Finally, we compare the performance of SPREAD vs GIA in terms of the variance of the acquired sample set as well as the number of acquired samples per budget. Figures 9 and Figure 10 show these results respectively. In the first case, the SPREAD algorithm acquired a sample set with a variance of 56% on average greater than the variance of the set acquired by GIA. However, GIA acquired on average 9% more samples than SPREAD.

V. Conclusions and Future Research

This work presents SPREAD, an incentive algorithm in the context of crowd sensing. This algorithm works in two stages, in the first stage, SPREAD creates a graph using the participant location. Then,

TABLE I
PARAMETERS OF DEPLOYED USERS

Cluster	1	2	3	4	5
Location μ	(300, 800)	(300, 400)	(600, 600)	(800, 800)	(900, 300)
Location $Covariance$	$\begin{bmatrix} 2500 & 0 \\ 0 & 2600 \end{bmatrix}$	$\begin{bmatrix} 2500 & 0 \\ 0 & 2600 \end{bmatrix}$	$\begin{bmatrix} 2600 & 0 \\ 0 & 2600 \end{bmatrix}$	$\begin{bmatrix} 2500 & 0 \\ 0 & 2600 \end{bmatrix}$	$\begin{bmatrix} 2600 & 0 \\ 0 & 2600 \end{bmatrix}$
True valuation	$\mu = 5, \sigma = 1$	$\mu = 9, \sigma = 1$	$\mu = 13, \sigma = 1$	$\mu = 17, \sigma = 1$	$\mu = 20, \sigma = 1$

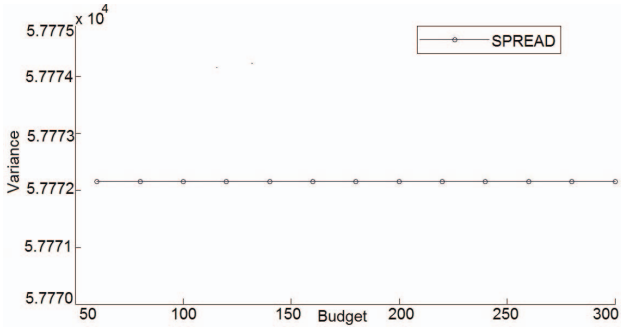

Fig. 7. Experiment 2, SPREAD, sample set variance under different budget

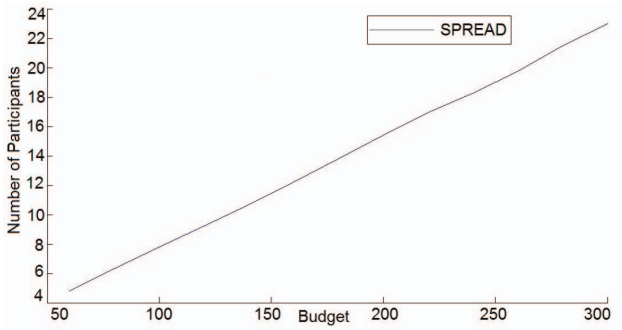

Fig. 8. Experiment 2, SPREAD, samples acquired under different budget

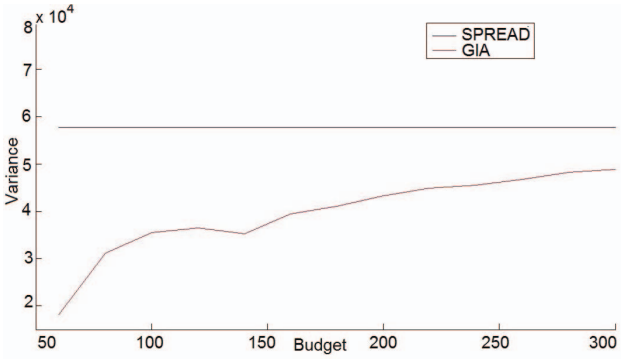

Fig. 9. Experiment 3, SPREAD vs GIA, variance of sample sets acquired by using different budget values

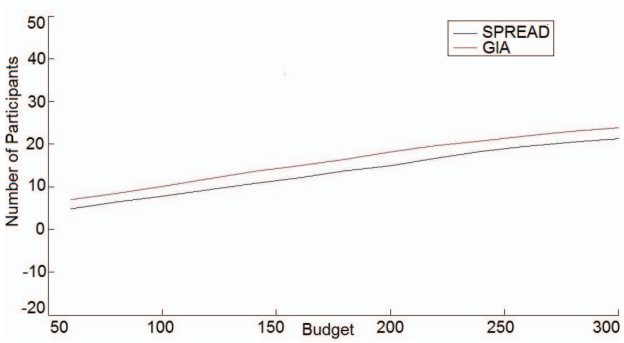

Fig. 10. Experiment 3, SPREAD vs GIA, number of samples acquired by using different budget values

a greedy version of Set Cover algorithm is utilized to select a set of potential candidates to be acquired. Finally, the Algorithm 2 acquires a sample set in each round. SPREAD acquires on average a sample set with a variance 23 times greater than the sample set acquired by RADP-VPC-RC and 56% greater than the sample set acquired by GIA. Besides, SPREAD is able to keep a high level of variance even under low budgets, and scale linearly on the number of samples acquired under budget increments.

Future work is under way to model a flexible scheme which include not only the rules of reverse auctions, but also scenarios such the admission of new participants in the middle of a set of rounds. We also want to investigate mobility models to represent the natural movement of the participants. Additionally, we would like to investigate the application of Reinforcement Learning as well as Game Theory as possible modeling frameworks for incentives mechanisms for CS. Finally, we would like to implement the proposed system in a mobile environment to explore the system's usability as well as its performance in real word applications.

REFERENCES

[1] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin. Diversity in smartphone usage. In *MobiSys*, pages 179–194, 2010.

[2] Fabio Gori, Gianluigi Folino, Mike S. M. Jetten, and Elena Marchiori. Mtr: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. *Bioinformatics*, 27(2):196–203, 2011.

[3] Luis G. Jaimes, Idalides J. Vergara-Laurens, and Miguel A. Labrador. A location-based incentive mechanism for participatory sensing systems with budget constraints. In *PerCom*, pages 103–108, 2012.

[4] Samir Khuller, Anna Moss, and Joseph Naor. The budgeted maximum coverage problem. *Inf. Process. Lett.*, 70(1):39–45, 1999.

[5] Stacey Kuznetsov and Eric Paulos. Participatory sensing in public spaces: activating urban surfaces with sensor probes. In *Conference on Designing Interactive Systems*, pages 21–30, 2010.

[6] Juong-Sik Lee and Baik Hoh. Dynamic pricing incentive for participatory sensing. *Pervasive and Mobile Computing*, 6(6):693–708, 2010.

[7] Diego Mendez and Miguel A. Labrador. Density maps: Determining where to sample in participatory sensing systems. In *MUSIC*, pages 35–40, 2012.

[8] Hong Nhat Pham, Back Sun Sim, and Hee Yong Youn. A novel approach for selecting the participants to collect data in participatory sensing. In *SAINT*, pages 50–55, 2011.

[9] Sasank Reddy, Deborah Estrin, and Mani B. Srivastava. Recruitment framework for participatory sensing data collections. In *Pervasive*, pages 138–155, 2010.

[10] Katie Shilton, Nithya Ramanathan, Sasank Reddy, Vidyut Samanta, Jeff Burke, Deborah Estrin, Mark H. Hansen, and Mani B. Srivastava. Participatory design of sensing networks: strengths and challenges. In *PDC*, pages 282–285, 2008.

[11] Zhanwei Sun, Chi Harold Liu, Chatschik Bisdikian, Joel W. Branch, and Bo Yang. Qoi-aware energy management in internet-of-things sensory environments. In *SECON*, pages 19–27, 2012.