Resource-Aware Approaches for Truth Analysis in Crowdsourcing

Xiaomei Zhang, Yibo Wu and Guohong Cao
Department of Computer Science and Engineering
The Pennsylvania State University, University Park, PA, 16802
Email: {xqz5057,yxw185,gcao}@cse.psu.edu

***Abstract*—Although crowdsourcing can provide a large amount of information through mobile devices and mobile users, the information provided by them may be inaccurate. Various truth analysis techniques have been proposed to identify truth from the noisy data either in a heuristic manner or using statistical models. However, if the available data are limited or have large conflicts, it is difficult to identify the truth or ensure the data credibility (quality). In this paper, we address this problem by utilizing the communication networks to adaptively collect data from mobile users, especially when the existing data are not enough to ensure data credibility. Considering the requirement on data credibility and the constraint of network resources, we quantify the tradeoff between the enhanced data credibility and the increased network overhead, and propose resource-aware approaches for truth analysis. Specifically, we formalize two problems in resource-constrained mobile opportunistic networks: *max-credibility* which aims to maximize data credibility with some network overhead, and *min-overhead* which aims to achieve a specified data credibility while minimizing the network overhead. Simulation and experimental results demonstrate the effectiveness of the proposed solutions in terms of data credibility and network overhead.**

I. Introduction

Crowdsourcing [1][2][3] has received considerable attention in recent years due to the unprecedented popularity of mobile devices and the rapid growth of wireless communication technology. Based on the sensing capabilities provided by mobile devices, crowdsourcing relies on individual volunteers to collect data about themselves and their surroundings. For example, through crowdsourcing, real-time traffic conditions can be monitored and shared with other drivers [4]. In disaster recovery, through crowdsourcing, the building/road damage condition and human injury/loss can be identified, which can be used to help first responders plan their recovery paths and resources.

Although crowdsourcing can provide a large amount of information through mobile devices and mobile users, the reliability of these information sources is usually unknown a priori and the information provided by them may be inaccurate. An important problem in crowdsourcing is to assess the reliability of the mobile users and identify the truth from the reported data. To address this problem, various truth analysis techniques have been proposed by researchers. For example, [5][6][7] designed techniques based on a basic heuristic, i.e., iteratively estimating the truth of the information by analyzing the reliability of the information source and estimating the reliability of information sources based on the correctness of their provided information. Some other researchers also investigated statistical models for truth finding, such as bayesian inference [8] and expectation-maximization (EM) [9][10].

The aforementioned truth analysis techniques only depend on the currently available or observed data. However, if the available data are limited or have large conflicts, it is difficult to identify the truth or ensure the data credibility (quality). In this paper, we address this problem by utilizing the communication networks to adaptively collect data from mobile users, especially when the existing data are not enough to ensure data credibility. Specifically, based on a feedback loop between information and communication networks, we are able to utilize the communication networks to collect the right information from mobile users and make sure enough data are collected to improve the data credibility.

Although data credibility can be improved by utilizing communication networks to collect extra data, it also increases the communication overhead. The communication overhead for data collection can be significant in some cases. For example, in mobile opportunistic networks (also known as delay tolerant networks [11][12]) which are commonly used in battlefield or disaster-recovery, the network resources are limited due to the dynamics of network topology and the lack of end-to-end routing path.

Considering the requirement on data credibility and the constraint of network resources, we quantify the tradeoff between the enhanced data credibility and the increased network overhead, and propose resource-aware approaches for truth analysis. Specifically, we formalize two problems in resource-constrained mobile opportunistic networks: the *max-credibility* problem which aims to maximize data credibility with some network overhead, and the *min-overhead* problem which aims to achieve a specified data credibility while minimizing the network overhead. Overall, the contributions of the paper are summarized as follows:

- Different from existing truth analysis approaches, we utilize the communication network to adaptively collect data from some users, so that high-quality data are collected to achieve the required data credibility.
- We formalize the max-credibility problem and propose a solution to select users based on their data collection

This work was supported in part by Network Science CTA under grant W911NF-09-2-0053 and by the National Science Foundation (NSF) under grant CNS-1421578.

capabilities such as network reachability and user reliability. A maximum likelihood estimation (MLE)-based technique is also proposed to identify truth from the collected data.
- We formalize the min-overhead problem and propose a solution which iteratively selects users for data collection until the required data credibility is achieved.
- The proposed solutions are evaluated based on trace-driven simulations and a testbed consisting of 20 Bluetooth-enabled smartphones. Both simulation and experimental results demonstrate the effectiveness of the proposed solutions in terms of data credibility and network overhead.

The rest of the paper is organized as follows. In Section II, we present the problem formalization and the network model. Section III presents the max-credibility problem and our proposed solutions. Section IV presents the min-overhead problem and our proposed solutions. In Section V, we evaluate the performance of the proposed solutions. Section VI reviews related work, and Section VII concludes the paper.

II. Preliminaries

In this section, we first formulate the problems addressed in the paper, and then describe the network model and user reliability.

A. Problem Formulation

The system model includes a server and a set of mobile users who communicate with each other through a mobile opportunistic network, as shown in Figure 1. Similar to [13], we assume the server knows the mobile users existing in the network. In this paper, we focus on the sensing tasks requiring real-valued measurements. First, the server creates sensing tasks specifying the data required and the time constraint, and sends each sensing task to a set of selected mobile users through the mobile opportunistic network. Then, the queried users collect data as specified by the sensing task if they have the sensing capability, and send the collected data back to the server. Note that the time constraint is at a much larger time scale, and it does not mean real time. It is used to reduce the unnecessary traffic due to the large delay variation in mobile opportunistic networks.

Since the reliability of mobile users is usually unknown a prior, the data provided by them may be inaccurate and the data credibility cannot be ensured. To ensure the data credibility, the server usually queries multiple users to collect data, and then estimates the true value based on truth analysis techniques. However, in mobile opportunistic networks, querying more nodes also induces more network overhead. This problem becomes worse considering that mobile users may have to send back extra information to improve the *data provenance*; e.g., voice or video related to the collected data or the context (temporal and spacial information) related to this data collection. This will significantly increase the data size and hence increase the network overhead. Considering the requirements on data credibility and the constraint of

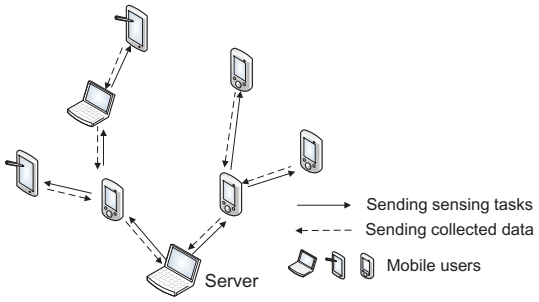


Fig. 1: Sending tasks and collecting data through mobile opportunistic network.

network resources in mobile opportunistic networks, our main objective is to quantify the tradeoff between the improved data credibility and the increased network overhead, and propose resource-aware approaches to improve data credibility.

To formulate the problem, we first explain how to quantify *data credibility* and *network overhead*.

Definition* 1. *Data credibility *measures the quality of the data for a sensing task. We use the estimation accuracy to represent the data credibility. A smaller estimation error $|\mu - \hat{\mu}|$ stands for a higher data credibility, where μ is the ground truth and $\hat{\mu}$ is the estimated value from the collected data.*

In mobile opportunistic networks, due to the sparse node connectivity, the performance of data collection is limited by the transmission opportunities which only happen when nodes move into the communication distance. A large data item may have to be cut into small packets to make sure that they can be transmitted during a short contact time. Therefore, we use the total number of transmissions (i.e., transmission hops) for all these data packets to quantify the network overhead.

***Definition* 2.** *The **network overhead** of a sensing task is quantified as the total number of transmission hops used for collecting data from mobile users. If the number of transmission hops for collecting data from user i is h_i, the network overhead is represented as $\sum_{i \in V_s} h_i$, where V_s is the set of queried users.*

In this definition, the transmission hops for collecting data from a user include the round-trip transmission hops for transmitting the task to the user and sending back the collected data to the server.

To quantify the tradeoff between data credibility and network overhead, we formulate two problems: the ***max-credibility*** problem and the ***min-overhead*** problem.

Problem* 1. *Max-credibility: *finding the set of mobile users to query so that the data credibility is maximized while the network overhead is below a specific value.*

The max-credibility problem can be formalized as follows:

$$\min \quad |\mu - \hat{\mu}| \tag{1}$$

$$s.t. \quad \sum_{i \in V} s_i \cdot h_i \leq \beta \tag{2}$$

$$s_i \in \{0, 1\}, \quad \forall i \in V \tag{3}$$

where $s_i \in \{0, 1\}$ denotes whether i is selected for collecting data, V denotes the set of mobile users in the system, and β denotes a specific value for network overhead. Maximizing data credibility equals to minimizing estimation error as depicted in Formula (1). Formula (2) ensures that the network overhead is below β.

Problem 2. Min-overhead*: using minimum network overhead to collect data so that a specified data credibility is achieved.*

The min-overhead problem can be formalized as follows:

$$\min \quad \sum_{i \in V} s_i \cdot h_i \tag{4}$$

$$s.t. \quad |\mu - \hat{\mu}| \leq \epsilon \tag{5}$$

$$s_i \in \{0, 1\}, \quad \forall i \in V \tag{6}$$

The objective of the optimization is to minimize the network overhead (Formula (4)), while achieving the specified data credibility, i.e., limiting the estimation error to be smaller than ϵ as shown in Formula (5).

In this paper, the collected data for sensing tasks are assumed to be numerical values. The magnitude of the data may vary tremendously for different sensing tasks. To ensure different sensing tasks can be processed using the same statistical model, the data value for a sensing task is normalized by a *base number* σ, which is a constant and relevant to the data magnitutue specific to each sensing task. It can be learned using MLE as presented in the next section.

With data normalization, the estimation error $|\mu - \hat{\mu}|$ in the two problems should be modified to the normalized estimation error $\frac{|\mu - \hat{\mu}|}{\sigma}$. This does not affect the objective (Formula (1)) in the max-credibility problem, since minimizing the estimation error is equal to minimizing the normalized estimation error. For the min-overhead problem, the constraint of limiting the estimation error by ϵ (Formula (5)) becomes limiting the normalized error by ϵ: $\frac{|\mu - \hat{\mu}|}{\sigma} \leq \epsilon$

B. Models

1) Network Model: The network studied in this paper is a mobile opportunistic network $G(V, E)$, where V is the set of nodes (i.e., mobile users) and E is the set of edges. The edge $e_{u,v}$ in E represents the pairwise contact process between nodes u and v ($u, v \in V$). The inter-contact time between nodes u and v has been experimentally validated in [14] to follow an exponential distribution with rate parameter λ_{uv}. The contact process between u and v follows a homogeneous Poisson process with rate λ_{uv}.

2) Reliability of Users: The mobile users in the system have different *reliability*. The reliability determines the quality of the data provided by the user. Here, we denote the reliability of a user i as r_i ($0 < r_i \leq 1$). A reliability of 1 means that the user is totally reliable and likely to provide high-quality data. A user with lower reliability is likely to provide data deviating more from the ground truth. Similar to [8], we assume that the random observation of user i for task j follows normal distribution $N(\mu_j, (\sigma_j / r_i)^2)$, where μ_j is the ground truth of task j and σ_j / r_i is the standard deviation (σ_j is the base number of task j). The assumption of normal distribution has been experimentally validated in [15]. The distribution has a smaller variance as the reliability r_i is higher. Even if $r_i = 1$, the random observation of user i still has variance σ_j^2, which is caused by the unavoidable random errors in real-world observations.

III. Max-Credibility with Network Resource Constraint

In this section, we first study the problem of maximizing data credibility with limited network resource. The objective of the max-credibility problem is to minimize the estimation error as depicted in Formula (1). However, it is impossible to directly calculate the estimation error since the ground truth of the task is unavailable to server. Therefore, we propose a heuristic-based max-credibility approach to address the problem. In the approach, the server first selects users that are more likely to provide accurate data with limited network overhead. After data are collected from these users, an MLE-based technique is further utilized to identify the truth from these data.

A. User Selection

To ensure the selected users can provide data that achieve high data credibility, our max-credibility approach prioritizes the users that are more likely to provide accurate data with limited network overhead. Specifically, users are selected based on their data collection capabilities such as *network reachability* and *user reliability*.

1) Network reachability: The network reachability measures if the data can be successfully transmitted within a time constraint. The transmission includes the round-trip transmission between the server (denoted as s) and user i. To transmit data between two nodes in the network, we simply choose the path with the minimum expected delay. If the time constraint for task j is $\mathcal{T}_j$, we try to ensure the one-way transmission delay between the server and the user is smaller than $\mathcal{T}_j/2$. Based on the assumption that the pairwise contact processes follows Poisson processes which are mutually independent, the delay on the transmission path in a mobile opportunistic network satisfies hypo-exponential distribution [14][12]. The probability that the data can be successfully transmitted from the server to user i within $\mathcal{T}_j/2$ is:

$$P(D_{s \to i} < \mathcal{T}_j/2) = \sum_{i=1}^{l} C_i^{(l)} \cdot (1 - e^{-\lambda_i \mathcal{T}_j/2}) \tag{7}$$

where the coefficient $C_i^{(l)} = \prod_{k=1, k \neq i}^{l} \frac{\lambda_k}{\lambda_k - \lambda_i}$. In this formula, l is the number of hops on the transmission path, and the rate parameter on each hop is λ_i ($1 \leq i \leq l$). Since the transmission delay from user i back to server s has the same hypo-exponential distribution as the transmission delay from s to i, we have

$$P(D_{i \to s} < \mathcal{T}_j/2) = P(D_{s \to i} < \mathcal{T}_j/2) \tag{8}$$

Then, we simply compute the probability that the data can be successfully transmitted before $\mathcal{T}_j$ as:

$$p_{i,j}^n = P(D_{s \to i} < \mathcal{T}_j/2) \cdot P(D_{i \to s} < \mathcal{T}_j/2) \tag{9}$$

2) Reliability: If user i has the capability to collect data, the reliability of user i determines the accuracy of the data it collects. The collected data value is accurate if the normalized error from the ground truth is smaller than ϵ, i.e., $\frac{|x_{ij}-\mu_j|}{\sigma_j} < \epsilon$, where x_{ij} is the data collected by i. Based on the assumption that the random observation of user i has normal distribution $N(\mu_j, (\sigma_j/r_i)^2)$, where σ_j is the base number for task j and r_i is the reliability for user i, the probability that the collected data value is accurate is:

$$\begin{aligned} p^r_{i,j} &= P(\frac{|x_{ij}-\mu_j|}{\sigma_j} < \epsilon) \\ &= \int_{\mu_j-\epsilon\sigma_j}^{\mu_j+\epsilon\sigma_j} \frac{1}{\sigma_j/r_i\sqrt{2\pi}} e^{-\frac{(x_{ij}-\mu_j)^2}{2\sigma_j^2/r_i^2}} dx_{ij} \quad (10) \\ &= \Phi(\epsilon r_i) - \Phi(-\epsilon r_i) \quad (11) \end{aligned}$$

where $\Phi(*)$ is the cumulative distribution function (CDF) of the standard normal distribution. Even though the unknown ground truth μ_j and base number σ_j are used in the deviation, the result $\Phi(\epsilon r_i) - \Phi(-\epsilon r_i)$ is not influenced by μ_j and σ_j.

Then, the server integrates the two properties and computes the overall probability for user i to provide accurate data before task expires, i.e., $p_{i,j} = p^n_{i,j} \cdot p^r_{i,j}$

The probability $p_{i,j}$ is also referred to as the *success probability* for user i.

In the beginning, the server usually does not know users' reliability in priori. In this case, it can simply assume users are reliable and use $r_i = 1$ when computing $p^r_{i,j}$. As more data have been collected from users, the server may have some knowledge about users' reliability based on our MLE-based estimation. In this case, the inclusion of $p^r_{i,j}$ in $p_{i,j}$ can assist the selection of more reliable users.

Afterwards, the server selects the users to query based on the following optimization problem:

$$\max \quad 1 - \prod_{i \in V} (1 - p_{i,j})^{s_i} \quad (12)$$

$$s.t. \quad \sum_{i \in V} s_i \cdot h_i \leq \beta \quad (13)$$

$$s_i \in \{0, 1\}, \quad \forall i \in V \quad (14)$$

where h_i denotes the network overhead for collecting data from user i, and s_i denotes whether i is selected. The optimization problem maximizes the probability that at least one selected user can provide accurate data (12), with the constraint that the total network overhead is less than β (13). By solving the optimization problem, the server can prioritize the nodes that not only have a higher probability to provide accurate data but also consume less network overhead. This optimization problem can be solved using dynamic programming similar to a 0-1 knapsack problem in pseudo-polynomial time.

B. Finding Truth using MLE

After data are collected from the selected users, the truth can be identified with an MLE-based truth analysis technique. Specifically, the MLE-based truth analysis can be used to learn the unknown parameters like ground truth and the base number of each task, and the reliability of each user.

1) Background for MLE: The maximum-likelihood estimation (MLE) is commonly used to estimate the unknown parameters of a statistical model. In general, given a set of observed data and the underlying statistical model, MLE selects the set of values for the model parameters that maximize the likelihood function, which is the probability that the data are observed under the resulting distribution. Given the set of observed data $X = \{x_1, x_2, ..., x_n\}$ and the set of parameters Θ, the likelihood function is represented as

$$L(\Theta; X) = f(x_1, x_2, ..., x_n|\Theta) = \prod_{i=1}^{n} f(x_i|\Theta) \quad (15)$$

To maximize the likelihood function, the log-likelihood function is used to simplify the process of derivation. Then, the MLE for Θ is computed as

$$\hat{\Theta} = \text{argmax} \log L(\Theta; X) = \text{argmax} \sum_{i=1}^{n} \log f(x_i|\Theta) \quad (16)$$

2) Mathematical Formulation and Derivation: To utilize MLE in our problem, we first set up the statistical model. In our statistical model, the observed data X is the set of data collected from the selected users. Assuming the server has collected data for m tasks, we have $X = \{X_1, X_2, ...X_m\}$, where X_j is the set of data values observed for task j. The unknown parameters in the model include the ground truth μ_j, the base number σ_j for each task j, and the reliability r_i of each user i, i.e., $\Theta = \{\mu_1, \mu_2, ..., \mu_m, \sigma_1, \sigma_2, ..., \sigma_m, r_1, r_2, ..., r_n\}$.

Here, the ground truth of the each task is treated as unknown parameters in the statistical model, so that we can estimate the true value by estimating the set of unknown parameters Θ using MLE.

Let $d_{ij} = \{0, 1\}$ denote if user i has provided data for task j, and x_{ij} denote the data user i has provided for task j. The *probability density function (pdf)* that x_{ij} is observed is:

$$f(x_{ij}|\Theta) = \frac{1}{\sigma_j/r_i\sqrt{2\pi}} e^{-\frac{(x_{ij}-\mu_j)^2}{2\sigma_j^2/r_i^2}} \quad (17)$$

From Equation (17), the *pdf* that the data in X_j are observed for task j is calculated as:

$$f(X_j|\Theta) = \prod_{i=1}^{n} (f(x_{ij}|\Theta))^{d_{ij}} = \prod_{i=1}^{n} (\frac{1}{\sigma_j/r_i\sqrt{2\pi}} e^{-\frac{(x_{ij}-\mu_j)^2}{2\sigma_j^2/r_i^2}})^{d_{ij}} \quad (18)$$

Here, x_{ij} is only valid as $d_{ij} = 1$, i.e., when user i has observation x_{ij} for task j. As $d_{ij} = 0$, user i has no observation, so x_{ij} is meaningless. We can simply let $x_{ij} = 0$ as $d_{ij} = 0$ to make (18) a legal equation.

Then, the *pdf* that $X = \{X_1, X_2, ...X_m\}$ is observed can be computed based on Equation (18):

$$f(X|\Theta) = \prod_{j=1}^{m} f(X_j|\Theta) = \prod_{j=1}^{m} \prod_{i=1}^{n} (\frac{1}{\sigma_j/r_i\sqrt{2\pi}} e^{-\frac{(x_{ij}-\mu_j)^2}{2\sigma_j^2/r_i^2}})^{d_{ij}} \quad (19)$$

From Equation (19), the log-likelihood function $\log L(\Theta; X)$ is given by:

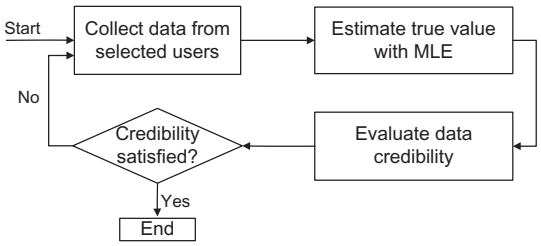


Fig. 2: An overview of the iterative approach

$$\begin{aligned}\log L(\Theta; X) = &\quad \log f(X|\Theta) \\ = &\quad \sum_{i=1}^{m}\sum_{i=1}^{n} d_{ij} \cdot [\log(\frac{1}{\sigma_j/r_i\sqrt{2\pi}}) - \frac{(x_{ij}-\mu_j)^2}{2\sigma_j^2/r_i^2}]\end{aligned} \tag{20}$$

To get $\hat{\Theta}$ that maximize $\log L(\Theta; X)$, we set the derivatives to be zero, i.e., $\frac{\partial \log L}{\partial \mu_j} = 0$, $\frac{\partial \log L}{\partial \sigma_j} = 0$, and $\frac{\partial \log L}{\partial r_i} = 0$, which yields,

$$\mu_j = \frac{\sum_{i=1}^{n} d_{ij} r_i^2 x_{ij}}{\sum_{i=1}^{n} d_{ij} r_i^2} \tag{21}$$

$$\sigma_j = (\frac{\sum_{i=1}^{n} d_{ij}(x_{ij}-\mu_j)^2 r_i^2}{\sum_{i=1}^{n} d_{ij}})^{\frac{1}{2}} \tag{22}$$

$$r_i = (\frac{\sum_{j=1}^{m} d_{ij}}{\sum_{j=1}^{m} d_{ij}(x_{ij}-\mu_j)^2/\sigma_j^2})^{\frac{1}{2}} \tag{23}$$

It is impossible to directly solve these equations to obtain the MLEs $\hat{\mu_j}$, $\hat{\sigma_j}$ and $\hat{r_i}$. Instead, we approach it in an iterative manner: given the values $\mu_j^{(k)}$, $\sigma_j^{(k)}$ and $r_i^{(k)}$, we use the equations (21)-(23) to obtain the values $\mu_j^{(k+1)}$, $\sigma_j^{(k+1)}$ and $r_i^{(k+1)}$; the iterative process continues until the results converge. The initial values of the iterative process is set as:

$$\mu_j^{(0)} = \frac{\sum_{i=1}^{n} x_{ij} d_{ij}}{\sum_{i=1}^{n} d_{ij}},\ \sigma_j^{(0)} = \frac{\sum_{i=1}^{n} |x_{ij} - \mu_j^{(0)}| d_{ij}}{\sum_{i=1}^{n} d_{ij}},\ r_i^{(0)} = 1$$

IV. Achieving Specified Credibility with Min-Overhead

In this section, we study another important problem which aims to achieve a specified data credibility with minimum network overhead. User selection in this problem is more challenging since it is difficult to evaluate whether the data credibility can be satisfied with the selected users. For example, if only a small group of users are selected to keep the overhead in a low level, the required data credibility may not be guaranteed; if many users are selected to ensure data credibility, a large amount of network overhead may be incurred.

To address this problem, we propose an iterative user selection approach: in each iteration, only a limited group of users are selected, and the data credibility are evaluated based on all collected data; the iterative process continues selecting users until the required data credibility is achieved.

In this section, we first have an overview of the iterative approach, and then discuss in detail how to select users and evaluate credibility in each iteration.

A. An Overview of the Iterative Approach

Figure 2 shows an overview of the iterative approach. In each iteration, the server first selects a group of users from the unselected users to collect data. Then, the true value is estimated based on all the collected data values including those collected from previous iterations. Afterwards, the server evaluates the data credibility and checks if the required data credibility is satisfied. If so, the iterative process ends. Otherwise, the server starts a new iteration and selects another group of users.

In this iterative approach, we propose solutions to select the users to query in each iteration and estimate the true value from the collected data, and find techniques to evaluate the data credibility.

B. Selecting Users and Estimating Truth

To minimize the network overhead, we limit the network overhead in each iteration when selecting users. Specifically, the network overhead in each iteration is limited by β^o, which is a pre-defined parameter and can be set flexibly. With the limitation on network overhead, the problem becomes selecting a group of users that are more likely to provide high-quality data, which is equivalent to the max-credibility problem studied in Section III.

When collecting data from users, there is a time constraint $\mathcal{T}_j$ for each sensing task j. Here, we set the time constraint for collecting data in one iteration to be τ^o with $\tau^o < \mathcal{T}_j$. The parameter τ^o is also pre-defined and can be set flexibly. As τ^o is smaller, more iterations may be performed within $\mathcal{T}_j$.

To estimate the truth from the collected data, the MLE-based truth analysis approach presented in Section III-B is applied here. The observed data applied to MLE include all data values collected in the current and previous iterations.

C. Evaluating Data Credibility

Since the ground truth is unknown, it is impossible to directly calculate data credibility from the collected data. Instead, we calculate data credibility in a probabilistic manner, and ensure that the specified data credibility reaches a confidence level.

For each task j, instead of achieving a specified data credibility by constraining the normalized error by ϵ, $\frac{|\mu_j - \hat{\mu}_j|}{\sigma_j} \leq \epsilon$, we ensure the normalized error to be smaller than ϵ with confidence $1-\alpha$, i.e., $P(\frac{|\mu_j - \hat{\mu}_j|}{\sigma_j} \leq \epsilon) > 1-\alpha$, which is equal to:

$$P(\mu_j \in [\hat{\mu}_j - \epsilon\sigma_j, \hat{\mu}_j + \epsilon\sigma_j]) > 1-\alpha \tag{24}$$

In this formula, $[\hat{\mu}_j - \epsilon\sigma_j, \hat{\mu}_j + \epsilon\sigma_j]$ is in fact the $1-\alpha$ confidence interval for the ground truth μ_j. As long as we find the $1-\alpha$ confidence interval for μ_j is smaller than $[\hat{\mu}_j - \epsilon\sigma_j, \hat{\mu}_j + \epsilon\sigma_j]$, the inequity in formula (24) can be satisfied, so that the requirement on data credibility can be ensured. Thereafter, evaluating data credibility has become calculating the confidence interval for μ_j.

The confidence interval for μ_j can be calculated based on one of the asymptotic properties of MLE, i.e., *asymptotic normality* [16]:

Theorem* 1. *Asymptotic Normality: *Distribution of MLE estimators for a parameter θ is asymptotically normal with mean θ and variance var(θ), which can be approximated by the inverse of the Fisher information $I(\theta)$:*

$$I(\theta) = E_\theta((\frac{\partial}{\partial\theta}\log f(X|\theta))^2) = -E_\theta(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)) \tag{25}$$

Accordingly, the MLE $\hat{\mu}_j$ is asymptotically normal with mean μ_j, and its variance var(μ_j) is approximated by:

$$\frac{1}{I(\mu_j)} = -\frac{1}{E_{\mu_j}(\frac{\partial^2}{\partial\mu_j^2}\log f(X|\mu_j))} = \frac{\sigma_j^2}{\sum_{i=1}^n d_{ij}r_i^2} \tag{26}$$

From the property of normal distribution, it is easy to obtain the $1-\alpha$ confidence interval for μ_j, i.e.,

$$\begin{aligned}&[\hat{\mu}_j - Z_{\alpha/2}(\frac{1}{\sqrt{I(\mu_j)}}), \hat{\mu}_j + Z_{\alpha/2}(\frac{1}{\sqrt{I(\mu_j)}})]\\ &= [\hat{\mu}_j - Z_{\alpha/2}\frac{\sigma_j}{\sqrt{\sum_{i=1}^n d_{ij}r_i^2}}, \hat{\mu}_j + Z_{\alpha/2}\frac{\sigma_j}{\sqrt{\sum_{i=1}^n d_{ij}r_i^2}}]\end{aligned} \tag{27}$$

where $Z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution.

As long as the $1-\alpha$ confidence interval for μ_j is smaller than $[\hat{\mu}_j - \epsilon\sigma_j, \hat{\mu}_j + \epsilon\sigma_j]$, the requirement on data credibility is satisfied. Otherwise, the server will start another iteration.

V. Performance Evaluations

In this section, we evaluate the performance of the proposed max-credibility approach and min-overhead approach using trace-driven simulations and testbed-based experiments.

A. Trace-Driven Simulations

In our simulation, the mobile opportunistic network is based on a real network trace *Infocom 06* [17] which records the pairwise contact information between 98 mobile users in a conference environment. The server is chosen as the one with the highest centrality in the network, where centrality measures the popularity of a node in the network. The reliability r_i of user i is randomly generated within $[0.2, 1]$ with uniform distribution.

In the experiments, we generate 100 sensing tasks. The ground truth μ_j for task j is randomly generated within $[0, 10]$. The base number σ_j is generated within $[0.2, 5]$. For each sensing task j, the data value observed by user i, i.e., x_{ij}, follows normal distribution $N(\mu_j, (\sigma_j/r_i)^2)$.

1) Results on the Max-Credibility Approach: The proposed max-credibility approach is compared with the following approaches:

- *Max-credibility without considering network reachability*: This is based on our proposed max-credibility approach but does not consider the factor of network reachability when selecting users.
- *Max-credibility without considering user reliability*: It is also based on the proposed max-credibility approach, but does not consider the user reliability in the statistical model. All users are assumed to be reliable.

In addition to the above two approaches, we also compare our max-credibility approach with the existing approaches for truth analysis. Since existing approaches are all designed for categorical data, to be applicable to numerical data, we assume that two data values are equal as long as their difference is within a fixed value (set to ϵ). These approaches use different methods to calculate reliability and credibility, which are defined as follows.

- *Hubs and Authorities (Sums)* [18]: The reliability of a source is the sum of the credibility of the data items it provides, and the credibility of a data item is the sum of the reliability of sources that provide the data.
- *Average·Log* [7]: Different to *Hubs and Authorities*, it computes the reliability of each source by multiplying the average credibility of its provided data item and the logarithm of the number of its provided data item.
- *TruthFinder* [5]: The credibility of an observed data item is the probability that it is accurate and the reliability of the source is the probability that it provides accurate data. The credibility of a data item is computed as the probability that at least one source can provide accurate data. A source's reliability is computed by averaging the credibility of its provided data.

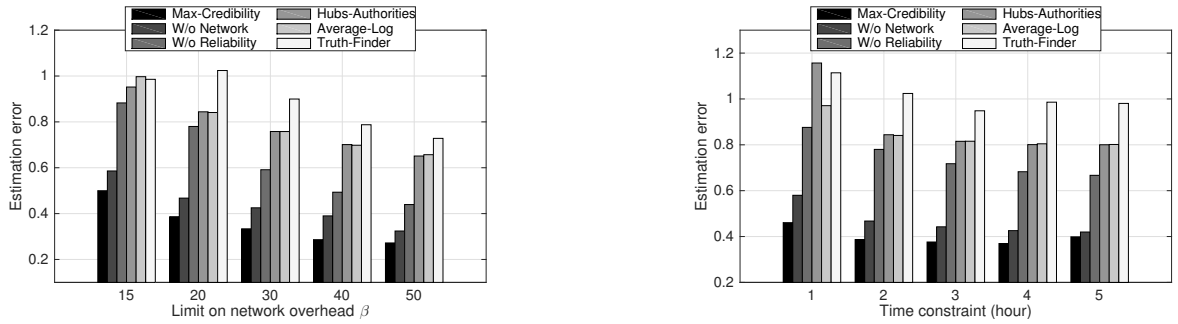
To evaluate the performance of the max-credibility approach, we measure the average *normalized estimation error* of the truth value. The normalized estimation error measures the deviation of the estimated value from the ground truth, and may be larger than 1. Since maximizing data credibility is equal to minimizing the estimation error, a smaller estimation error is always preferred.

In the experiments, we respectively vary the two parameters, i.e., the limit on network overhead β and the time constraint of tasks. The first experiment varies β from 15 to 50 while setting time constraint to two hours, and the second experiment varies time constraint from one to five hours and limits the network overhead to 20. The results are respectively shown in Figure 3a and Figure 3b.

From the results we can see the max-credibility approach has less estimation error compared with the approaches that eliminate the factor of network reachability or the factor of reliability, indicating the importance of both factors on the proposed approach. We also find that eliminating the factor of reliability results to higher performance downgrade than eliminating the factor of network reachability, highlighting the significance of reliability in our approach.

Our max-credibility approach also achieves much less estimation error compared with the three existing approaches on truth analysis. There are two reasons. First, by incorporating the factors of network reachability and reliability, more effective users can be selected. Second, the truth can be estimated effectively by applying the MLE-based truth analysis technique.

2) Results on Min-Overhead: Basically, the min-overhead approach is an iterative approach based on the max-credibility approach. Therefore, in respect to the five approaches in comparison as listed in Section V-A1, there are also five corresponding min-overhead approaches in comparison. For the two approaches that eliminate one factor from max-credibility,



(a) (b)

Fig. 3: Comparison on normalized estimation error with variations of (a) limit on network overhead, (b) time constraint.

we determine when the iteration ends by computing confidence interval. For the three existing approaches in comparison, there is not a straightforward way to compute confidence interval, so we simply let the iteration ends as there are a fixed number of users sending results back.

In these experiments, 100 sensing tasks are generated. The time constraint is set to 12 hours. The requirement on the data credibility is $\frac{|\mu-\hat{\mu}|}{\sigma} < 0.5$ and we set the confidence level to be $1-\alpha = 95\%$. Since the confidence level is high, we find the requirement on data credibility $\frac{|\mu-\hat{\mu}|}{\sigma} < 0.5$ can be satisfied in most cases, which means the success ratio is near 100%. The objective of min-overhead is to minimize the network overhead while achieving the required data credibility. Thereafter, we mainly measure the average network overhead consumed for each task.

In the first experiment, we compare their performance with the variation of the parameter β^o, the limit on network overhead in each iteration. The time constraint in each iteration τ^o is set to two hours. As can be seen from Figure 4a, the min-overhead approach consumes the minimum network overhead in all cases, illustrating the superiority of our approach in minimizing overhead. Moreover, we also notice that the network overhead of our approach reaches minimum when $\beta^o = 20$. As β^o is very small, the MLE estimator usually has bad estimation in the first few iterations, since only a few data values are collected. Therefore, it needs more iterations to collect data in order to re-estimate and the resulted overhead is high. As β^o is very large, the performance is also bad, since many unnecessary data values are collected and much network overhead is wasted.

The second experiment changes the other parameter τ^o, the time constraint in each iteration. β^o is set to 20. From Figure 4b, we can observe a clear superiority of our approach over other approaches. An interesting observation is that less network overhead is consumed as τ^o is larger. There are two reasons. One is that only limited iterations can be completed before the time constraint. For example, as τ^o is three hours, only four iterations can be completed before the 12 hours time constraint, and the network overhead consumed is constrained by $4 * \beta^o = 80$. The other reason is that when τ^o is small, there is not enough time for the requested users to send data back in the first few iterations, resulting to many unnecessary users being requested in the following iterations.

When compared with the three existing approaches, our min-overhead approach consumes much less network overhead, demonstrating the effectiveness of our approach in realizing high data credibility with less network overhead.

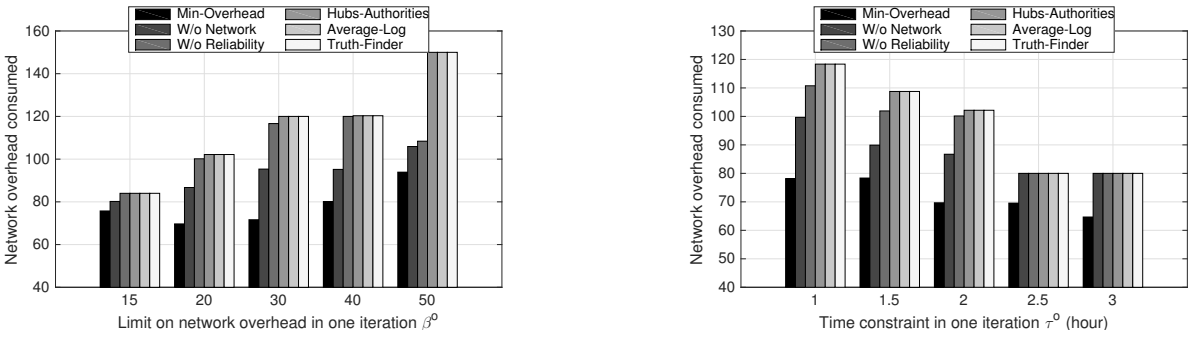
B. Testbed-based Experiment

1) Testbed and Experiment Setting: To further evaluate the performance, we also deployed a testbed of mobile opportunistic network with 20 graduate students in two departments on our campus [19]. In the system, students are distributed with Bluetooth-enabled Android smartphones, with which two students are able to communicate and share data with each other whenever they opportunistically move into the Bluetooth communication range. Finally, about seven months of data on networking are collected.

These 20 students are also required to provide answers to 89 questions about their everyday life and basic knowledge in various topics, e.g., the availability of the parking lots on campus, the estimated driving hours to another city in the local state, or the average salary for the software engineers in US. Since the provided answers are noisy, these data can be ideally used for the evaluation of the truth analysis techniques. To evaluate the proposed approaches, we assume that these questions are the sensing tasks provided by the server in the mobile opportunistic network. If a user is queried with a sensing task, the collected data value is his or her answer to the corresponding question.

Since data from 20 students may not be enough for truth analysis, we also surveyed 40 students in other departments on campus to finish the questions. Their rough locations are also recorded. Based on the networking information we have collected in the testbed, we have captured the contact features between the students in the same department and between students from different departments. First, we experimentally verify the claim in [14] that the pairwise contact process follows Poisson process. Moreover, we find the inverse of the pairwise contact rate $\frac{1}{\lambda}$, i.e., the average inter-contact time, follows log-normal distributions either for the node pairs inside the same department or the node pairs in different departments, with which we are able to generate the contact rates between nodes. Let λ_{in} and λ_{out} respectively denote the pairwise contact rate between nodes inside the same department and that between nodes in different departments, we have

$$\ln \frac{1}{\lambda_{in}} \sim N(\mu_{in}, \sigma_{in}), \text{ and } \ln \frac{1}{\lambda_{out}} \sim N(\mu_{out}, \sigma_{out}) \quad (28)$$



(a) (b)

Fig. 4: Comparison of the min-overhead performance of four approaches with the variations of (a) the network overhead limit in one iteration (β^o) and (b) the time constraint in one iteration (τ^o).

The four parameters $\mu_{in}, \sigma_{in}, \mu_{out}, \sigma_{out}$ are estimated based on our collected networking information in the testbed. Table I shows the estimation on the parameters and the 95% confidence interval of the estimation (the fourth row). A smaller confidence interval means a better fit to the log-normal distribution. As we can see, the parameter estimations have relatively small confidence intervals with range less than 0.5. This result demonstrates the inverse of pairwise contact rate (i.e., the average inter-contact time) can be well approximated by the log-normal distributions. We also believe if there are more students participated in the testbed, the distribution fit can be even more accurate.

TABLE I: **The parameters of log-normal distributions**

Inside department ($\ln \frac{1}{\lambda_{in}}$)		Between department ($\ln \frac{1}{\lambda_{out}}$)	
μ_{in}	σ_{in}	μ_{out}	σ_{out}
2.09	1.15	3.51	0.37
[1.86, 2.33]	[1.01,1.34]	[3.26,3.76]	[0.26, 0.64]

Based on the extracted contact features, we generate synthetic contact information (contact rate) among the 40 students as well as between the 40 students and 20 students in the testbed. Specifically, for each pair of students u, v, their contact rate λ_{uv} is generated based on the log-normal distributions in (28). Then the detailed contact process between them is generated to follow Poisson process with rate λ_{uv}. Finally, a mobile opportunistic network of 60 mobile users is built by integrating the synthetic contact information and the authentic networking information collected in the testbed.

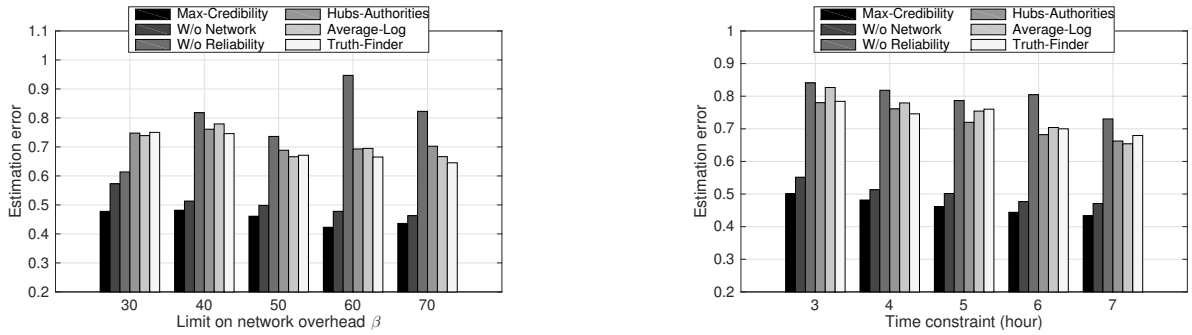
2) Results: Based on the testbed and the collected data, we first evaluate our approaches by comparing the performance on the max-credibility problem. The approaches listed in Section V-A1 are compared. The results are shown in Figure 5. The first experiment varies the limit on the network overhead β and set the time constraint to be four hours (Figure 5a). The second experiment varies the time constraint and set the limit on network overhead to be 40 (Figure 5b). As can be seen from the results, our approach achieves smaller estimation error than the other approaches, similar to what we observed in the trace-driven simulation. These results further demonstrate the effectiveness of our approach on maximizing credibility in real network implementation.

Since the most important part of our min-overhead approach is its ability to iteratively and adaptively collect data from users. We next evaluate our min-overhead approach by comparing the iterative approach with the non-iterative approach. In this experiment, we set a stringent requirement on data credibility: $\frac{|\mu-\hat{\mu}|}{\sigma} < 0.5$ and set the confidence level to be 90%. The performance is evaluated based on the success ratio, which is the percentage of tasks that can successfully satisfy the requirement on data credibility. The result is shown in Figure 6. As can be seen, the iterative approach consistently achieves higher success ratio than the non-iterative approach under different network overhead limit. The superior performance on success ratio is because the iterative approach can adaptively collect data from the mobile opportunistic networks when the requirement on data credibility is not satisfied, but a non-iterative approach is not flexible and cannot collect data adaptively. Even though the non-iterative approach may achieve a high success ratio by requesting a lot of users, it is still difficult to specify what and how many users are queried. Our approach addresses this problem through adaptive collection, achieving high data credibility with low network overhead.

VI. Related Work

Truth analysis in social sensing or crowdsourcing has received considerable attention recently. Some solutions including Hubs and Authorities [18], Averge·Log [7] and TruthFinder [5] are based on a simple heuristic: the correctness of information is estimated based on the reliability of the information sources, and then the reliability of the information source is estimated based on the correctness of their provided information. Other than the heuristic-based approaches, researchers also investigated statistical models for truth finding in order to achieve more convincing results. For example, the bayesian inference is utilized in [8]. An expectation-maximization (EM) based scheme is utilized in [9][10].

The research on crowdsourcing mostly focused on the design of incentive mechanisms and user recruitment shcemes. For example, [1][20] designed incentive schemes for the smartphone-based crowdsourcing. Other work [13][21] designed schemes to recruit users to achieve high information quality and minimize total expenses or resources. However, they failed to consider how user reliability influences the quality of information, and lacked a recruiting mechanism to collect enough data to assure high information quality. Our



(a) (b)

Fig. 5: Comparison of normalized estimation error for the max credibility problem in testbed-based experiment: (a) with different constraint on network overhead β, and (b) with different time constraint.

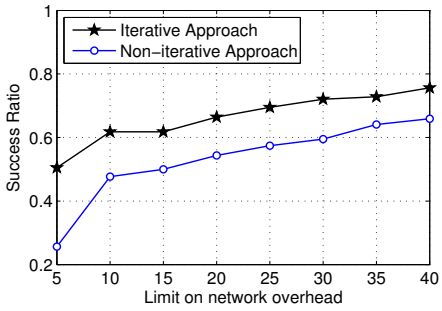


Fig. 6: The comparison on success ratio between the iterative approach and non-iterative approach.

paper addresses this problem by adaptively collecting data from users until the required information quality is achieved.

VII. Conclusion

An important problem in crowdsourcing is to assess the reliability of the mobile users and identify the truth from the reported data. Different from existing truth analysis approaches, we utilize the communication network to adaptively collect data from some users, so that high-quality data are collected to achieve the required data credibility. Considering the requirement on data credibility and the constraint of network resources, we quantify the tradeoff between the enhanced data credibility and the increased network overhead as two problems: max-credibility which aims to maximize data credibility with some network overhead, and min-overhead which aims to achieve a specified data credibility while minimizing the network overhead. For max-credibility, we propose a solution to select users based on their data collection capabilities such as network reachability and user reliability, and propose a MLE-based technique to identify truth from the collected data. For Min-overhead, we propose a solution which iteratively selects users for data collection until the required data credibility is achieved. The proposed solutions are evaluated based on trace-driven simulations and a testbed. Both simulation and experimental results demonstrate the effectiveness of the proposed solutions in terms of data credibility and network overhead.

References

[1] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing," in *ACM MobiCom*, 2012.

[2] Y. Wang, W. Hu, Y. Wu, and G. Cao, "Smartphoto: a resource-aware crowdsourcing approach for image sensing with smartphones," in *ACM MobiHoc*, 2014.

[3] Y. Wu, Y. Wang, W. Hu, X. Zhang, and G. Cao, "Resource-aware photo crowdsourcing through disruption tolerant networks," in *IEEE ICDCS*, 2016.

[4] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson, "Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones," in *ACM Sensys*, 2009.

[5] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, 2008.

[6] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, "Corroborating information from disagreeing views," in *ACM WSDM*, 2010.

[7] J. Pasternack and D. Roth, "Knowing what to believe (when you already know something)," in *COLING*. Association for Computational Linguistics, 2010.

[8] B. Zhao and J. Han, "A probabilistic model for estimating real-valued truth from conflicting sources," *Proc. of QDB*, 2012.

[9] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *ACM IPSN*, 2012.

[10] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *IEEE ICDCS*, 2013.

[11] K. Fall, "A delay-tolerant network architecture for challenged internets," in *ACM SIGCOMM*, 2003.

[12] W. Gao, Q. Li, B. Zhao, and G. Cao, "Social-aware multicast in disruption-tolerant networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 20, no. 5, pp. 1553–1566, 2012.

[13] M. Karaliopoulos, O. Telelis, and I. Koutsopoulos, "User recruitment for mobile crowdsensing over opportunistic networks," in *IEEE INFOCOM*, 2015.

[14] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in delay tolerant networks: a social network perspective," in *ACM MobiHoc*, 2009.

[15] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han, "A confidence-aware approach for truth discovery on long-tail data," *Proceedings of the VLDB Endowment*, vol. 8, no. 4, pp. 425–436, 2014.

[16] G. Casella and R. L. Berger, *Statistical inference*. Duxbury Pacific Grove, CA, 2002, vol. 2.

[17] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, 2007.

[18] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[19] W. Gao, W. Hu, and G. Cao, "Interest-based data dissemination in opportunistic mobile networks: Design, implementation and evaluation," in *Opportunistic Mobile Social Networks, Book Chapter*, 2014.

[20] I. Koutsopoulos, "Optimal incentive-driven design of participatory sensing systems," in *IEEE INFOCOM*, 2013.

[21] Z. He, J. Cao, and X. Liu, "High quality participant recruitment in vehicle-based crowdsourcing using predictable mobility," in *IEEE INFOCOM*, 2015.