Nearly Optimal Probabilistic Coverage for Roadside Data Dissemination in Urban VANETs

Yawei Hu[1,2], Mingjun Xiao[1,2](✉), An Liu[3], Ruhong Cheng[1,2], and Hualin Mao[1,2]

[1] School of Computer Science and Technology, University of Science and Technology of China, Hefei, People's Republic of China
xiaomj@ustc.edu.cn
[2] Suzhou Institute for Advanced Study, University of Science and Technology of China, Suzhou, People's Republic of China
[3] School of Computer Science and Technology, Soochow University, Suzhou, People's Republic of China

Abstract. Data disseminations based on Roadside Access Points (RAPs) in vehicular ad-hoc networks attract lots of attentions and have a promising prospect. In this paper, we focus on a roadside data dissemination, including three basic elements: RAP Service Provider (RSP), mobile vehicles and requesters. The RSP has deployed many RAPs at different locations in a city. A requester wants to rent some RAPs, which can disseminate their data to vehicles with some probabilities. Then, it tries to select the minimal number of RAPs to finish the data dissemination, in order to save the expenses. Meanwhile, the selected RAPs need to ensure that the probability of each vehicle receiving data successfully is no less than a threshold. We prove that this RAP selection problem is NP-hard, since it's a meaningful extension of the classic Set Cover problem. To solve this problem, we propose a greedy algorithm and give its approximation ratio. Moreover, we conduct extensive simulations on real world data to prove its good performance.

Keywords: Vehicular ad-hoc network · Data dissemination · Roadside access points selection

1 Introduction

Vehicular ad-hoc networks (VANETs) as a new paradigm of mobile ad-hoc networks, can offer convenient data service for passengers and have attracted attentions of many researchers. Further, with the development of intelligent transportation system, some cities have deployed many RAPs at different locations, such as bus stations, taxi pickup points and intersections. These RAPs usually have good capacities of storage and communication. When a vehicle enters a RAP's communication range, the RAP can deliver the information stored in its

H. Gao et al. (Eds.): DASFAA 2016 Workshops, LNCS 9645, pp. 238–253, 2016.
DOI: 10.1007/978-3-319-32055-7_20

memory to the vehicle via WiFi or other protocols of short-range wireless communication. Obviously, the introduction of RAPs can improve the connectivity of VANETs. Moreover, by using these RAPs in urban areas, some large-size data can be disseminated to passengers in vehicles at a low cost. Due to this advantage, much research ([1–3]) has studied the new form of data dissemination based on RAPs in VANETs.

Consider a scenario of roadside data dissemination. There is a RAP Service Provider (RSP), for instance, certain government agency, who has deployed many RAPs at different locations in a city. If a requester hopes to disseminate its data, it can rent parts of the RAPs from the RSP to finish the dissemination. For example, a shop can be a requester, which wants to disseminate advertisements of its new commodities. On the one hand, they need to select as less RAPs as possible for this requester to conduct the data dissemination, in order to save the requester's expenses. On the other hand, they also need to let these RAPs cover adequate vehicles, so that the data can be received by as many passengers as possible. Moreover, as the communication connections between RAPs and mobile vehicles are unstable and intermittent, the selected RAPs also need to ensure that the probability of each vehicle receiving the data successfully is no less than a reasonable threshold.

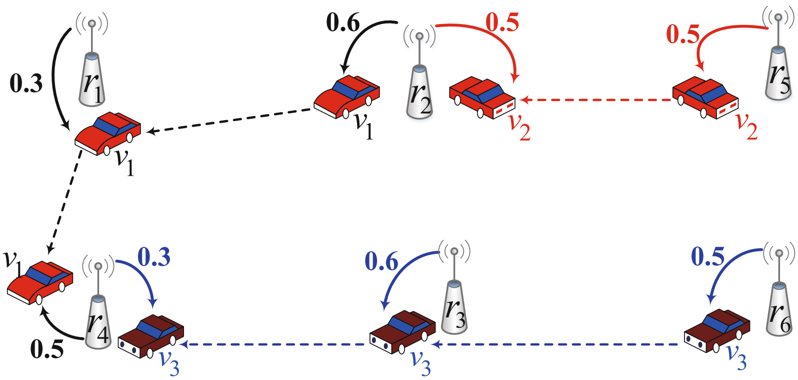


Fig. 1. A example of the roadside data dissemination. (Three vehicles pass six RAPs, during which the vehicles can receive data with some probabilities.)

Let's take Fig. 1 as an example. There are 6 RAPs ($r_1, \cdots, r_6$) and 3 mobile vehicles (v_1, v_2, v_3). Suppose that a requester hopes to disseminate its data to the three vehicles, and expects that the probability of each vehicle receiving the data successfully is not less than the threshold 0.7. In this example, we assume that the requester selects $\{r_2, r_3, r_4, r_5\}$ to perform the data dissemination. Although each vehicle cannot receive the data successfully from each single RAP with the probability no less than 0.7, their joint probability exceeds the threshold. For instance, v_1's probabilities of receiving the data successfully from r_2 and r_4 are 0.6 and 0.5, respectively. Their joint probability is $1-(1-0.6)*(1-0.5)=0.8$, beyond the threshold 0.7.

For the schemes of data dissemination in VANETs, most of them pay attentions to the delivery ratio or delay of messages. However, our roadside data dissemination focuses on how to select the minimal number of RAPs to conduct the dissemination. Unlike other schemes of data dissemination, our scheme can ensure that the probability of each vehicle receiving data successfully is no less than a predefined threshold. For example, in [2], the authors select several seed vehicles to disseminate advertisement data to other vehicles, and the seed vehicle selection is based on the degree centrality [4] of each vehicle. Nevertheless, the solution cannot guarantee that each vehicle's probability of receiving advertisement data successfully is large enough. From another perspective, the data dissemination can be regarded as the problem of coverage. For example, in [2], the dissemination is achieved by selecting some seed vehicles to cover as many vehicles as possible. Their "cover" means that there is at least one social contact between a single seed vehicle and each vehicle. In contrast, our problem is to ensure the joint dissemination probability from all selected RAPs no less than the threshold. Actually, our problem will lead to a combining probabilistic set cover mixed by non-linear programming, which is also different from the classic Set Cover problem.

In this paper, we first describe the moving pattern of a mobile vehicle, by using a set of RAPs which this vehicle often passes by. The mobile vehicles can receive data successfully from these passed RAPs with some probabilities. Then, we define the RAP selection problem for roadside data dissemination and prove its NP-hardness. The problem is to select the minimal number of RAPs to cover all of the vehicles, while ensuring that each vehicle can receive data successfully from the RAPs with probability no less than a threshold. Finally, we propose a greedy algorithm to solve the RAP selection problem, analyze the complexity and approximation ratio of this algorithm, and conduct extensive simulations to prove its good performance.

We highlight our main contributions as follows:

1. We propose a scheme for roadside data dissemination based on probabilistic coverage of some selected RAPs.
2. We introduce a new optimization problem, which is the RAP selection problem for roadside data dissemination. We prove that this problem is NP-hard.
3. We propose a greedy approximation algorithm to solve the RAP selection problem, give the corresponding approximation ratio of this algorithm and conduct extensive simulations to prove its good performance.

The remainder of this paper is organized as follows. Section 2 presents the related work. In Sect. 3, the network model, the definition of RAP selection problem and the proof of the NP-hardness of this problem are described. We propose a greedy algorithm, analyze its complexity and approximation ratio in Sect. 4. In Sect. 5, the evaluation of this algorithm is showed, followed by the conclusion of this paper in Sect. 6. Partially complex proofs are moved to the Appendix.

2 Related Work

The scenario of data dissemination based on RAPs in VANETs is different from data dissemination in Delay Tolerant Networks or Mobile Social Networks [5–7]. In VANETs, J. Qin et al. [2] considered to select seed vehicles to diffuse advertisement data to others. They analyzed the sociality of the vehicular network and proved the dynamic and temporal correlations of sociality. Based on the analysis, they proposed a greedy scheme to solve it. Z. Li et al. [3] proposed a scheme for advertisement data diffusion with an incentive-centered architecture. The architecture encourages the advertisement providers to trade off the effect and cost of their advertisements messages, aiming to avoid unnecessary distractions to drivers and message storms in VANETs. In [1], H. Zheng et al. designed a system, which can disseminate advertisement data via some placed RAPs. When a driver receives a shop's advertisements from RAPs, he or she may detour to the shop and the detour probability depends on the detour distance. The RAP placement needs to balance the tradeoff between the traffic density and the detour probability. S.-B. Lee et al. [8] proposed a secure incentive framework to avoid the noncooperative behavior of selfish or malicious vehicle nodes in the advertisement data dissemination. Unlike these schemes, our paper uses the RAPs to disseminate data to the vehicles passing by.

Our roadside data dissemination is relevant to the data dissemination based on roadside units (RSUs) in VANETs. For example, in [9], J. Jeong et al. considered to forward data from stationary APs to moving vehicles. The forwarding scheme took into account the AP's location and the destination of the vehicle's trajectory, and selected a target point as packet-and-vehicle-rendezvous-point. M. Sardari et al. in [10] proposed a message dissemination paradigm, in which each RSU encoded a huge message into k data packets and forwarded them to vehicles, then the vehicles can decode a specific RSU's message by collecting sufficient packets. K. Liu et al. [11] focused on accessing information stored at RSUs and paid attention to the channel division. Comparing with these papers, we consider how to select the minimal RAPs to finish the data dissemination from RAPs to vehicles.

Our problem, i.e., the RAP selection, is related to the deployment of RSUs in VANETs. In [12], T. Wu et al. studied the RSU placement problem for vehicular networks in a highway-like scenario. Their placement strategy maximized the aggregate throughput in the network, taking into account the impact of wireless interference, vehicle population distribution, and vehicle speeds in the formulation. In [13], B. Aslam et al. focused on the placement problem in the scenario of urban vehicular network environment. They proposed two methods aiming to minimize the reporting time of event for RSUs, with incorporating the density and speed of vehicle, as well as the occurrence likelihood of an event in urban. By comparison, our problem is based on the joint probabilistic coverage of RAPs.

Therefore, our roadside data dissemination is different from other studies of advertisement or data dissemination, and cannot be solved directly by existing solutions.

3 Network Model and Problem Definition

In this section, we set up the network model, define the RAP selection problem for roadside data dissemination and analyze the hardness of this problem.

3.1 Network Model

We first describe the set of i vehicles, that is, $V=\{v_1, v_2, \cdots, v_i\}$. We assume that the vehicles have been equipped with wireless communication devices. Consequently, they can communicate with the RAPs. Then, we use the set $R=\{r_1, r_2, \cdots, r_j\}$ to denote all the RAPs which are deployed by the RSP at different locations in a city. As we know, a vehicle often visits some locations frequently due to the sociality of the vehicle (actually, the driver or the passenger). For simplicity, we describe the *moving pattern* of each vehicle with a set of locations where the vehicle passes and the RAPs are placed. Namely, we use a subset of R, i.e., $R(v_i)=\{r_{i_1}, r_{i_2}, \cdots, r_{i_k}, \cdots\}$ to describe the move of the vehicle v_i. Next, for $v_i \in V$ and $r_j \in R(v_i)$, we use $p_j^i (0<p_j^i<1)$ to indicate the probability that v_i successfully receives data from r_j. Although a vehicle v_i may pass by certain RAP twice or more, we only consider the vehicle passing by each RAP in $R(v_i)$ once. This assumption is reasonable. If v_i passes by r_j n times, we can use the $1-(1-p_j^i)^n$ to replace the p_j^i, and meanwhile, treat it as one-time passing. In addition, we assume that the value of p_j^i can be derived from the history records.

We construct a graph G to describe the model. The vertex set of G is $V \cup R$. For an arbitrary pair of v_i and r_j, if $r_j \in R(v_i)$, we add an edge (v_i, r_j) into the edge set of G, and attach p_j^i as the weight of (v_i, r_j). Let's take Fig. 2(a) as an example. The vehicle node set is $V=\{v_1, v_2, \cdots, v_5\}$, the RAP node set is $R=\{r_1, r_2, r_3, r_4\}$. v_1 passes by the RAPs in $R(v_1)=\{r_1, r_2\}$, v_2 passes by the RAPs in $R(v_2)=\{r_2, r_3\}$, and so on. By the way, we denote $V(\cdot)$ as the set of neighbor vehicle nodes of one RAP, denote $R(\cdot)$ as the set of neighbor RAP nodes of one vehicle, denote $deg(\cdot)$ as the degree of a node. In Fig. 2(a), $V(r_1)=\{v_1, v_3\}$, $R(v_3)=\{r_1, r_3, r_4\}$, $deg(v_1)=2$.

3.2 Problem Definition

In our scheme of roadside data dissemination, a requester hopes that the vehicles in V can successfully receive its data from the RAPs with probability. Therefore, it's necessary to select as less RAPs as possible to conduct the data dissemination to these vehicles, and ensure that each vehicle's probability of receiving data successfully is not less than a threshold. We use τ to denote this threshold $(0<\tau<1)$, use S to denote the set of all selected RAPs from R. Moreover, for a vehicle v_i, it may receive data from each RAP in $S \cap R(v_i)$, and we denote $S(v_i)= S \cap R(v_i)$. As a result, given a selected RAP set S, v_i's probability of receiving data successfully from the RAPs in $S(v_i)$ is $Pr(v_i|S)=1-\prod_{r_j \in S(v_i)}(1-p_j^i)$. Especially, if $S(v_i)=\emptyset$, $Pr(v_i|S)=0$.

In this scheme, the problem we need to solve is how to select minimal number of RAPs and ensure that each vehicle's probability of receiving data successfully is not less than the threshold τ. In summary, we illustrate this problem of RAP selection as follow.

$$\begin{aligned} minimize &: |S| \\ subject\ to &: \bigcup_{r_j \in S} V(r_j) = V; \\ & Pr(v_i|S) \geq \tau\ for\ \forall v_i \in V; \\ & Pr(v_i|S) = 1 - \prod_{r_j \in S(v_i)} (1 - p_j^i); \\ & S \subseteq R. \end{aligned} \tag{1}$$

Here, we assume that $Pr(v_i|R) \geq \tau$ for $\forall v_i \in V$. If there is a vehicle v_k satisfying $Pr(v_k|R) < \tau$, the v_k cannot receive data successfully with the probability no less than τ, even if all RAPs in R are selected. Actually, there is no feasible solution to this case. In this paper, we only study the cases which exist feasible solutions.

3.3 Problem Hardness Analysis

Theorem 1. *The RAP selection problem for roadside data dissemination is NP-hard.*

Proof. We simplify the RAP selection problem by assuming $p_j^i = 1$ for $\forall v_i \in V$ and $\forall r_j \in R$. Hence, $Pr(v_i|S) \geq \tau$ for $\forall v_i \in V$ always holds. Therefore, the simplified problem is to select the minimal number of RAPs from R to cover all vehicles in V. It is obvious that $\bigcup_{r_j \in R} V(r_j) = V$ and $V(r_j) \subseteq V$. Let's denote $V_R = \{V(r_1), V(r_2), \cdots, V(r_j)\}$, i.e., V_R is a collection of V's subsets. For the RAP selection problem, if to make the selection of the RAP r_k corresponds to the selection of V's subset $V(r_k)$, this problem finally can be illustrated as the following (2).

$$\begin{aligned} minimize &: |V_S| \\ subject\ to &: \bigcup_{V(r_k) \in V_S} V(r_k) = V; \\ & V_S \subseteq V_R. \end{aligned} \tag{2}$$

Based on [14], we can conclude that Eq. 2 is same as the problem of Set Cover. Therefore, the RAP selection problem is NP-hard. Moreover, we can find that, the RAP selection problem is different from the Set Cover problem and is a meaningful extension of it.

4 Greedy Algorithm and Performance Analysis

In this section, we propose a greedy algorithm to solve the RAP selection problem, and analyze the correctness, complexity and approximation ratio of this algorithm.

4.1 Greedy Algorithm

Before the algorithm, we define a utility function $f(\cdot): 2^R \rightarrow Q^+$, which is a mapping from a RAP set to a real utility value. It indicates the probabilistic coverage utility of a given RAP set, in other words, the sum of probabilities that each vehicle in V successfully receives data from the given set of RAPs. Concretely, the utility is defined as follows.

$$f(S) = \theta * \sum_{v_i \in V} min\{Pr(v_i|S), \tau\} \tag{3}$$

where $\theta = max\{\frac{1}{\theta_1}, \frac{1}{\tau - \theta_2}, \frac{1}{\theta_3}\}$, $\theta_1 = p_{min} * (1 - p_{max})^{d_{max}}$, $\theta_2 = max\{Pr(v_i|S) \mid \forall v_i \in V, S \subseteq R, Pr(v_i|S) < \tau\}$, $\theta_3 = \frac{\tau * |V|}{|R|}$, $d_{max} = max\{deg(v_i) \mid \forall v_i \in V\}$, $p_{max} = max\{p_j^i \mid \forall v_i \in V, \forall r_j \in R\}$, $p_{min} = min\{p_j^i \mid \forall v_i \in V, \forall r_j \in R\}$.

In Eq. 3, $min\{Pr(v_i|S), \tau\}$ is the probabilistic coverage utility of S to the vehicle v_i, i.e., the (joint) probability that v_i successfully receives data from the RAPs in S. Moreover, along with the increase of the probability, the utility value $min\{Pr(v_i|S), \tau\}$ will become larger and larger. When the probability exceeds the threshold τ, the utility value will not change any more. θ is a parameter that we defined for the approximation ratio analysis in the following part.

Let S be the selected RAP set. When a RAP r_k is added to S, some vehicles which haven't been covered will be covered by r_k, or some vehicles' probabilities of receiving data successfully will increase. Both of them result in the growth of the utility. Then, the *basic idea* of our algorithm is to select the RAP r_k which maximizes the growth of the utility, and add it to the set S in each round. The concrete algorithm is showed in Algorithm 1, where S is initialized to be $\emptyset$. In each round, the RAP r_k which maximize $f(S + \{r_k\})$ is selected and added into S. The algorithm terminates when $f(S) = \theta * \tau * |V|$.

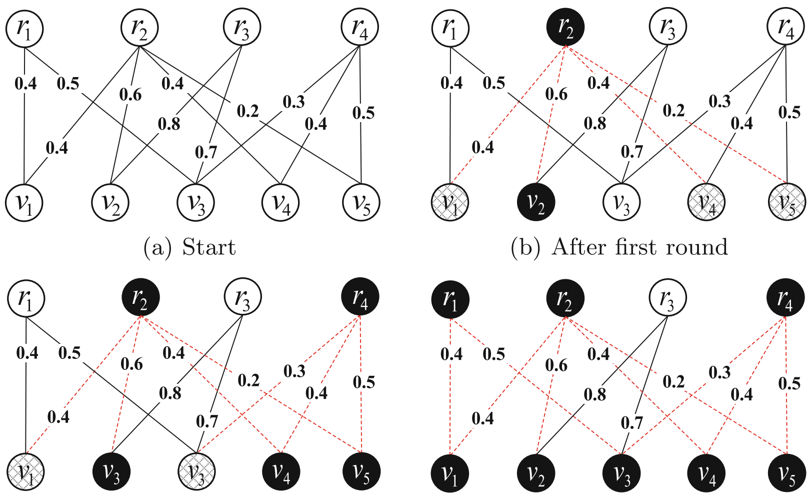
Algorithm 1. Greedy Algorithm for RAP Selection

Input: V, R, S, τ, p_j^i for $\forall v_i$ and $\forall r_j$;
Start:
$S = \emptyset$;
while $f(S) < \theta * \tau * |V|$ **do**
    choose the RAP $r_k \in R \backslash S$ which maximize $f(S + \{r_k\})$;
    $S = S \cup \{r_k\}$;
End
Output: S

We use the example in Fig. 2(a) to illustrate Algorithm 1 and let $\tau = 0.6$. In addition, there is a small trick in Algorithm 1. Since θ is a constant and only be used in the theoretical analysis of approximation ratio, we can simply let $\theta = 1$ in the real implementation of Algorithm 1. That will not change the final result. Note that, we denote $f_i(S) = min\{Pr(v_i|S), \tau\}$. Figure 2(b)$\sim$(d) show the corresponding results.

1. First round: $f(\{r_1\}) = 0.9$, $f(\{r_2\}) = 1.6$,$f(\{r_3\}) = 1.2$, $f(\{r_4\}) = 1.2$. Hence, we select r_2 and add it to S. After this round, $Pr(v_1|S) = 0.4$, $Pr(v_2|S)=0.6$,$Pr(v_3|S)=0$, $Pr(v_4|S)=0.4$, $Pr(v_5|S)=0.2$. That is to say, v_2 is covered and its probability of receiving data successfully is not less than 0.6, v_3 isn't covered, v_1, v_4, v_5 are covered with some probabilities of receiving data successfully less than 0.6.
2. Second round: $f(S\cup\{r_1\})=2.3$, $f(S\cup\{r_3\})=2.2$, $f(S\cup\{r_4\})=2.5$. Therefore, r_4 is selected and added into S. After this round, both v_4 and v_5 can receive data successfully with the probability no less than 0.6.
3. Third round: $f(S \cup \{r_1\})=3$, $f(S \cup \{r_3\})=2.8$. Therefore, r_1 is selected and added into S.

   After this round, we have $f(S)=3=\theta*\tau*|V|$. Therefore, Algorithm 1 terminates. As $f_i(S)=0.6$ for $\forall v_i \in V, S=\{r_1, r_2, r_4\}$ is a feasible solution.



(a) Start (b) After first round

(c) After second round (d) After third round

Fig. 2. The results of Algorithm 1 being conducted on the example in Fig. 2(a). (The white RAP nodes are not selected and the black RAP nodes are selected. The white vehicle nodes are not covered, the black vehicle nodes are covered and their probabilities of receiving data successfully are not less than 0.6, the grid vehicle nodes are covered but their probabilities of receiving data successfully are less than 0.6.)

4.2 Correctness, Complexity and Approximation Ratio

We first show the correctness of Algorithm 1 by theoretical analysis. On the one hand, in Algorithm 1, only one RAP is added into S in each round. In the worst case, all of RAPs in R are added into S after $|R|$-th round. For each vehicle, if all of RAPs are selected, its probability of receiving data successfully is not less than τ (assumed before). In this case, $f(S)$ must be $\theta * \tau * |V|$, since $min\{Pr(v_i|S), \tau\} = \tau$ for $\forall v_i \in V$. Therefore, Algorithm 1 terminates for sure. On the other hand, if $f(S) = \theta * \tau * |V|$, $min\{Pr(v_i|S), \tau\}$ must equal to τ for $\forall v_i \in V$. Hence, each vehicle in V can receive data successfully from the RAPs in S with the probability no less than τ, namely, S is a feasible solution of the RAP selection problem. In turn, if S is a feasible solution, $f(S)$ must be $\theta{*}\tau{*}|V|$ after each RAP in S is selected. In summary, our Algorithm 1 is correct.

Then, we analyze the complexity of Algorithm 1. In Algorithm 1, the loop body will run $|R|$ times in the worst case. During each round of Algorithm 1, in order to choose the maximal $f(S+\{r_k\})$ for $\forall r_k \in R \backslash S$, the algorithm needs to test all of $|R \backslash S|$ cases. Moreover, Algorithm 1 also need to compute $f(S{+}\{r_k\})$, so as to compare the values of $f(S{+}\{r_k\})$ for $\forall r_k \in R \backslash S$. The complexity of compute $f(S)$ is $O(|V|{*}|R|)$, according to Eq. 3. In sum, the complexity of Algorithm 1 is $O(|V|{*}|R|^3)$.

Finally, we give the approximation ratio of Algorithm 1, via the following Theorem 2, and its detail proofs are showed in appendix.

Theorem 2. *Our Algorithm 1 produces an approximation solution with a ratio of* $1 + ln(\frac{\theta * \tau * |V|}{opt})$ *from the optimal, where opt is the number of selected RAPs produced by the optimal solution.*

5 Evaluation

We conduct extensive simulations to evaluate the performance of our proposed algorithm. The compared algorithms, the traces that we used, the simulation settings, and the results are presented as follows.

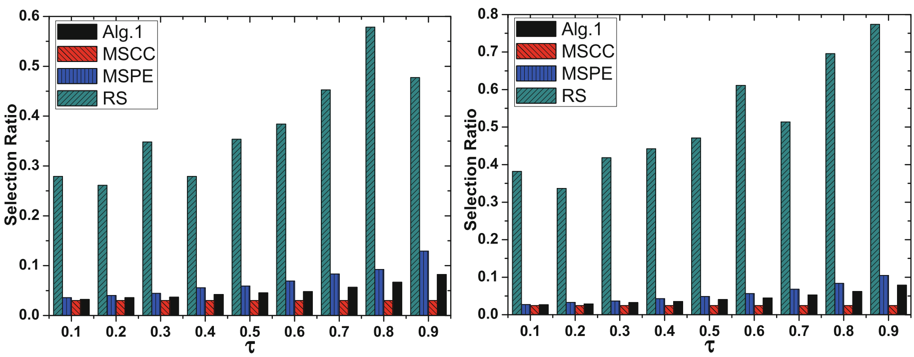
5.1 Compared Algorithms and Traces

In order to evaluate the performance of our proposed algorithm, some compared algorithms need to be introduced. As we know, it's an effective method to solve the problem of coverage, by using the degree centrality. In [2], the authors also took into account the degree centrality of each vehicle, in order to select the seeds. Based on this fact, we design two compared algorithms: MSCC (Minimum Selection for Covering Completely) and MSPE(Minimum Selection with Probability Ensuring). Differing from our Algorithm 1, which selects the RAP r_k from $R \backslash S$ to maximize $f(S + \{r_k\})$ in each round until all of the vehicles are covered with some probabilities of receiving data successfully no less than the threshold, MSCC selects the RAP r_k with maximum degree, i.e., the RAP with maximum $deg(r_k)$ for $\forall r_k \in R \backslash S$ in each round, until all of the vehicles are covered. Note that, MSCC do not ensure that each vehicle can receive data successfully with probability no less than the threshold. Obviously, it's the minimal guarantee that a vehicle can receive data. MSPE is similar with MSCC, which also select the RAP with maximum degree in each round, and terminates when each vehicle can receive data successfully with enough probability, i.e., no less than the threshold. Finally, we also design RS (Random Selection), which randomly selects a RAP in each round until each vehicle in V is covered with probability of receiving data successfully no less than the threshold. Moreover, in order to improve the performance of RS, its values showed in Figs. 3, 4 and 5 are the optimal among 10 running results.

We evaluate the metric of these four schemes, by using the real data of bus systems in two cities, Hefei and Shanghai in China. The data of Hefei includes 125 bus lines and 899 bus stations, and the data of Shanghai includes 503 bus lines and 3850 stations. We regard each bus line as a mobile vehicle, each station as a location where a RAP is deployed. Therefore, the moving pattern of a vehicle is the set of the stations passed by corresponding bus line. Further, the probabilities that each vehicle receives data successfully from the RAPs are generated randomly.

5.2 Evaluation Metrics, Methods and Results

First, we use the metric *Selection Ratio* to compare these algorithms. Selection ratio refers to the ratio of the number of the selected RAPs to the total number of all RAPs. We compare their performances with different threshold τ, and θ is set to 1.0. The detail results are showed in Fig. 3.



(a) on the data of hefei (b) on the data of shanghai

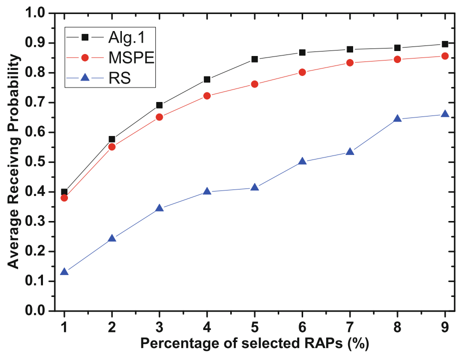
Fig. 3. Performance comparison: selection ratio vs. τ

In both Fig. 3(a) and (b), we can see that Algorithm 1 achieves good performance and precedes MSPE and RS, with $\tau = \{0.1, \cdots, 0.9\}$. MSCC selects the minimal number of RAPs, about 3 % in Fig. 3(a) and 2.4 % in Fig. 3(b), without ensuring the probabilities. And its values don't change with the increment of τ, therefore can be regard as benchmark. RS has the worst performance. Its ratio much exceeds the others and fluctuates strongly. MSPE has close performance to Algorithm 1 when τ is rather small, and becomes worse and worse with τ increasing. When $\tau = 0.9$, the results of Algorithm 1 showed in Fig. 3(a) and

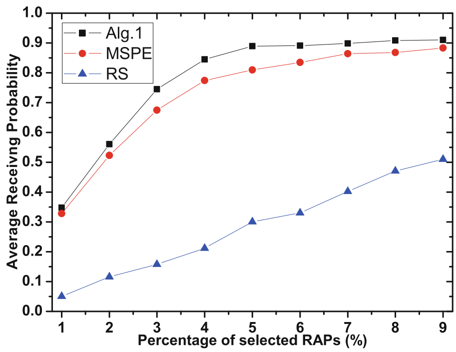
(b) are about 36 % and 25 % smaller than MSPE, respectively. The reason is that, when τ is small, most of vehicles can receive data successfully with some probabilities no less than τ from one single RAP. Along with τ increases, more and more vehicles need to be covered jointly by two or more RAPs, and at this moment, taking into account only degree without the probability attached by the edge is partial. However, for Algorithm 1, its RAP selection in each round considers the coverage utility to all of vehicles, therefore, it achieves the best performance.

We also conclude from Fig. 3, that the selection ratio of Algorithm 1 doesn't rise severely as the threshold τ increases. Therefore, an appropriate and large value of τ can be chose without worrying about the consequent bad performance.

Second, we also define the metric of *average receiving probability* to compare Algorithm 1, MSPE, and RS. Here, we don't care that, whether or not the probability of each vehicle of receiving data successfully is less than the threshold τ. And we only select parts of RAPs (1 %, $\cdots$, 9 %), and compute the corresponding average probability of all of vehicles receiving data successfully. The results are depicted in Fig. 4(a) and (b). From Fig. 4(a) and (b), we can find



(a) on the data of hefei

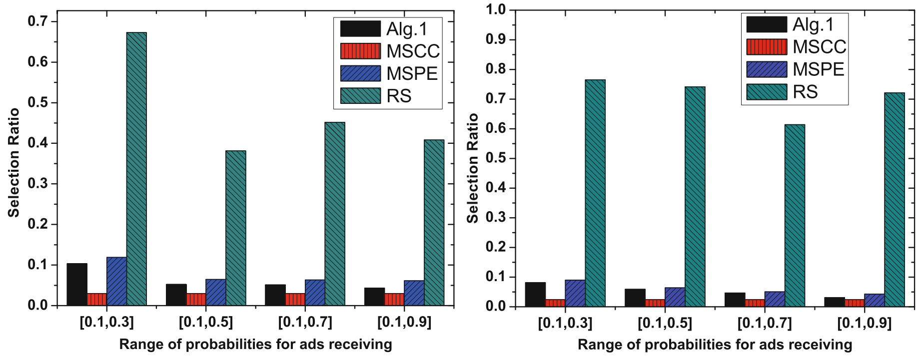


(b) on the data of shanghai

Fig. 4. Performance comparison: average receiving probability vs. percentage of selected RAPs

that, with percentage of selected RAPs increasing, the three values of average receiving probabilities increase. Among them, RS increases slowly and show fluctuation with the lowest average probability. However, the average probabilities of Algorithm 1 and MSPE increases fast when selected RAPs is less than roughly 5 %, then approach slowly to about 0.9. in spite of this, Algorithm 1 is still better than MSPE since its average probability reaches to 0.9 faster. It's because that Algorithm 1 considers the coverage utility of each RAP to all vehicles in each round, instead of just the degree of a RAP.

Finally, we also examine the metric of selection ratio of these schemes, by randomly selecting the probability between any pair of vehicle and RAP, i.e., p_j^i



(a) on the data of hefei (b) on the data of shanghai

Fig. 5. Performance comparison: selection ratio vs. range of probabilities for data receiving

for $\forall v_i \in V$ and $\forall r_j \in R(v_i)$, from $[0.1, 0.3]$, $[0.1, 0.5]$, $[0.1, 0.7]$ $[0.1, 0.9]$. Here, $\tau = 0.5$, $\theta = 1.0$. Such settings have practical significance, because it simply describe the environment of VANETs with different levels of connectivity. Figure 5(a) and (b) plot the results, from which we still can get that Algorithm 1 achieves the best performance.

6 Conclusion

In this paper, we focus on the roadside data dissemination in VANETs. Its scenario includes three basic elements: RSP, mobile vehicles and requesters. We first introduce the model of the roadside data dissemination. Based on the model, we propose a RAP selection problem for roadside data dissemination, which is to select the minimal number of RAPs to cover all vehicles in model, and ensure that each vehicle can receive data successfully with probability no less than a threshold. Then we prove that the RAP selection problem is NP-hard by converting it to the Set Cover problem. Next, we propose a greedy approximation algorithm to solve this problem, analyze the complexity and approximation ratio of this proposed algorithm. Finally, we conduct extensive simulations to show the good performance of our greedy algorithm. The results indicate that our algorithm can select the minimum number of RAPs to finish the roadside data dissemination, than other compared schemes.

Acknowledgment. This research was supported in part by the National Natural Science Foundation of China (NSFC) (Grant No. 61572457, 61572336, 61502261, 61379132), and the Natural Science Foundation of Jiangsu Province in China (Grant No. BK20131174, BK2009150).

Appendix: The Detailed Proof of Theorem 2

We first give two important properties of our utility function $f(\cdot)$.

Lemma 1. *$f(\emptyset)=0$ and $f(\cdot)$ is an increasing function.*

Proof. 1. We define $f_i(S)=min\{Pr(v_i|S),\tau\}$. If $S=\emptyset$, $Pr(v_i|S)=0$ for $\forall v_i\in V$. Hence, $f_i(S)=0$ for $\forall v_i\in V$. Further, $f(S)=\sum_{v_i\in V} f_i(S)=0$.
2. Giving any two sets X and Y, and suppose $X\subseteq Y\subseteq A$, obviously we have $Pr(v_i|X)\le Pr(v_i|Y)$. Consequently, we have $f_i(X)\le f_i(Y)$ for $\forall v_i\in V$. Moreover, because of $\theta>0$, $f(X)=\theta*\sum_{v_i\in V} f_i(X)\le\theta*\sum_{v_i\in V} f_i(Y)=f(Y)$ when $X\subseteq Y\subseteq R$. Therefore, $f(\cdot)$ is an increasing function.

Then, we prove that our utility function $f(\cdot)$ is submodular by giving the below Lemma 2, since we all know that, submodular function relates closely to greedy algorithm.

Lemma 2. *$f(\cdot)$ is a submodular function.*

Proof. Given $X\subseteq Y\subseteq R$, $\forall r_k\in R\backslash Y$, if $\Delta_{r_k}f(X)=f(X+\{r_k\})-f(X)\ge\Delta_{r_k}f(Y)=f(Y+\{r_k\})-f(Y)$, $f(\cdot)$ is a submodular function. Before proving this conclusion, we first compare the size of $\Delta_{r_k}f_i(X)=f_i(X+\{r_k\})-f_i(X)$ and $\Delta_{r_k}f_i(Y)=f_i(Y+\{r_k\})-f_i(Y)$, by dividing all possibilities into the following six cases. Note that, $Pr(v_i|(X+\{r_k\}))\ge Pr(v_i|X)$, $Pr(v_i|(Y+\{r_k\}))\ge Pr(v_i|Y)$, $Pr(v_i|Y)\ge Pr(v_i|X)$ and $Pr(v_i|(Y+\{r_k\}))\ge Pr(v_i|(X+\{r_k\}))$.

1. $\tau<Pr(v_i|X)$, $\tau<Pr(v_i|Y)$. Hence, $\Delta_{r_k}f_i(X)-\Delta_{r_k}f_i(Y)=[min\{Pr(v_i|(X+\{r_k\})),\tau\}-min\{Pr(v_i|X),\tau\}]-[min\{Pr(v_i|(Y+\{r_k\})),\tau\}-min\{Pr(v_i|Y),\tau\}]=(\tau-\tau)-(\tau-\tau)=0$;
2. $Pr(v_i|X)\le\tau<Pr(v_i|(X+\{r_k\}))$, $\tau<Pr(v_i|Y)$. Hence, $\Delta_{r_k}f_i(X)-\Delta_{r_k}f_i(Y)=(\tau-Pr(v_i|X))-(\tau-\tau)\ge 0$;
3. $Pr(v_i|X)\le\tau<Pr(v_i|(X+\{r_k\}))$, $Pr(v_i|Y)\le\tau<Pr(v_i|(X+\{r_k\}))$. Hence, $\Delta_{r_k}f_i(X)-\Delta_{r_k}f_i(Y)=(\tau-Pr(v_i|X))-(\tau-Pr(v_i|Y))=Pr(v_i|Y)-Pr(v_i|X)\ge 0$;
4. $Pr(v_i|(X+\{r_k\}))\le\tau$, $\tau<Pr(v_i|Y)$. Hence, $\Delta_{r_k}f_i(X)-\Delta_{r_k}f_i(Y)=[Pr(v_i|(X+\{r_k\}))-Pr(v_i|X)]-(\tau-\tau)\ge 0$;
5. $Pr(v_i|(X+\{r_k\}))\le\tau$, $Pr(v_i|(Y+\{r_k\}))\le\tau$. Hence, $\Delta_{r_k}f_i(X)-\Delta_{r_k}f_i(Y)=[Pr(v_i|(X+\{r_k\}))-Pr(v_i|X)]-[Pr(v_i|(Y+\{r_k\}))-Pr(v_i|Y)]$;
6. $Pr(v_i|(X+\{r_k\}))\le\tau$, $Pr(v_i|Y)\le\tau<Pr(v_i|(Y+\{r_k\}))$. Hence, $\Delta_{r_k}f_i(X)-\Delta_{r_k}f_i(Y)=[Pr(v_i|(X+\{r_k\}))-Pr(v_i|X)]-[\tau-Pr(v_i|Y)]\ge[Pr(v_i|(X+\{r_k\}))-Pr(v_i|X)]-[Pr(v_i|(Y+\{r_k\}))-Pr(v_i|Y)]$.

For the cases 5 and 6 above, we need to continue analyzing the size of $Pr(v_i|(S+\{r_k\}))-Pr(v_i|S)$ for $\forall v_i\in V$ and $\forall r_k\in R\backslash S$. We define $\Delta_{r_k}Pr(v_i|S)=Pr(v_i|(S+\{r_k\}))-Pr(v_i|S)$. In fact, $\Delta_{r_k}Pr(v_i|S)$ for $\forall v_i$ is the increment of its probability of receiving data successfully when a new RAP r_k is added to S. It is obvious that r_k only influences the vehicles in $V(r_k)$. In order to compute $\Delta_{r_k}Pr(v_i|S)$, there exists three possible case below:

1. v_i is covered by r_k and (part or all of) the RAPs in S. Hence, $\Delta_{r_k} Pr(v_i|S)$ $= [1 - \prod_{r_j \in (S+\{r_k\})(v_i)} (1 - p_j^i)] - [1 - \prod_{r_j \in S(v_i)} (1 - p_j^i)] = \prod_{r_j \in S(v_i)} (1 - p_j^i) - (1 - p_k^i) \prod_{r_j \in S(v_i)} (1 - p_j^i) = p_k^i * \prod_{r_j \in S(v_i)} (1 - p_j^i)$;
2. v_i is covered by r_k but isn't covered by any RAPs in S. Hence, $\Delta_{r_k} Pr(v_i|S) = [1 - (1 - p_k^i)] - 0 = p_k^i$;
3. v_i isn't covered by r_k, namely, wether or not r_k is selected has no influence on v_i. Obviously, $\Delta_{r_k} Pr(v_i|S) = 0$.

Based on the analyses above, we can compute $\Delta_{r_k} Pr(v_i|X) - \Delta_{r_k} Pr(v_i|Y)$ by dividing it into following four possible cases. Note that, $X \subseteq Y$ and $X(v_i) \subseteq Y(v_i)$.

1. v_i isn't covered by any RAPs in Y, but is covered by r_k. Hence, $\Delta_{r_k} Pr(v_i|X)$– $\Delta_{r_k} Pr(v_i|Y) = p_k^i - p_k^i = 0$;
2. v_i isn't covered by any RAPs in X, but r_k and RAPs in $Y \backslash X$ cover v_i. Hence, $\Delta_{r_k} Pr(v_i|X) - \Delta_{r_k} Pr(v_i|Y) = p_k^i - p_k^i * \prod_{r_j \in Y(v_i)} (1 - p_j^i) \geq 0$;
3. r_k and RAPs in X and Y cover v_i. Hence, $\Delta_{r_k} Pr(v_i|X) - \Delta_{r_k} Pr(v_i|Y) = p_k^i * [\prod_{r_j \in X(v_i)} (1 - p_j^i) - \prod_{r_j \in Y(v_i)} (1 - p_j^i)] \geq 0$ since $X(v_i) \subseteq Y(v_i)$;
4. v_i isn't covered by r_k. Hence, $\Delta_{r_k} Pr(v_i|X) - \Delta_{r_k} Pr(v_i|Y) = 0 - 0 = 0$.

In summary, we have $\Delta_{r_k} Pr(v_i|X) - \Delta_{r_k} Pr(v_i|Y) \geq 0$. Finally, we also can conclude that $\Delta_{r_k} f_i(X) - \Delta_{r_k} f_i(Y) \geq 0$ in all six cases. Moreover, as $\theta > 0$, $\Delta_{r_k} f(X) - \Delta_{r_k} f(Y) = \theta * \sum_{v_i \in V} [\Delta_{r_k} f_i(X) - \Delta_{r_k} f_i(Y)] \geq 0$. Therefore, we have the conclusion that $f(\cdot)$ is a submodular function.

According to Lemmas 1 and 2, we know that, $f(\cdot)$ is polymatroid since it's an increasing submodular function with $f(\emptyset) = 0$. Similarly, we can easily prove that the cardinality function $c(X) = |X|$ is polymatroid. Given two polymatroid functions $g(\cdot)$ and $h(\cdot)$ on 2^E, the problem of Minimum Submodular Cover with Submodular Cost (MSC/SC) is defined, which is the minimization problem $min\{h(X)|g(X) = g(E), X \subseteq E\}$[15].

In this paper, for the utility function $f(\cdot)$, given a selected RAP set S, if $f(S) = \theta * \tau * |V|$, S is a feasible solution of the RAP selection problem and $f(S) = f(V) = \theta * \tau * |V|$. Therefore, the RAP selection problem can be described as: $min\{c(S)|f(S) = f(V), S \subseteq V\}$, where $c(\cdot)$ is the cardinality function. That is to say, our selection problem is a MSC/SC problem. Based on this fact, we give the following Theorem 3.

Theorem 3. *[15] Suppose $g(\cdot)$ is a polymatroid function on 2^E, and $g(E) \geq opt$ where opt is the cost of a minimum submodular cover. For a greedy algorithm, if the selected x in each round always satisfies that $\frac{g(X+\{x\}) - g(X)}{c(\{x\})} \geq 1$, then the greedy solution is a $1 + \rho ln(\frac{g(E)}{opt})$-approximation, where $\rho = 1$ if $c(\cdot)$ is modular (i.e., linear).*

Now, we give the following crucial Lemma 3.

Lemma 3. *Give the set $S \subseteq R$, which is the set of RAPs selected by Algorithm 1 after r-th round, and the RAP r_k selected during $(r+1)$-th round, we have $\frac{f(S+\{r_k\}) - f(S)}{c(\{r_k\})} \geq 1$.*

Proof. 1. It is obvious that the cardinality function $c(\cdot)$ is linear and $c(\{r_k\})=1$.
2. In each round of Algorithm 1, if $f(S)<\theta{*}\tau{*}|V|$, a new RAP will be selected and added into the set S. In fact, if Algorithm 1 doesn't terminate after r-th round, there must exist a vehicle v_i with $Pr(v_i|S)<\tau$. Otherwise, if $Pr(v_i|S)\geq\tau$ for $\forall v_i\in V$, then $f(S)=\theta{*}\tau{*}|V|$, consequently, Algorithm 1 will terminate. Moreover, for this vehicle v_i and the selected RAP r_k, since $f(S{+}\{r_k\})$ is maximized during $(r{+}1)$-th round, we can suppose r_k must cover this v_i, i.e., $p_k^i>0$, without loss of generality. That is to say, if the RAP r_k is selected during $(r{+}1)$-th round, r_k must cover at least one vehicle v_i with $Pr(v_i|S)<\tau$. Otherwise, we have $f(S{+}\{r_k\}){-}f(S)=0$ for $\forall r_k\in R\backslash S$, which conflicts with the fact that there must exist a vehicle v_i with $Pr(v_i|S)<\tau$ if Algorithm 1 doesn't terminate. Based on these analyses, we have

$f(S{+}\{r_k\}){-}f(S) = \theta{*}\sum_{v_i\in V}[min\{Pr(v_i|(S{+}\{r_k\})),\tau\}{-}min\{Pr(v_i|S),\tau\}]$
$\geq \theta*(min\{Pr(v_i|(S+\{r_k\})),\tau\} - min\{Pr(v_i|S),\tau\}) = \theta*(min\{Pr(v_i|(S+\{r_k\})),\tau\} - Pr(v_i|S)) = \theta*min\{Pr(v_i|(S+\{r_k\})) - Pr(v_i|S),\tau - Pr(v_i|S)\}$
$=\theta*min\{p_k^i*\prod_{r_j\in S(v_i)}(1-p_j^i),\tau - Pr(v_i|S)\}\geq min\{\theta*\theta_1,\theta*(\tau-\theta_2)\}\geq 1$

To sum up, $\frac{f(S+\{r_k\})-f(S)}{c(\{r_k\})}\geq f(S{+}\{r_k\}){-}f(S)\geq 1$

Finally, suppose that S_{opt} is an optimal solution of the RAP selection problem and $c(S_{opt})=|S_{opt}|=opt$ which is number of RAPs in S_{opt}. Consequently, we have $f(R)=\theta{*}\tau{*}|V|\geq\frac{\tau*|V|}{\theta_3}\geq|R|\geq opt$ since $\theta\geq\frac{1}{\theta_3}$. Therefore, we can conclude that our proposed Algorithm 1 is a $1{+}\ln(\frac{\theta*\tau*|V|}{opt})$-approximation, by applying Theorem 3.

References

1. Zheng, H., Jie, W.: Optimizing roadside advertisement dissemination in vehicular cyber-physical systems. In: IEEE International Conference on Distributed Computing Systems (ICDCS) (2015)
2. Qin, J., Zhu, H., Zhu, Y., Li, L., Xue, G., Li, M.: Post: Exploiting dynamic sociality for mobile advertising in vehicular networks. In: IEEE INFOCOM (2014)
3. Li, Z., Liu, C., Chigan, C.: On secure vanet-based ad dissemination with pragmatic cost and effect control. IEEE Trans. Intell. Transp. Syst. **14**(1), 124–135 (2013)
4. Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Networks **1**(3), 215–239 (1979)
5. Guo, L., Zhang, C., Yue, H., Fang, Y.: A privacy-preserving social-assisted mobile content dissemination scheme in DTNs. In: IEEE INFOCOM (2013)
6. Ning, T., Yang, Z., Hongyi, W., Han, Z.: Self-interest-driven incentives for ad dissemination in autonomous mobile social networks. In IEEE INFOCOM (2013)
7. Gao, W., Cao, G.: User-centric data dissemination in disruption tolerant networks. In: IEEE INFOCOM (2011)
8. Lee, S.-B., Park, J.-S., Gerla, M., Songwu, L.: Secure incentives for commercial ad dissemination in vehicular networks. IEEE Trans. Veh. Technol. **61**(6), 2715–2728 (2012)
9. Jeong, J., Guo, S., Gu, Y., He, T., Du, D.H.C.: TSF: Trajectory-based statistical forwarding for infrastructure-to-vehicle data delivery in vehicular networks. In: IEEE International Conference on Distributed Computing Systems (ICDCS) (2010)

10. Sardari, M., Hendessi, F., Fekri, F.: Infocast: A new paradigm for collaborative content distribution from roadside units to vehicular networks. In: IEEE SECON proceedings (2009)
11. Liu, K., Lee, V.C.S.: RSU-based real-time data access in dynamic vehicular networks. In: International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1051–1056 (2010)
12. Tsung-Jung, W., Liao, W., Chang, C.-J.: A cost-effective strategy for road-side unit placement in vehicular networks. IEEE Trans. Commun. **60**(8), 2295–2303 (2012)
13. Aslam, B., Amjad, F., Zou, C.C.: Optimal roadside units placement in urban areas for vehicular networks. In: IEEE Symposium on Computers and Communications (ISCC) (2012)
14. Feige, U.: A threshold of ln n for approximating set cover. J. ACM **45**(4), 634–652 (1998)
15. Wan, P., Dingzhu, D., Pardalos, P., Weili, W.: Greedy approximations for minimum submodular cover with submodular cost. Comput. Optim. Appl. **45**(2), 463–474 (2010)