This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

An Efficient Prediction-Based User Recruitment for Mobile Crowdsensing

En Wang; Yongjian Yang†; Jie Wu, *IEEE Fellow*; Wenbin Liu; Xingbo Wang

***Abstract*—Mobile crowdsensing is a new paradigm in which a group of mobile users exploit their smart devices to cooperatively perform a large-scale sensing job. One of the users' main concerns is the cost of data uploading, which affects their willingness to participate in a crowdsensing task. In this paper, we propose an efficient Prediction-based User Recruitment for mobile crowdsEnsing (PURE), which separates the users into two groups corresponding to different price plans: Pay as you go (PAYG) and Pay monthly (PAYM). By regarding the PAYM users as destinations, the minimizing cost problem goes to recruiting the users that have the largest contact probability with a destination. We first propose a semi-Markov model to determine the probability distribution of user arrival time at points of interest (PoIs) and then get the inter-user contact probability. Next, an efficient prediction-based user-recruitment strategy for mobile crowdsensing is proposed to minimize the data uploading cost. We then propose PURE-DF by extending PURE to a case in which we address the tradeoff between the delivery ratio of sensing data and the recruiter number according to Delegation Forwarding. We conduct extensive simulations based on three widely-used real-world traces: *roma/taxi*, *epfl*, and *geolife*. The results show that, compared with other recruitment strategies, PURE achieves a lower recruitment payment and PURE-DF achieves the highest delivery efficiency.**

***Index Terms*—Mobile crowdsensing, User recruitment, Semi-Markov, Uploading cost**

I. Introduction

The proliferation of smartphones, which have generally been equipped with multi-core processors and a set of sensors (e.g., camera, light sensor, chemical sensor and GPS) that allow them to be seen as powerful mobile sensors with sensing ability, have experienced explosive growth in recent years. Thanks to this, a new sensing paradigm called *mobile crowdsensing* is proposed [1] to recruit a group of mobile users who can jointly perform a large-scale sensing task through their smartphones. The aggregation and processing of the sensing data collected by mobile users' smartphones gives rise to diverse services ranging from traffic jam prediction and parking space management to indoor localization and environmental monitoring [2], [3], [4], [5], [6], [7].

E. Wang is with the Department of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China. E-mail: wangen@jlu.edu.cn.

†Y. Yang is the corresponding author and is with the Department of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China. E-mail: yyj@jlu.edu.cn.

J. Wu is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA. E-mail: jiewu@temple.edu.

W. Liu is with the Department of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China. E-mail: liuwb16@mails.jlu.edu.cn.

X. Wang is with the Department of Software, Jilin University, Changchun, Jilin 130012, China. E-mail: xbwang15@mails.jlu.edu.cn.

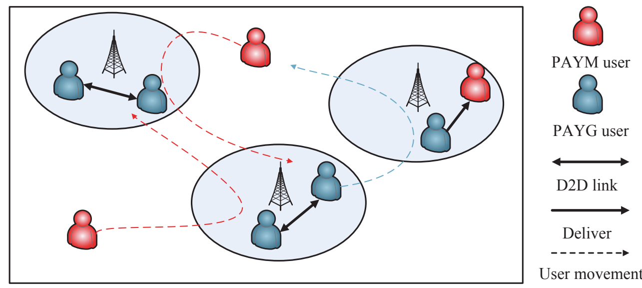


Fig. 1. User recruitment for mobile crowdsensing. The sensing tasks are assigned in PoIs, and we decide which PAYG user will take the sensing data. The PAYG user with the sensing data can either (1) encounter a PAYM user before the task deadline and upload the sensing data freely or (2) upload the sensing data costly (via 3G or 4G) when the deadline expires.

There has been much research on mobile crowdsensing including platform design [8], [9], [10], [11], user recruitment strategies [12], [13], [14], and incentive mechanisms [15], [16], [17], [18], [19]. Research on platform design focuses on proposing a framework or system for mobile crowdsensing, while research in terms of incentive mechanisms focuses on designing incentive mechanisms for crowdsensing to attract users to participate in the crowdsensing task. Among them, a common challenge for most mobile crowdsensing applications is to identify mobile users who can contribute the most value to the sensing task. Therefore, the user recruitment strategy is one of the most important topics of discussion. However, previous studies do not consider the uploading cost of sensing data. In this paper, we focus on proposing an efficient Prediction-based User Recruitment for mobile crowdsEnsing (PURE) where multiple users with a higher contact probability to the destinations can be recruited to cooperatively perform a common task, ensuring that the expected data-uploading cost is minimal. In PURE, users can be divided into the following two groups [20] according to their common price plans:

- Pay As You Go (PAYG): a user pays a data cost according to the amount of data transferred, e.g., \$0.2/MB.
- Pay Monthly (PAYM): a user can transfer an unlimited amount of data during a month-long period, e.g., \$8/month.

The price plan is decided by the user's preference and has no relationship to the sensing task. Assuming that a sensing task wants to collect real-time sensing data from many points of interest (PoIs) in an urban area (e.g., current traffic jam situations in some PoIs), and the sensing data needs to be uploaded before the given deadline. Some mobile users move around in the urban area every day. Users might pass by some

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

PoIs frequently, so they can be recruited to collect sensing data if they happen to be in the PoI when the sensing task is published.

If a mobile user participates in crowdsensing, the user will charge a recruiting payment from the publisher of the sensing task. However, this paper focuses on the uploading payment, not the recruiting payment. The recruiting payment is much less than uploading payment, because the recruiting payment is used to make up for the resource consumptions of the users' devices (energy, bandwidth, and memory), users do not need to pay extra money. Moreover, a user could only get the recruiting payment once. A user has two different ways to upload the sensing data: (1) encounter a PAYM user before the task deadline and upload the sensing data freely or (2) upload the sensing data costly (via 3G or 4G) when the deadline expires. As shown in Fig. 1, the *first problem* occurs, when the sensing task is published and the publisher needs to determine which users should be recruited in each PoI so that the total uploading cost can be minimized. To solve the first problem, we propose the user-recruitment strategy (PURE) which uses the semi-Markov model to predict which PAYG users have the highest contact probability with PAYM users so that we can recruit them for each PoI.

Through PURE, some PAYG users with higher contact probabilities with PAYM users are recruited. When they encounter a PAYM user before a deadline, they can upload sensing data freely. Otherwise, during the mobility process, selected PAYG users may encounter other PAYG users who are "better" (with a higher contact probability with PAYM users). The users selected by PURE could recruit the other "better" users by informing the publisher of the sensing task. The publisher will pay the recruiting payment. More recruiters leads to a higher delivery ratio, but also to a higher recruitment cost. The *second problem* is determining how to balance the tradeoff between the delivery ratio of sensing data and the recruiter number. To address this problem, we propose PURE-DF to improve the delivery ratio while controlling the total recruiter number, according to Delegation Forwarding.

The main contributions of this paper are briefly summarized as follows:

- We propose a points-of-interest (PoI) trajectory-prediction method that uses a semi-Markov process to determine the probability distribution of user arrival times at PoIs. Furthermore, inter-user contact probability is predicted according to the PoI prediction.
- We propose an efficient Prediction-based User Recruitment strategy for mobile crowdsEnsing (PURE), which groups two types of users with different data uploading schemes (Pay as you go (PAYG) and Pay monthly (PAYM)) that will cooperatively finish a crowdsensing task.
- Taking the relation between delivery ratio of sensing data and recruiter number into consideration, we propose PURE-DF, based on Delegation Forwarding [21], to address the trade-off.

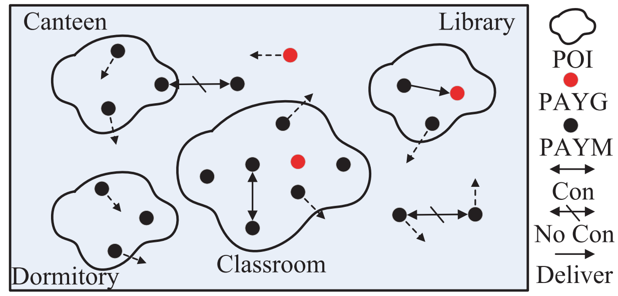


Fig. 2. A running example of mobile crowdsensing network model.

- We conduct extensive simulations based on three widely-used real-world traces: *roma/taxi*, *epfl*, and *geolife*. The results show that compared with other recruitment strategies, PURE achieves a lower recruitment payment and PURE-DF achieves the highest delivery efficiency.

The remainder of this paper is organized as follows: The system model (network, semi-markov and contact probability models) and problem formulation are presented in Section II. The recruitment strategies (PURE and PURE-DF) are proposed in Section III. In Section IV, we evaluate the performance of the proposed recruitment strategies through extensive simulations. We review the related work in Section V. We conclude the paper in Section VI.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Network Model

We consider a mobile crowdsensing environment that is composed of a crowd of mobile users, denoted by the set $U = \{u_1, u_2, \cdots, u_n\}$. The users could be divided into two groups according to their price plans: Pay as you go (U_G) and Pay monthly (U_M). PAYM users can upload the sensing data freely, while PAYG users can only upload the data costly. There are some points of interest (PoIs): $L = \{1, 2, \cdots, l\}$. In every PoI, two users can communicate with each other and can deliver or replicate the sensing data to each other. Every crowdsening task generates at some PoIs, and recruits one user in each PoI to upload the sensing data before the deadline.

Here, we say that two users are in contact when they are in the same PoI and can directly communicate with each other via PoI access points. When the PoI is very large, we assume that users could form a mobile ad hoc network because the PoI in mobile crowdsensing is commonly a dense region of users. Two users cannot communicate with each other when one of them is not in a PoI. The network model is shown in Fig. 2, where four PoIs —Canteen, Library, Dormitory, and Classroom— exist in the network. A PAYG user can deliver the sensing data to a PAYM user or replicate to another PAYG user when they are in the same PoI. In this paper, we assume that the communication duration and bandwidth are sufficient for each user to receive sensing tasks and deliver or replicate the sensing data. The main notations are illustrated in Table I.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

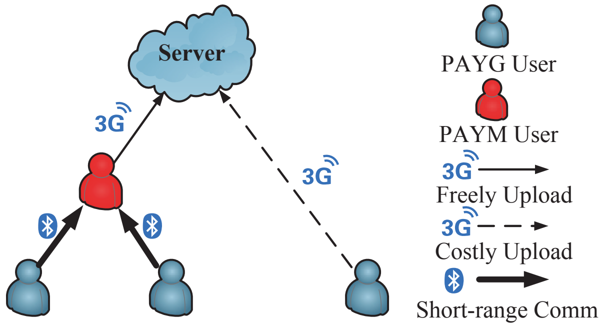


Fig. 3. Two kinds of uploading action for a PAYG user. The PAYG user with sensing data can either (1) encounter a PAYM user before the task deadline and upload the sensing data freely or (2) upload the sensing data costly (via 3G or 4G) when the deadline expires.

B. Problem

We consider the user recruitment problem for mobile crowdsensing in the above network model. Without loss of generality, a sensing task will assign some sensing data, denoted by $D = d_1, d_2, \cdots, d_l$, in all the PoIs at time t. The sensing task has an uploaded deadline T_d which means that the sensing data must be uploaded before time $t + T_d$. Only a user in a PoI at time t can be recruited to sense the data and perform one of the following two uploading actions before deadline (Fig. 3): (1) deliver the sensing data to a PAYM user and upload freely via the PAYM user or (2) upload the sensing data itself in a costly manner (via 3G or 4G).

Assuming that the payment for uploading through a PAYG user is 1 and that the payment for uploading through a PAYM user is 0, the sensing task prefers that a PAYG user upload the sensing data through a PAYM user rather than uploading by itself. Therefore, for each PoI without a PAYM user, we should recruit the PAYG user with the highest probability of contact with the PAYM users before the deadline.

The *first question* is how to decide which PAYG user in the PoI is the best recruiter. To address this, we use a semi-Markov model to predict the contact probability with PAYM users and to recruit the PAYG user with the highest probability. To do this, we propose a user recruitment strategy (PURE) in which multiple users can be recruited to cooperatively perform a common task, ensuring that the expected data-uploading payment is minimal.

We regard the initially recruited PAYG users as sources, and consider the PAYM users destinations. Sources can recruit other encountered PAYG users to deliver sensing data to a destination before the deadline. However, the sensing task should also pay the users that are recruited later (recruitment cost). More recruiters leads to a higher delivery ratio, but also to a higher recruitment cost.

The *second question* is how to balance the tradeoff between the delivery ratio and the recruiter number. To address the second question, we propose PURE-DF to improve the delivery ratio while controlling the total recruiter number, according to Delegation Forwarding. In PURE-DF, the PAYG user with the sensing data only recruits a PAYG user and replicates the sensing data if the encountered PAYG user's contact probability with PAYM users is greater than any seen by the sensing data so far.

TABLE I
MAIN NOTATION USED THROUGHOUT THE PAPER

Symbol	Meaning
U	whole user set
U_G	PAYG user set
U_M	PAYM user set
D	sensing data set
T_d	sensing data's uploaded deadline
l	total number of PoIs
S_n^k	state holding time for user k in the nth state
$Z_{ij}^k(t)$	probability of user k transiting from state i to state j at time t
P_{ij}^k	transition probability from state i to state j
$G_{ij}^k(t)$	probability that user k will transit from state i to state j before time unit t
$G_i^k(t)$	probability that user k will leave the PoI i on or before time unit t
$Q_{ij}^k(t)$	probability that the user k's current state is i, it will be in state j after t time units
$C_{ab}^i(h)$	contact probability between user a and user b at the PoI i and time h
$C_{ab}(h)$	probability that users a and b are in contact at the time h
$R_{ab}(h)$	probability of the first contact at time h between users a and b
F_{ab}	maximum probability of the first contact between a and b before the deadline T_d
$F_{U_{G_i}U_M}$	probability that a PAYG user could contact anyone of the PAYM users (U_M) before deadline

C. Semi-Markov Model

The states of user k are recorded as the PoIs $L^k = \{1, 2, 3, \cdots, l\}$, which represents the PoI id that user k is being now. l represents the total number of PoIs and the nth state of user k is recorded as L_n^k, which is the nth PoI in the path of user k. The beginning time of user k to be in the nth state is recorded as T_n^k, which represents the time to enter the nth PoI in the path of user k.

We model the mobility of user k with a time homogeneous semi-Markov [22], [23] (L_n^k,T_n^k) with discrete time because the probability of a user k transiting from state L_n^k to state L_{n+1}^k is independent of state L_{n-1}^k. Thus, process L_n^k is a standard discrete-time Markov Chain. The random variable T_n^k represents the time instant of the transition from L_n^k to L_{n+1}^k. Random variable $S_n^k = T_{n+1}^k - T_n^k$ describes the PoI sojourn time, or state holding time. These random variables are independent and identically distributed (i.i.d.), with distributions that do not change over time (time-homogeneous). For example, for a student, the expected breakfast time is 20 mins, and expected class time is 45 mins. So this can be different from the geometric or exponential distributions (semi-Markov).

The associated time-homogeneous semi-Markov kernel part is defined in Eq. 1, where $Z_{ij}^k(t)$ represents the probability of user k transiting from state i to state j at time t. It is not difficult to find that in Eq. 1, L_{n+1}^k depends on L_n^k, but has no relationship with L_{n-1}^k.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

$$\begin{aligned} Z_{ij}^k(t) &= P(L_{n+1}^k = j, S_n^k \le t | L_0^k \cdots L_n^k, T_0^k \cdots T_n^k) \\ &= P(L_{n+1}^k = j, S_n^k \le t | L_n^k = i) \end{aligned} \quad (1)$$

Suppose $P^k = P_{ij}^k$ is the transition probability matrix of the L_n^k embedded Markov chain for user k. Then, the transition probability from state i to state j is shown in Eq. 2, where num_i^k stands for the number of transitions from PoI i without considering the next PoI, and where num_{ij}^k is the number of transitions from PoI i to PoI j.

$$P_{ij}^k = P(L_{n+1}^k = j | L_n^k = i) = num_{ij}^k / num_i^k \quad (2)$$

The probability that user k will transit from state i to state j before time unit t is defined as $G_{ij}^k(t)$, which is shown in Eq. 3.

$$\begin{aligned} G_{ij}^k(t) &= P(S_n^k \le t | L_n^k = i, L_{n+1}^k = j) \\ &= \sum_{x=1}^{t} P(S_n^k = x | L_n^k = i, L_{n+1}^k = j)) \end{aligned} \quad (3)$$

Also, we can achieve the probability $G_i^k(t)$ that user k will leave the PoI i on or before time unit t as follows:

$$G_i^k(t) = P(S_n^k \le t | L_n^k = i) = \sum_{j=1, j \ne i}^{l} Z_{ij}^k(t). \quad (4)$$

$S_n^k = T_{n+1}^k - T_n^k$ describes the PoI state holding time. It is not difficult to find that $G_i^k(t)$ also indicates the distribution of the state-holding time at PoI i for user k, regardless of the next state.

According to Eqs. 1- 3, we could derive the associated time-homogeneous semi-Markov kernel part Z_{ij}^k, which is shown as Eq. 5.

$$\begin{aligned} Z_{ij}^k(t) &= P(L_{n+1}^k = j, S_n^k \le t | L_0^k \cdots L_n^k, T_0^k \cdots T_n^k) \\ &= P(S_n^k \le t | L_n^k{=}i, L_{n+1}^k{=}j) P(L_{n+1}^k{=}j | L_n^k{=}i) \\ &= G_{ij}^k(t) P_{ij}^k \end{aligned} \quad (5)$$

Let $Q_{ij}^k(t)$ be another time-homogeneous semi-Markov that describes the probability that the user k is currently in PoI i. After t time units, it will be in PoI j. Note that, $Q_{ij}^k(t)$ is different from $Z_{ij}^k(t)$ in Eq. 1. $Z_{ij}^k(t)$ represents the probability that the current state of user k is i, the next state is j, and the transiting time is not more than t. However, $Q_{ij}^k(t)$ is the probability that the user k's current state is i, and that it will be in state j after t time units. That is to say, the transition may be more than one hop; the transiting process from state i to j could pass through another one or more states. $Q_{ij}^k(t)$ provides an easy prediction of the user's location at an arbitrary time unit t, with knowing its current location. The derivation of $Q_{ij}^k(t)$ is described next.

It is worth noting that for any initial state i of user k, the transition in the further time t, $\sum_{j=1}^{l} Q_{ij}^k(t) = 1$. For the initial case $t = 0$ without any iteration, if $j \ne i$, then $Q_{ij}^k(0) = 0$. Similarly, if $j = i$, then $Q_{ij}^k(0) = 1$.

In order to achieve $Q_{ij}^k(t)$, we start with a special case: user k will not leave initial state i before time unit t. In other words, user k has no transition before time t. In this case, according to Eq. 4, the probability is shown in Eq. 6

$$P(S_n^k > t | L_n^k = i) = 1 - G_i^k(t) \quad (6)$$

Next, we consider a second case: user k has at least one transition between times 0 and t, and we assume that the first transition happens at time x to PoI r. Then, we can achieve the probability in this case as follows:

$$\begin{aligned} &P(L_t^k = j | L_0^k = i \ and \ at \ least \ one \ transition \ to \ r) \\ &\qquad = \sum_{r=1}^{l} \sum_{x=1}^{t} (Z_{ir}^k(x) - Z_{ir}^k(x-1)) Q_{rj}^k(t-x), \end{aligned} \quad (7)$$

where $(Z_{ir}^k(x) - Z_{ir}^k(x-1))$ means that the first transition happens at time x to PoI r.

Combine the above two cases, we achieve $Q_{ij}^k(t)$ as follows:

$$Q_{ij}^k(t) = \begin{cases} \sum_{r=1}^{l} \sum_{x=1}^{t} (Z_{ir}^k(x) - Z_{ir}^k(x-1)) Q_{rj}^k(t-x), j \ne i \\ 1 - G_i^k(t) + \\ \sum_{r=1, r \ne i}^{l} \sum_{x=1}^{t} (Z_{ir}^k(x) - Z_{ir}^k(x-1)) Q_{rj}^k(t-x), j = i \end{cases} \quad (8)$$

With regard to Eq. 8, we have the following explanations. Case 1: when $j \ne i$, $Z_{ir}^k(x) - Z_{ir}^k(x-1)$ means that the first transition step is from i to r, which happens at time x. Then, $Q_{rj}^k(t-x)$ means that the current state is r. After time $t-x$, the state turns to j. Case 2: when $j = i$, $1 - G_i^k(t)$ means that user k has no transition before time t. $\sum_{r=1, r \ne i}^{l} \sum_{x=1}^{t} (Z_{ir}^k(x) - Z_{ir}^k(x-1)) Q_{rj}^k(t-x)$ means that user k has at least one transition to r ($r \ne i$), and then get back to i at time t. According to this analysis, $Q_{ij}^k(t)$ can be correctly obtained by Eq. 8.

The derive process of $Q_{ij}^k(t)$ utilizes the thought of Dynamic Programming (DP). Because we know the initial value of $Q_{ij}^k(0)$, (1) $Q_{ij}^k(0) = 0$ when $j \ne i$ and (2) $Q_{ij}^k(0) = 1$ when $j = i$. Time t of $Q_{ij}^k(t)$ represents the phase number in DP. We can use the results of $t = 0$ to calculate the results of $t = 1$ and so on. Therefore, $Q_{ij}^k(t)$ could be achieved through $Q_{ij}^k(0)$ to $Q_{ij}^k(t-1)$.

Note that the derivation of $Q_{ij}^k(t)$ needs to use $Z_{ij}^k(t)$ matrix, which is derived through two other matrices: the state holding time probability matrix $G_{ij}^k(t)$ and the transition probability matrix P_{ij}^k. With these matrices, we can predict the future PoI of user k based on its current PoI using probability distributions $Q_{ij}^k(t)$. Therefore, the problem comes to get the matrices: $G_{ij}^k(t)$ and P_{ij}^k.

We retrieve both matrices of user k from its history path composed of 2-tuples (L_n^k,T_n^k). P is the transition probability

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

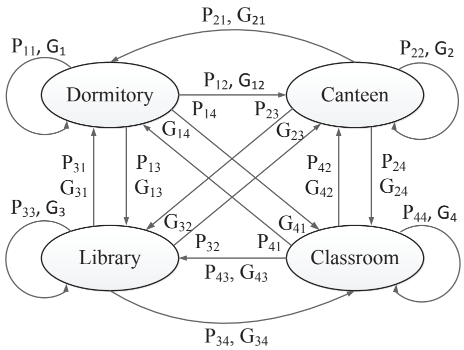


Fig. 4. The matrices P and G in the semi-Markov model.

matrix of the embedded Markov chain. G is the state-holding time probability distribution matrix. Fig. 4 shows an example of transition probability matrix and state-holding time probability distribution matrix for user k that visits four PoIs: Dormitory (PoI 1), Canteen (PoI 2), Library (PoI 3), and Classroom (PoI 4). For each PoI, the user can choose to stay for a while or to move to another PoI according to its preferred probability. For example, if the user is at the Dormitory (PoI 1), it can undertake one of the following four actions:

(1) move to the Canteen with probability P_{12} and transition time probability distribution $G_{12}(t)$
(2) move to the Library with probability P_{13} and transition time probability distribution $G_{13}(t)$
(3) move to the Canteen with probability P_{14} and the transition time probability distribution $G_{14}(t)$
(4) stay in the Dormitory with the probability P_{11} and the transition time probability distribution $G_1(t)$

It is not difficult to find that P_{ij} can be achieved by Eq. 2. For example, the state list of L_n^k is $2, 1, 2, 1, 2, 3, 5$, and $P_{21}^k = 2/3$. It is not difficult to find that, $num_{ij}^k \leq num_i^k$ and $P_{ij}^k \leq 1$. By keeping track of num_{ij}^k and num_i^k, each user can generate and refine its own P matrix over time. Similarly, G_{ij}^k can be obtained through Eq. 3. For example, the time list of transiting from state i to state j is $2, 3, 4, 4, 5, 6, 9$, then $G_{ij}^k(4) = 4/7$. Based on the above descriptions, user k can achieve $Q_{ij}^k(t)$ through the results of matrices P and G. Moreover, in the general Markov process, the state holding time is usually considered to be an exponential distribution. The semi-Markov used in our model eliminates this constraint and is more reflective of real-world processes.

D. Contact Probability Model

In this paper, we try to propose an efficient user recruitment strategy where multiple users can be recruited to cooperatively perform a common task, ensuring that the expected data uploading-cost is minimal. As previously described, we divide all the users into two groups corresponding to different price plans: PAYG users and PAYM users. In each POI, we attempt to recruit the PAYG user that has the highest probability of contacting a PAYM user before the deadline. In this way, a PAYG user can freely upload the sensing data through a PAYM user.

Next, the question is how to achieve the contact probability between a PAYG user and a PAYM user. Probability distributions $Q_{ij}^k(t)$ give the probability that the future PoI at time t of a user k will be j, with the condition that at time 0 the PoI was i. We assume that, the paths of users are independent of each other and that the most recently known state of user a is PoI l_a at time h_a, and that of user b is PoI l_b at time h_b. Therefore, the contact probability between user a and user b at PoI i and time h is shown as follows:

$$C_{ab}^i(h) = Q_{l_a i}^a(h - h_a) Q_{l_b i}^b(h - h_b), h > 0, \tag{9}$$

where $h > h_a > 0$, and $h > h_b > 0$. Then, the probability that users a and b are in contact at time h at any PoI is

$$C_{ab}(h) = \sum_{i=1}^{l} C_{ab}^i(h), h > 0. \tag{10}$$

Because the purpose of our study is to select the optimal PAYG user who has the highest contact probability with any PAYM user, the first contact probability is useful for us. We define the probability that two users begin their first contact at time h. Note that when we talk about users a and b beginning their first contact at time h, it means that they had no contact at any prior time unit. Assuming that user trajectories are independent, the probability of the first contact at time h between users a and b is defined as follows:

$$R_{ab}(h) = C_{ab}(h) \prod_{t=0}^{h-1} (1 - C_{ab}(t)), h > 0. \tag{11}$$

Denote the maximum data uploading deadline as T_d, which means that a PAYG user a with sensing data is required to contact a PAYM user b in time T_d. Thus, maximum probability of the first contact between a and b before the deadline T_d is shown in Eq. 12.

$$F_{ab} = \sum_{h=1}^{T_d} R_{ab}(h) \tag{12}$$

According to Eq. 12, for a PAYG user U_{G_i}, the probability of contacting a PAYM user (U_M) before the deadline is shown in Eq. 13, where m is the total number of PAYM users.

$$F_{U_{G_i} U_M} = 1 - \prod_{y=1}^{m} (1 - F_{U_{G_i} U_{M_y}}) \tag{13}$$

According to Eq. 13, we can decide which PAYG user is optimal to upload the sensing data in each PoI in order to minimize the total uploading cost.

III. User Recruitment Strategy

A. PURE

In this section, we propose an efficient Prediction-based User Recruitment for mobile crowdsEnsing (PURE) in order to address the following question: how can we recruit a user in each PoI in a way that will minimize the total uploading

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

Algorithm 1 PURE

Input:

PAYG users: U_G

PAYM users: U_M

```
1: Calculate the F_{U_{G_i} U_M} for each PAYG user in the POIs
2: for each PoI do
3:     if there is a PAYM user U_{G_i} in this PoI then
4:         upload the sensing data through the PAYM user
5:     else if there is no user in this PoI then
6:         no user could be recruited
7:     else
8:         recruit the U_{G_i} with highest F_{U_{G_i} U_M} in this PoI to sense the data
```

cost? To save uploading cost, a PAYG user prefers uploading sensing data through a PAYM user, rather than uploading by itself. The analysis above shows that, the problem in each PoI becomes how to find the PAYG user with the highest probability of encountering a PAYM user before the deadline. In the previous section, we calculated the contact probability by $F_{U_{G_i}U_M}$. Therefore, the recruitment strategy PURE is proposed in Algorithm 1.

As shown in Algorithm 1, users are divided into two groups: PAYG users and PAYM users. If there is a PAYM user in a PoI at the sensing task time, the PAYM user will be recruited to upload the sensing data freely. Otherwise, if there is no PAYG user in it, we could not sense any data in this PoI. If there is no PAYM user in the PoI, but one or more PAYG users in it, then the PAYG user with the highest $F_{U_{G_i}U_M}$ will be recruited to sense the data. Moreover, the running time for line 2 of Algorithm 1 is *O(l)*, where l is the total number of PoIs. And the running time for line 8 of Algorithm 1 is *O(n)*, where n is total number of PAYG users. The worst case is that PURE needs to select the best user among all the PAYG users. Therefore, it can be seen that Algorithm 1 can be implemented to run in *O(ln)* time.

Algorithm 2 PURE-DF

Input:

Recruit initial users $U_{G_i}, (0 < i \le C)$ through PURE (Algorithm 1)

Sensing data: $d_1, d_2, \cdots, d_l$

```
INITIALIZE ∀i : H_i ← F_{U_{G_i} U_M}
On contact between U_{G_i} with sensing data and U_{G_j} without sensing data
for k =1 to l do
    if d_k is currently held by U_{G_i}, and U_{G_j} without d_k then
        if H_i < F_{U_{G_j} U_M} then
            H_i ← F_{U_{G_j} U_M}
            recruit user U_{G_j}
            replicate d_k from U_{G_i} to U_{G_j}
```

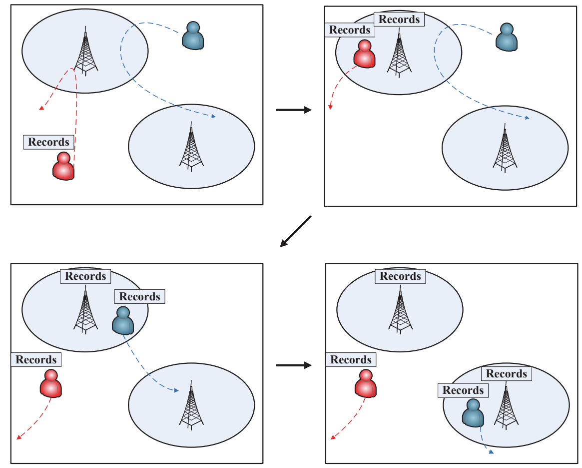


Fig. 5. Dissemination process of the PAYM users' parameters with PoI-assistant. We regard the PoIs as the stationary users and users exchange and update their records of the PAYM users' parameters. With the help of PoIs, the parameters could be disseminated more quickly. Therefore, the warm-start time can be further reduced.

B. PURE-DF

In this section, we propose PURE-DF by extending PURE to a case where we address the tradeoff between the delivery ratio of sensing data and the recruiter number according to the thought of Delegation Forwarding. PURE-DF first recruits the initial users U_{G_i} in each PoI through PURE, and then, PURE-DF decides whether to recruit the encountered U_{G_j} by the initial users U_{G_i} according to their contact probabilities with U_M before the uploading deadline.

Based on the contact probabilities with U_M users, two simple recruitment strategies are naturally proposed. The first is the PURE recruitment strategy, which only recruits the initial users U_{G_i}, and waits for them to forward the sensing data to U_M users. The second is Epidemic recruitment strategy, which recruits every encounter to assist in forwarding the sensing data to U_M users. It is not difficult to find that the PURE recruitment strategy uses fewer PAYG users, though its contact probability cannot be guaranteed. However, the Epidemic recruitment strategy can achieve the highest contact probability, but it also incurs the highest recruitment cost. Recruiting more users leads to higher contact probability, which means a lower uploading cost, but which also leads to higher recruitment cost. Therefore, taking the Delegation Forwarding strategy into consideration, we attempt to recruit the most efficient users to forward the sensing data. The pseudo-code of PURE-DF is shown in Algorithm 2.

In PURE-DF recruitment strategy, the initial PAYG users recruited by PURE will recruit PAYG user and replicate the sensing data only if the encountered PAYG user's contact probability with the PAYM users is greater than any seen by the sensing data so far. As shown in Algorithm 2, when user U_{G_i} encounters user U_{G_j}, it will recruit user U_{G_j} and replicate the sensing data to user U_{G_j} if and only if $F_{U_{G_j}U_M}$ is higher than highest contact probability existing in the threshold H_i. Moreover, the method to recruit initial users is the same as that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

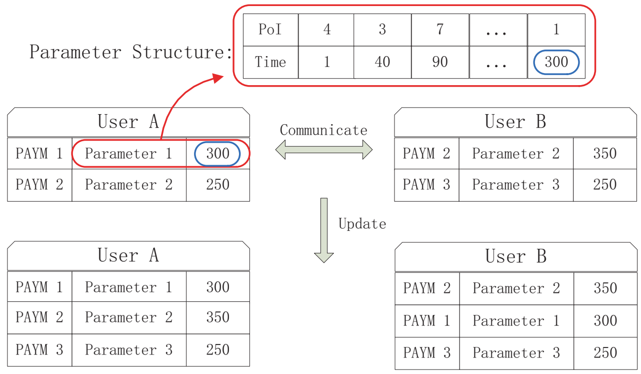


Fig. 6. An example where users A and B exchange and update the parameters of PAYM users according to the record time (i.e. 250, 300, and 350).

of PURE, so the running time is $O(ln)$. The other steps are the same as those of Delegation Forwarding, so the running time could be bounded by $O(\frac{5}{3}\sqrt{n})$ [21]. Therefore, it can be seen that Algorithm 2 can be implemented to run in $O(\frac{5}{3}ln^{\frac{3}{2}})$ time.

C. Parameters Collection

Each user k records its mobility history list: (L_n^k,T_n^k), when it enters the nth PoI, it will add the (L_n^k,T_n^k) to the tail of its list. After a period of time, each user could collect its mobility history list, which is shown in the parameter structure of Fig. 6. The last item of the mobility history list shows the newest state containing both the last PoI it enters and the updating time.

To predict user mobility in the PURE and PURE-DF algorithms, a pair of users in communication need to know each other's mobility history lists. More importantly, they also need to know the PAYM users' mobility history lists. For example, when a PAYG user A encounters a PAYG user B, they could exchange their mobility history lists, and they could calculate the transition probability matrix and the sojourn time probability distribution matrix. However, they could not know the PAYM users' mobility history lists. Therefore, they could not calculate the contact probability between user A and the PAYM users.

To easily solve this problem, regard the PAYM users as sources that will disseminate their mobility history lists to the network. The dissemination process is considered the warm-start time, after which they begin the PURE and PURE-DF algorithms to recruit users. However, there are still following two issues to be solved: (1) finding a dissemination strategy to make the PAYM users' mobility history lists cover all the users as soon as possible and (2) estimating the current position of the PAYM user to further calculate the contact probability of Eq. 13.

For issue (1), an easy solution is to exchange the PAYM users' mobility history lists through Epidemic dissemination strategy [24]. However, simulation results in Section IV-C show that, in order to cover all the users, we achieve an overlong warm-start time; this is not good enough. Therefore, in order to reduce the warm-start time, we propose a parameter dissemination strategy with PoI-assistant (shown in Fig. 5).

TABLE II
SIMULATION PARAMETERS

Parameter	Traces		
	roma/taxi	**epfl**	**geolife**
Simulation Time	800,850,900,950,1000		
Uploading Deadline	400,450,500,550,600		
Time Unit (s)	15	30	5
Number of PoIs	10	13	12
PoI Radius (m)	200	80	300
Number of Users	158	368	727
Number of PAYM Users	2∼6	4∼8	8∼12

Because users communicate with each other only in the PoI, we propose making the PAYM users' mobility history lists cover all the PoIs, instead of covering all the users. As shown in Fig. 5, all the PoIs are regarded as common users, they could also help disseminate the PAYM users' mobility history lists. The PAYM user disseminates its mobility history list to the encountered PoI, and every user in this PoI could save the parameter in its buffer and take it to another PoI. In this manner, the parameters could be disseminated more quickly and the warm-start time could be further reduced. Simulation results are shown in Section IV-C.

To ensure that all the PoIs have the newest mobility history lists of the PAYM users, we propose a parameter exchange and update strategy as shown in Fig. 6. The users including PAYG users, PAYM users, and PoIs, exchange and update the records of the PAYM users. For example, in Fig. 6, user A with records of PAYM1 and PAYM2 encounters user B with records of PAYM2 and PAYM3, they exchange their missing records (user A forwards the record of PAYM1 to user B and user B forwards the record of PAYM3 to user A), and they update their same record (PAYM2). A simple update action is implemented according to the record time (updating the record with the nearest record time). The record time is the time of the tail item in the parameter structure.

For issue (2), user A wants to calculate the contact probability to PAYM1. First, user A needs to estimate the current position of PAYM1, according to the mobility history list of PAYM1. As shown in Eq. 14, the current time is t, $T_{RecPAYM1}$ is the record time, and we know that the PoI of PAYM1 at $T_{RecPAYM1}$ is $PoI_{CurPAYM1}$. According to Eq. 8, we could achieve the probability that PAYM1 is currently in j is $Q^{PAYM1}_{PoI_{CurPAYM1}j}(t-T_{RecPAYM1})$. For every PoI j, we regard the PoI with the highest $Q^{PAYM1}_{PoI_{CurPAYM1}j}(t-T_{RecPAYM1})$ as the current position of PAYM1.

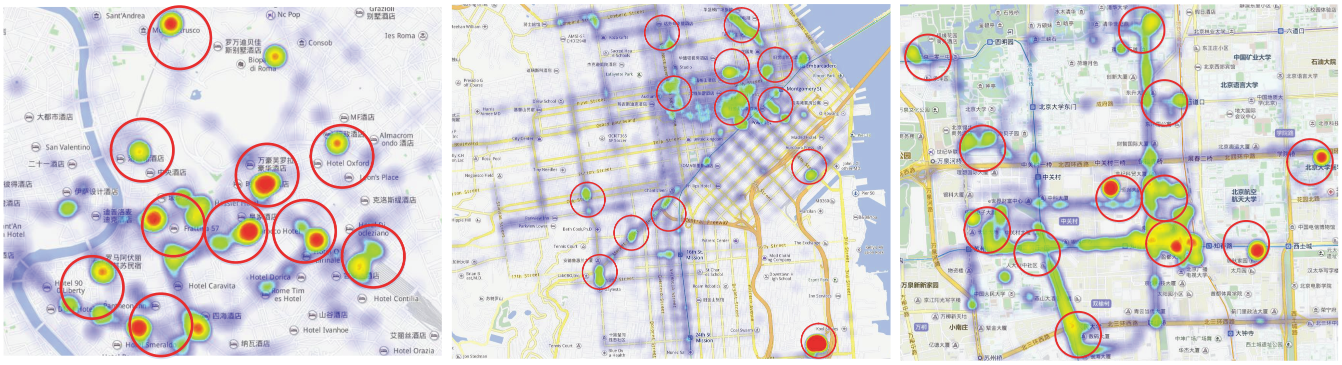
$$Q^{PAYM1}_{PoI_{CurPAYM1}j}(t-T_{RecPAYM1}) = P(L_t^{PAYM1}=j|L^{PAYM1}_{T_{RecPAYM1}}=PoI_{CurPAYM1}) \quad (14)$$

IV. PERFORMANCE EVALUATION

A. The Traces Used and Settings

We adopt three widely-used real-world traces, *roma/taxi trace set* [25], *epfl trace set* [26], and *geolife trace set* [27], [28] to test the performances of the proposed recruitment

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing



(a) roma/taxi (b) epfl (c) geolife

Fig. 7. The PoI areas in Baidu map of the three real-world data sets.

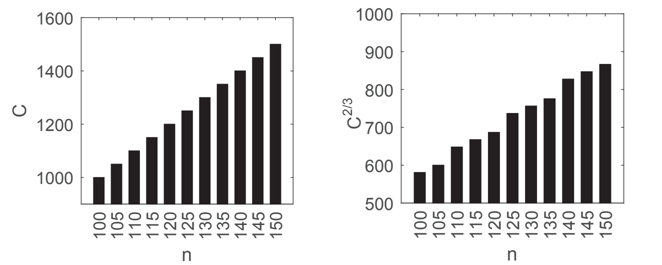
strategies. The *roma/taxi trace set* includes 320 taxi drivers that work in the center of Rome, Italy. The traces present the positions of drivers. Each taxi driver has a tablet that periodically retrieves a GPS position and sends it to a central server. The *epfl trace set* contains mobility traces of taxi cabs in San Francisco, USA. It contains GPS coordinates of approximately 500 taxis collected over 30 days in the San Francisco Bay Area. The *geolife trace set* contains 17,621 trajectories with a total distance of about 1.2 million kilometers and a total duration of 48,000+ hours. These trajectories were recorded by different GPS loggers and GPS phones.

We first address these data sets by filtering some abnormal user traces (discontinuous records or remote areas). Then, we put the traces into Baidu map according to the GPS. Through invoking the JavaScript API in Baidu map, we draw a thermodynamic diagram. We paint red to the area that is covered by more than 400 times and find the associate PoIs in each data set (Fig. 7). We divide all the users into PAYG users and PAYM users. We regard the initial PAYG users with sensing data as sources, and we consider the PAYM users destinations. The detailed simulation parameters in this network environment (simulation time, number of PoIs, number of PAYM users) are listed in Table II. In particular, 800, 850 in simulation time means the number of time slice, which is the period of collecting data. For the *roma/taxi* data set, the collecting period is 15s, for the *efpl* data set 30s, and for the *geolife* data set 5s.

B. Algorithms and Performances in Comparison

In order to validate the time complexity of Algorithms 1 and 2, in real-world trace *roma/taxi*, we test the relationship (as shown in Fig. 8) between the number of PAYG users (n) and the time complexity (C), which is achieved by counting the number of elementary operations performed by the algorithm (assignment, comparison and judgment operations). The results in Fig. 8 show that, the time complexity of PURE is *O(ln)*, and that of PURE-DF is $O(ln^{\frac{3}{2}})$. The experimental results match the theoretical results.

To demonstrate the performance of the proposed recruitment strategies, we carry out simulations focusing on the following



(a) Algorithm 1 PURE (b) Algorithm 2 PURE-DF

Fig. 8. Experimental results of the time complexity for Algorithms 1 and 2. The time complexity is defined as C. n is the total number of PAYG users.

two aspects: (1) payment in PURE and (2) delivery efficiency in PURE-DF.

For the first part, in order to test whether PURE could achieve a lower payment, we compare two recruitment strategies: PURE and RR. PURE, proposed in this paper, recruits one user with the highest contact probability with a PAYM user in every PoI. RR (Recruitment Randomly) randomly recruits a user in each PoI.

For the second part, we attempt to test whether PURE-DF could achieve the highest delivery efficiency compared to four other recruitment strategies: PURE, PURE-DF, EP [24], and SAW [29]. PURE has just been described, and in PURE-DF recruitment strategy, the initial PAYG users recruited by PURE will recruit another PAYG user and replicate the sensing data only if the encountered PAYG user's contact probability with PAYM users is greater than any seen by the sensing data so far. In EP (Epidemic) recruitment strategy, the initial PAYG users recruited by PURE will recruit every encounter and replicate the sensing data to it. In SAW (Spray and Wait) recruitment strategy, the initial PAYG users recruited by PURE will make a fixed number of sensing data copies (8 in the simulation), recruit every encounter, and spray half of its copies to each encounter until its copies reduce to 1. The encounters will undertake the same spray and wait action.

While a range of data is gathered from the simulations, we take the following five main performance metrics into consideration:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

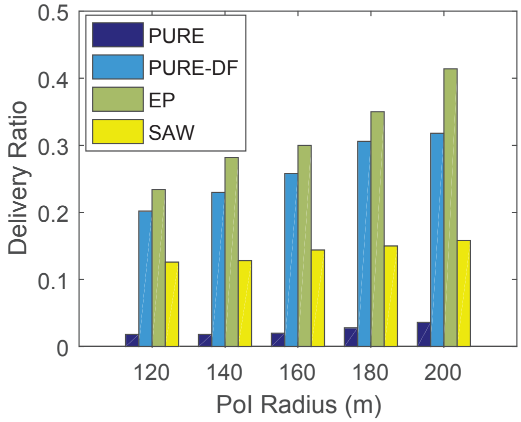
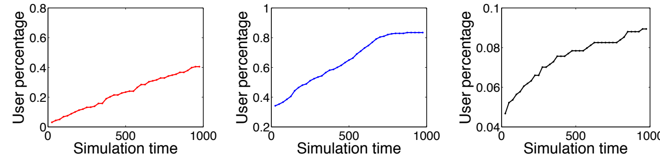
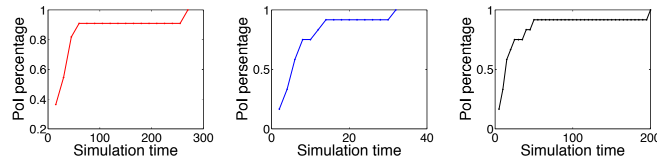


Fig. 9. Relationship between delivery ratio and PoI radius.



(a) *roma/taxi trace set* (b) *epfl trace set* (c) *geolife trace set*

Fig. 10. Parameters' dissemination results without PoI-assistant.



(a) *roma/taxi trace set* (b) *epfl trace set* (c) *geolife trace set*

Fig. 11. Parameters' dissemination results with PoI-assistant.

(1) Payment, which is the total cost for uploading the sensing data of all the PoIs.
(2) Delivery Ratio, which is the ratio between the sensing data number uploaded by PAYM users and the total sensing number data generated by all the PoIs.
(3) Delivery Cost, which is the number of recruiters that assist in delivering the sensing data.
(4) Average Delay, which is the average elapsed time of the successfully-delivered (uploaded by PAYM users) sensing data.
(5) Delivery Efficiency, which is the result of delivery ratio divided by delivery cost.

C. Simulation Results

1) PoI Radius: To determine the influence of different PoI sizes, in real-world trace *roma/taxi*, we first test the delivery ratio along with the growth of PoI radius. As shown in Fig. 9, there is an upward trend of delivery ratio along with the growth of PoI radius because a larger PoI radius leads to a better communication condition. Moreover, the ranking of delivery ratio performance is EP>PURE-DF>SAW>PURE, which proves that the recruitment strategy proposed in this paper could be efficiently and widely used in a PoI-based network environment.

2) Warm-start Time: To determine the warm-start time in the three different real-world data sets: *roma/taxi*, *epfl*, and *geolife*, we test the situation of parameter collection along with the change of simulation time.

As described in Section III-C, an easy solution to exchange the PAYM users' mobility history lists is using Epidemic dissemination strategy, which utilizes every possible connection to replicate the records in its buffer to every ever-encountered user. However, if we only use PAYG and PAYM users to exchange the parameters, the dissemination speed in this method is unsatisfactory; we need an overlong time to make the PAYM users' mobility history lists cover all the users. As shown in Fig. 10, we test the user percentage, defined as the infected user number divided by the total user number, along with the growth of simulation time. The simulation results show that, after 1,000 simulation time, the user percentages of the three data sets are 41%, 83%, and 9%, respectively. This means that we need an overlong warm-start time to get a 100% user percentage. Therefore, this dissemination strategy without PoI-assistant is useless.

In order to reduce the warm-start time, we propose a parameter dissemination strategy with PoI-assistant. This dissemination strategy regards the PoIs as the users.Because users could communicate with each other only in the PoI, we propose covering all the PoIs with the PAYM users' mobility history lists instead of covering all the users. Then, every pair of users that encounter one another could obtain the mobility history lists of the PAYM users via the current PoI. As shown in Fig. 11, the PoI percentage is defined as the infected PoI number (having the PAYM users' mobility history lists) divided by the total PoI number. With the growth of the simulation time, the PoI percentage quickly reaches 100%. For the different data sets to reach 100%, the simulation times in roma/taxi, epfl, and geolife are 270, 32, and 200, respectively. The simulation results show that the dissemination strategy with PoI-assistant is useful in reducing warm-start times. In the following simulations, the warm-start times for each of the data sets are designed as 300, 50, and 250, respectively.

3) Payment in PURE: To evaluate the performances of PURE and RR, we first conduct three groups of simulations using the *roma/taxi*, *epfl*, and *geolife* traces. PURE recruits a PAYG user with the highest contact probability with the PAYM users in each PoI, while RR randomly recruits a PAYG user in each PoI. Assuming that, the payment for sensing data uploaded via a PAYM user is 0, while the payment for uploading by a PAYG user is 1. The payments as a function of the number of PAYM users, sensing data uploading deadline (message TTLs), and simulation time are shown in Fig. 12-Fig. 14 for roma/taxi, epfl, and geolife, respectively.

It is not difficult to find that, PURE achieves a lower payment compared with RR. This is because the recruiters in PURE have a higher probability of contacting a PAYM user. Therefore, the sensing data uploaded by PAYG users is less than that to RR. As seen in Fig. 12, where the first part displays

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

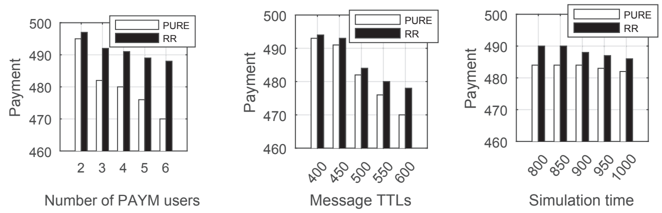


Fig. 12. Payment comparisons on the *roma/taxi trace set*.

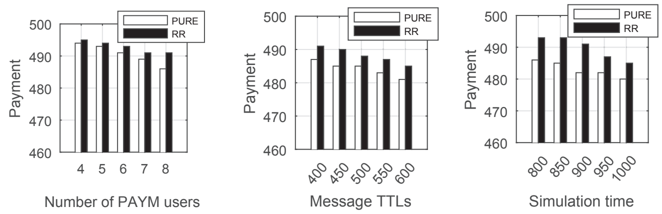


Fig. 13. Payment comparisons on the *epfl trace set*.

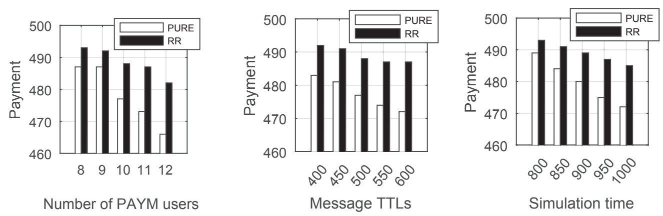


Fig. 14. Payment comparisons on the *geolife trace set*.

the variation of payments along with the growth of the number of PAYM users. The payments for PURE are lower than that of RR. It is worth noting that, the payment difference between PURE and RR becomes larger with the growth of the number of PAYM users. This is not strange because more PAYM users improve the precision of the prediction in PURE. Therefore, more sensing data can be delivered to PAYM users and can be uploaded by them freely.

Moreover, there is a downward trend of payment along with the increase of the number of PAYM users for RR and PURE. This is because more PAYM users could save more uploading payment. In the second part of Fig. 12, there is still a downward trend of payment along with the growth of message TTL because a larger message TTL leads to a higher delivery ratio to the PAYM users which, in turn, reduces the uploading payment. Similarly, in the third part of Fig. 12, a downward trend of payment also appears along with the growth of simulation time, because longer simulation time leads to a lower recruitment payment.

4) Delivery Efficiency in PURE-DF: In this section, we regard the PAYG users recruited by PURE as source users and treat the PAYM users as destination users. We attempt to recruit as few users as possible to achieve the best possible delivery performance from sources to destinations. A larger number of recruiters leads to a higher delivery ratio, but also to a higher recruitment cost. To evaluate the performance of PURE-DF, we test the delivery ratio, delivery cost, average delay, and delivery efficiency in three real-world data sets. The simulation results are shown in Fig. 15-Fig. 17.

As seen in Fig. 15, the ranking of delivery ratio performance is EP>PURE-DF>SAW>PURE, which is reasonable because EP recruits every encounter to assist in delivering the sensing data so the delivery ratio of EP is the highest. However, the delivery cost of EP is also the highest, which is also shown in Fig. 15. EP utilizes every replication opportunity to disseminate the sensing data, and therefore, the delivery cost is higher than in any other strategy. In addition, the delivery ratio of PURE-DF is higher than that of SAW because PURE-DF will recruit a PAYG user and replicate the sensing data only if the encountered PAYG user's contact probability with the PAYM users is greater than any seen by the sensing data so far. SAW just replicates the sensing data to the encountered PAYG user without considering the contact probability with the PAYM users. Therefore, the recruiters of PURE-DF have a higher contact probability with destinations than SAW does. As a result, the delivery ratio of PURE-DF is higher than that of SAW.

Most importantly, PURE-DF achieves the highest delivery efficiency of all four recruitment strategies. This means that each recruiter of PURE-DF can do the more average contribution to the delivery ratio of sensing data, compared with other recruitment strategies. In other words, PURE-DF recruits as few users as possible to achieve the best delivery performance. Even if EP achieves the highest delivery ratio, its delivery efficiency is the lowest among all the recruitment strategies because EP recruits too many users thereby increasing its recruitment cost. Fig. 16 and Fig. 17 show performance comparisons on the *epfl* and *geolife* trace sets, respectively.

V. Related Work

Most of mobile crowdsensing works [30] focus on the following two aspects: (1) how to stimulate users to participate in a crowdsensing task and (2) which users should be recruited to finish the crowdsensing task.

A. Incentive Mechanisms

Game theoretic model and auction-based mechanism have been widely used in designing incentive mechanisms for mobile crowdsensing systems. Li *et al.* [31] propose a private incentive mechanism that protects the privacy of each user's bid against the other honest-but-curious users. Zhang *et al.* [32] propose incentivizing a number of workers to label a set of binary tasks under strict budget constraint. Yang *et al.* [33] present an incentive mechanism through a Stackelberg game, where the crowdsourcer is the leader and the users are the followers. Jin *et al.* [34] propose INCEPTION, a novel mobile crowdsensing system framework that integrates an incentive, a data aggregation, and a data perturbation mechanism. Wen *et al.* [35] present an incentive mechanism according to a quality-driven auction aimed at the mobile crowdsensing system. The worker is paid based on the quality of sensed data instead of working time, as adopted in the literature. Wu *et al.* [36] propose a quality-of-video oriented pricing incentive scheme,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

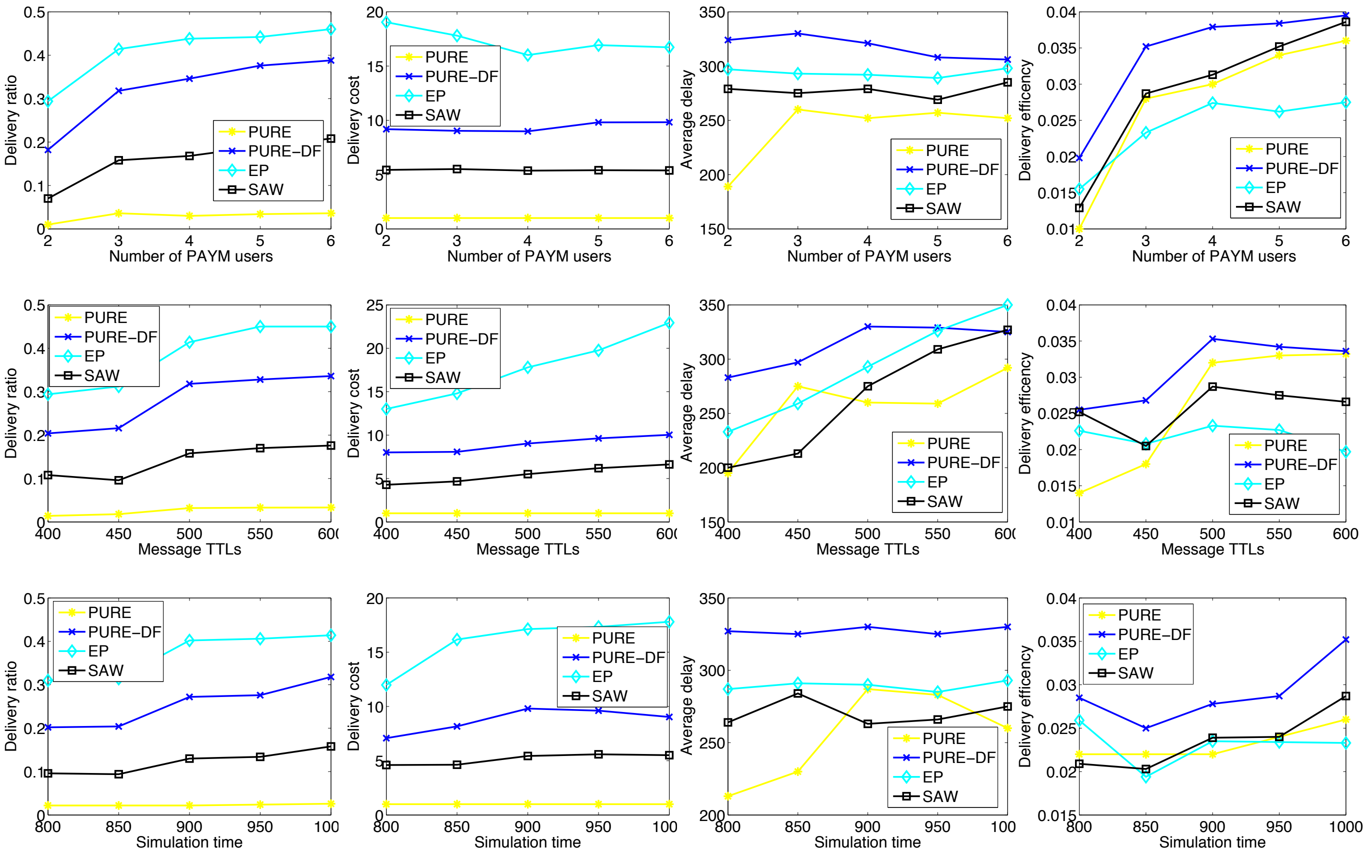


Fig. 15. Performance comparisons on the *roma/taxi trace set*: delivery ratio & delivery cost & average delay & delivery efficiency.

namely Vbargain, to stimulate mobile users to deliver video data collaboratively.

The above research focus on stimulating the users to participate in a crowdsensing task. These could be regarded as the preliminary works of this paper, and are an important part in mobile crowdsensing.

B. Recruitment Strategy

There is also plenty of research that focuses on user recruitment strategy. Xiong *et al.* [37] recruit some users in each sensing period for piggyback crowdsensing task participation, so that the resulting solution achieves near-maximal coverage quality without exceeding incentive budget. Xiao *et al.* [38] first formulate the deadline-sensitive user recruitment problem as an NP-hard problem. Then, they propose a greedy algorithm to solve this problem. Wang *et al.* [20] propose the design and implementation of a mobile crowdsensing data uploading mechanism (ecoSense) to help reduce additional 3G data costs incurred by the whole crowd of sensing participants. He *et al.* [39] present a new participant recruitment strategy for vehicle-based crowdsourcing, according to the predicted vehicle trajectory. This strategy guarantees that the system can perform well through the currently recruited participants for a period of future time. Xiao *et al.* [40] propose an oFfline Task Assignment (FTA) algorithm and an oNline Task Assignment (NTA) algorithm to assign tasks in mobile crowdsensing, according to the mobility model of users in mobile social networks. Pu *et al.* [41] formulate an online worker recruitment problem to maximize the expected sum of service quality. Hien To *et al.* [42] present a framework for crowdsourcing hyper-local information, where only the workers who have already been within the spatiotemporal vicinity of a task are eligible candidates to report data. Li *et al.* [43] make use of mobile crowdsourced data obtained from location-based social network services to study influence maximization in location-based social networks. Karaliopoulos *et al.* [44] decide which mobile users to select in order to generate the required space-time paths across the network for collecting data from a set of fixed locations.

The above works focus on proposing a user recruitment strategy to efficiently finish the crowdsensing task. However, almost all the works utilize the WiFi APs to upload the sensing data and ignore freely uploading the data through mobile users.

VI. CONCLUSION

We have looked into the problem of user recruitment in mobile crowdsensing campaigns drawing on opportunistic networking methods. First, we divide the users in the network into PAYG (uploading costly) users and PAYM users (uploading freely), and we formalize this problem as recruiting the user of the highest contact probability with PAYM users. Then, according to the semi-Markov model, we propose an efficient

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

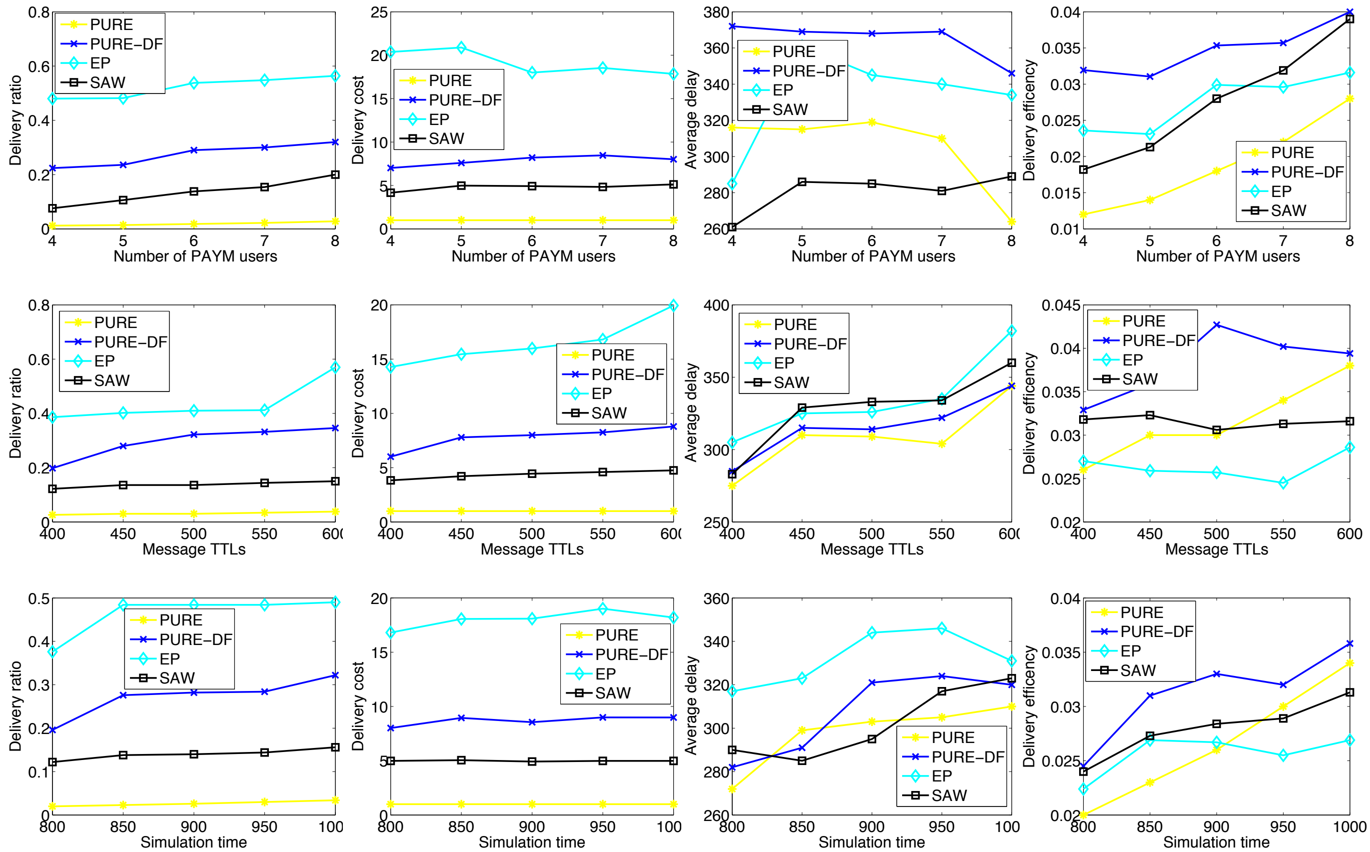


Fig. 16. Performance comparisons on the *epfl trace set*: delivery ratio & delivery cost & average delay & delivery efficiency.

Prediction-based User Recruitment for mobile crowdsEnsing (PURE), where the PAYG user's contact probability with destinations is achieved and where multiple users can be recruited to cooperatively perform a common task, ensuring that the expected data-uploading cost is minimal. Moreover, we propose PURE-DF by extending PURE to a case in which we address the tradeoff between the delivery ratio of sensing data and the recruiter number according to the thought of Delegation Forwarding. We conduct extensive simulations based on three widely-used real-world traces: roma/taxi, epfl and geolife. The results show that compared with other recruitment strategies, PURE achieves a lower recruitment payment and PURE-DF achieves the highest delivery efficiency.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under Grant No. 61272412, and by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 20120061110044. This work is also supported in part by NSF grants CNS 1449860, CNS 1461932, CNS 1460971, CNS 1439672, CNS 1301774, and ECCS 1231461.

REFERENCES

[1] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.

[2] Y. Gao, W. Dong, K. Guo, X. Liu, Y. Chen, X. Liu, J. Bu, and C. Chen, "Mosaic: A low-cost mobile sensing system for urban air quality monitoring," in *Proc. of IEEE INFOCOM 2016*.

[3] Y. Hu, G. Dai, J. Fan, Y. Wu, and H. Zhang, "Blueaer: A fine-grained urban pm2.5 3d monitoring system using mobile sensing," in *Proc. of IEEE INFOCOM 2016*.

[4] R. Du, C. Fischione, and M. Xiao, "Flowing with the water: On optimal monitoring of water distribution networks by mobile sensors," in *Proc. of IEEE INFOCOM 2016*.

[5] E. Koukoumidis, L.-S. Peh, and M. Martonosi, "Signalguru: leveraging mobile phones for collaborative traffic signal schedule advisory," in *Proc. of ACM MobiSys 2011*.

[6] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, "A Survey of Mobile Phone Sensing," *IEEE Communications Magazine*, vol. 48, no. 9, pp. 140–150, 2010.

[7] J. Wan, J. Liu, Z. Shao, A. V. Vasilakos, M. Imran, and K. Zhou, "Mobile crowd sensing for traffic prediction in internet of vehicles," *Sensors*, vol. 16, no. 1, 2016.

[8] C. Liu, L. Zhang, Z. Liu, K. Liu, X. Li, and Y. Liu, "Lasagna: Towards deep hierarchical understanding and searching over mobile sensing data," in *Proc. of ACM MobiCom 2016*.

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

[10] R. Gao, M. Zhao, T. Ye, F. Ye, and Y. Wang, "Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing," in *Proc. of ACM MobiCom 2014*.

[11] D. E. Difallah, M. Catasta, and G. Demartini, "The dynamics of micro-task crowdsourcing: The case of amazon mturk," in *Proc. of ACM WWW 2015*.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2017.2702613, IEEE Transactions on Mobile Computing

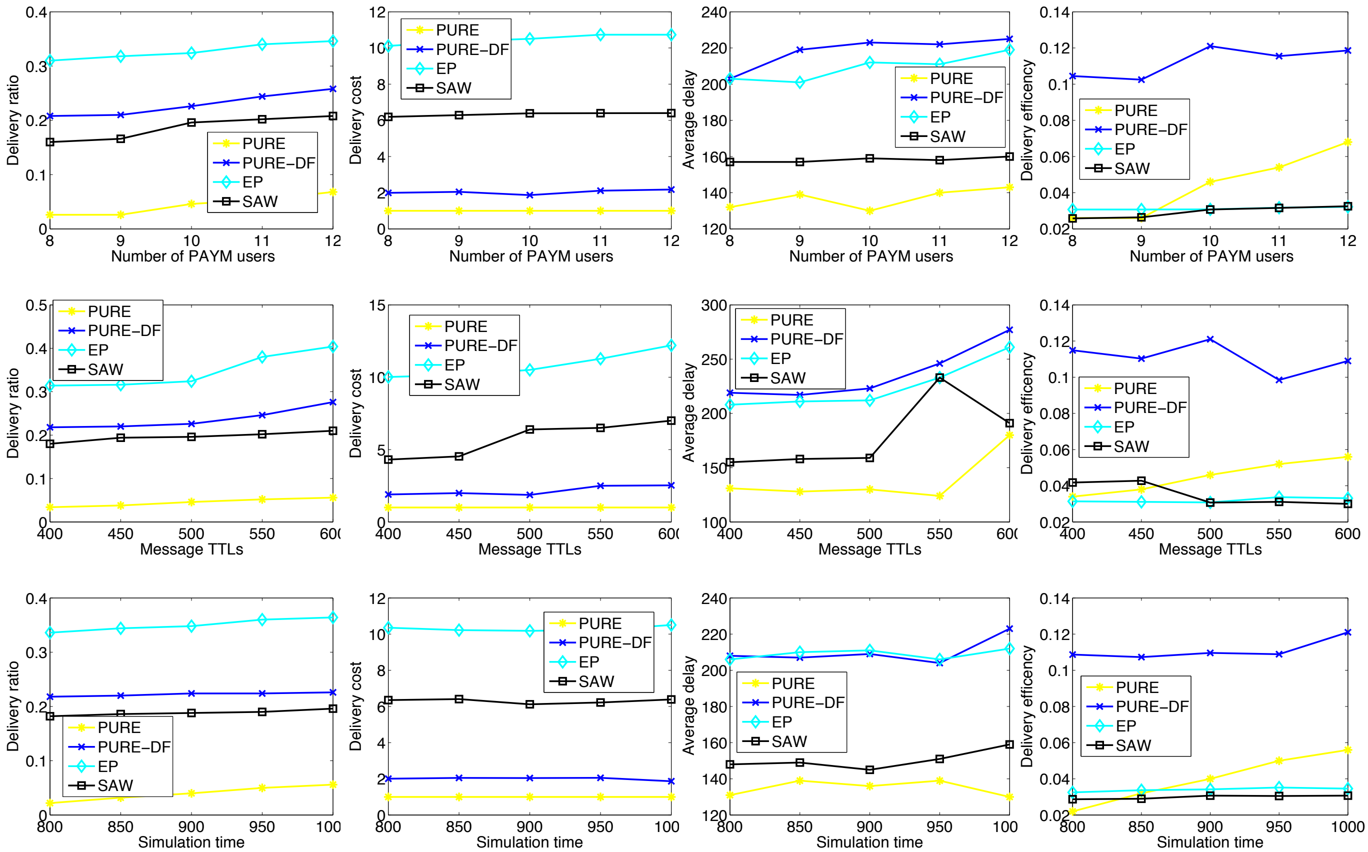


Fig. 17. Performance comparisons on the *geolife trace set*: delivery ratio & delivery cost & average delay & delivery efficiency.

[12] D. Zhang, H. Xiong, LeyeWang, and G. Chen, “Crowdrecruiter: Selecting participants for piggyback crowdsensing under probabilistic coverage constraint,” in *Proc. of ACM UbiComp 2014*.

[13] H. Xiong, D. Zhang, G. Chen, L. Wang, V. Gauthier, and L. E. Barnes, “iCrowd: Near-Optimal Task Allocation for Piggyback Crowdsensing,” *IEEE Transactions on Mobile Computing*, vol. 15, no. 8, pp. 2010–2022, 2016.

[14] B. Guo, C. Chen, D. Zhang, and Z. Yu, “Mobile crowd sensing and computing: when participatory sensing meets participatory social media,” *IEEE Communications Magazine*, vol. 54, no. 2, pp. 131–137, 2016.

[15] A. Singla and A. Krause, “Truthful incentives in crowdsourcing tasks using regret minimization mechanisms,” in *Proc. of ACM WWW 2013*.

[16] X. Gong, X. Chen, J. Zhang, and H. V. Poor, “Exploiting Social Trust Assisted Reciprocity (STAR) Toward Utility-Optimal Socially-Aware Crowdsensing,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 1, no. 3, pp. 195–208, 2015.

[17] S. Reddy, D. Estrin, M. Hansen, and M. Srivastavai, “Examining micro-payments for participatory sensing data collections,” in *Proc. of ACM UbiComp 2010*.

[18] P. Loiseau, G. Schwartz, J. Musacchio, S. Amin, and S. S. Sastry, “Incentive Mechanisms for Internet Congestion Management: Fixed-Budget Rebate versus Time-of-Day Pricing,” *IEEE/ACM Transactions on Networking*, vol. 22, no. 2, pp. 647–661, 2013.

[19] D. Yang, G. Xue, G. Fang, and J. Tang, “Incentive mechanisms for crowdsensing: Crowdsourcing with smartphones,” *IEEE/ACM Transactions on Networking*, pp. 1–13, 2015.

[20] L. Wang, D. Zhang, H. Xiong, J. P. Gibson, C. Chen, and B. Xie, “ecoSense: Minimize Participants Total 3G Data Cost in Mobile Crowdsensing Using Opportunistic Relays,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–14, 2016.

[21] V. Erramilli, M. Crovella, A. Chaintreau, and C. Diot, “Delegation forwarding,” in *Proc. of ACM MobiHoc 2008*.

[22] Q. Yuan, I. Cardei, and J. Wu, “An Efficient Prediction-Based Routing in Disruption-Tolerant Networks,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 1, pp. 19–31, 2012.

[23] E. Wang, Y. Yang, and L. Li, “A Clustering Routing Method Based on Semi-Markov Process and Path-Finding Strategy in DTN,” *Chinese Journal of Computers*, vol. 38, no. 3, pp. 483–499, 2015.

[24] A. Vahdat and D. Becker, “Epidemic Routing for Partially-Connected Ad Hoc Networks,” Duke University, Tech. Rep.

[25] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, and A. Rabuffi, “CRAWDAD dataset roma/taxi (v. 2014-07-17),” Downloaded from http://crawdad.org/roma/taxi/20140717, 2014.

[26] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, “CRAWDAD dataset epfl/mobility (v. 2009-02-24),” Downloaded from http://crawdad.org/epfl/mobility/20090224, Feb. 2009.

[27] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma., “Mining interesting locations and travel sequences from gps trajectories,” in *Proc. of ACM WWW 2009*.

[28] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma., “Understanding mobility based on gps data.” in *Proc. of ACM UbiComp 2008*.

[29] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, “Spray and wait: An efficient routing scheme for intermittently connected mobile networks,” in *Proc. of ACM SIGCOMM 2005*.

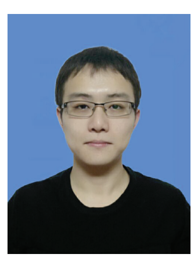
[30] Z. Yu, H. Xu, Z. Yang, and B. Guo, "Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 151–158, 2016.
[31] H. Jin, L. Su, B. Ding, K. Nahrstedt, and N. Borisov, "Enabling privacy-preserving incentives for mobile crowd sensing systems," in *Proc. of IEEE ICDCS 2016*.
[32] Q. Zhang, Y. Wen, X. Tian, X. Gan, and X. Wang, "Incentivize crowd labeling under budget constraint," in *Proc. of IEEE INFOCOM 2015*.
[33] D. Yang, G. Xue, X. Fang, and J. Tang, "Incentive Mechanisms for Crowdsensing: Crowdsourcing With Smartphones," *IEEE/ACM Transactions on Networking*, pp. 1–13, 2015.
[34] H. Jin, L. Su, H. Xiao, and K. Nahrstedt, "Inception: Incentivizing privacy-preserving data aggregation for mobile crowd sensing systems," in *Proc. of ACM MobiHoc 2016*.
[35] Y. Wen, J. Shi, Q. Zhang, X. Tian, Z. Huang, H. Yu, Y. Cheng, and X. Shen, "Quality-Driven Auction-Based Incentive Mechanism for Mobile Crowd Sensing," *IEEE Transactions on Vehicular Technoloty*, vol. 64, no. 9, pp. 4203–4214, 2015.
[36] H. Wu, L. Liu, X. Zhang, and H. Ma, "Quality of video oriented pricing incentive for mobile video offloading," in *Proc. of IEEE INFOCOM 2016*.
[37] H. Xiong, D. Zhang, G. Chen, L. Wang, and V. Gauthier, "Crowdtasker: Maximizing coverage quality in piggyback crowdsensing under budget constraint," in *Proc. of IEEE Percom 2015*.
[38] M. Xiao, J. Wu, H. Huang, L. Huang, and C. Hu, "Deadline-sensitive user recruitment for mobile crowdsensing with probabilistic collaboration," in *Proc. of IEEE ICDCS 2016*.
[39] Z. He, J. Cao, and X. Liu, "High quality participant recruitment in vehicle-based crowdsourcing using predictable mobility," in *Proc. of IEEE INFOCOM 2015*.
[40] M. Xiao, J. Wu, L. Huang, Y. Wang, and C. Liu, "Multi-task assignment for crowdsensing in mobile social networks," in *Proc. of IEEE INFOCOM 2015*.
[41] L. Pu, X. Chen, J. Xu, and X. Fu, "Crowdlet: Optimal worker recruitment for self-organized mobile crowdsourcing," in *Proc. of IEEE INFOCOM 2016*.
[42] H. To, L. Fan, L. Tran, and C. Shahabi, "Real-time task assignment in hyper-local spatial crowdsourcing under budget constraints," in *Proc. of IEEE Percom 2016*.
[43] J. Li, Z. Cai, M. Yan, and Y. Li, "Using crowdsourced data in location-based social networks to explore influence maximization," in *Proc. of IEEE INFOCOM 2016*.
[44] M. Karaliopoulos, O. Telelis, and I. Koutsopoulos, "User recruitment for mobile crowdsensing over opportunistic networks," in *Proc. of IEEE INFOCOM 2015*.

En Wang received his B.E. degree in software engineering from Jilin University, Changchun, in 2011, his M.E. degree in computer science and technology from Jilin University, Changchun, in 2013, and his Ph.D. in computer science and technology from Jilin University, Changchun, in 2016. He is currently a lecturer in the Department of Computer Science and Technology at Jilin University, Changchun. He is also a visiting scholar in the Department of Computer and Information Sciences at Temple University in Philadelphia. His current research focuses on the efficient utilization of network resources, scheduling and drop strategy in terms of buffer-management, energy-efficient communication between human-carried devices, and mobile crowdsensing.

Yongjian Yang received his B.E. degree in automatization from Jilin University of Technology, Changchun, Jilin, China, in 1983; his M.E. degree in computer communication from Beijing University of Post and Telecommunications, Beijing, China, in 1991; and his Ph.D. in software and theory of computer from Jilin University, Changchun, Jilin, China, in 2005. He is currently a professor and a PhD supervisor at Jilin University, the Vice Dean of the Software College of Jilin University, Director of Key lab under the Ministry of Information Industry, Standing Director of the Communication Academy, and a member of the Computer Science Academy of Jilin Province. His research interests include: network intelligence management, wireless mobile communication and services, and wireless mobile communication.

Jie Wu is the Associate Vice Provost for International Affairs at Temple University. He also serves as Director of the Center for Networked Computing and Laura H. Carnell professor in the Department of Computer and Information Sciences. Prior to joining Temple University, he was a program director at the National Science Foundation and was a distinguished professor at Florida Atlantic University. His current research interests include mobile computing and wireless networks, routing protocols, cloud and green computing, network trust and security, and social network applications. Dr. Wu regularly publishes in scholarly journals, conference proceedings, and books. He serves on several editorial boards, including IEEE Transactions on Service Computing and the Journal of Parallel and Distributed Computing. Dr. Wu was general cochair/chair for IEEE MASS 2006, IEEE IPDPS 2008, IEEE ICDCS 2013, and ACM MobiHoc 2014, as well as program co-chair for IEEE INFOCOM 2011 and CCF CNCC 2013. He was an IEEE Computer Society Distinguished Visitor, ACM Distinguished Speaker, and chair for the IEEE Technical Committee on Distributed Processing (TCDP). Dr. Wu is a CCF Distinguished Speaker and a Fellow of the IEEE. He is the recipient of the 2011 China Computer Federation (CCF) Overseas Outstanding Achievement Award.

Wenbin Liu received his B.S. degree in physics from Jilin University, Changchun, Jilin, China, in 2012. He is currently a Ph.D. candidate in the Department of Computer Science and Technology, Jilin University, Changchun, Jilin. His current research focuses on communications in body area networks, delay tolerant networks, and crowdsensing.



Xingbo Wang received his B.E.degree in software engineering from Jilin University,Changchun, Jilin, China, in 2015. He is currently a postgraduate student in software engineering from Jilin University, Changchun, Jilin, China. His current research focuses on crowdsensing and delay tolerant networks.