POST: Exploiting Dynamic Sociality for Mobile Advertising in Vehicular Networks

Jun Qin[1], Hongzi Zhu[1], Yanmin Zhu[1], Li Lu[2], Guangtao Xue[1], Minglu Li[1]

[1] Shanghai Jiao Tong University, China
[2] University of Electronic Science and Technology of China

qin.jun@sjtu.edu.cn, {hongzi, yzhu, xue-gt, li-ml}@cs.sjtu.edu.cn, luli2009@uestc.edu.cn

***Abstract*—Mobile advertising in vehicular networks is of great interest with which timely information can be fast spread into the network. Given a limited budget for hiring seed vehicles, how to achieve the maximum advertising coverage within a given period of time is NP-hard. In this paper, we propose an innovative scheme, POST, for mobile advertising in vehicular networks. The POST design is based on two key observations we have found by analyzing three large-scale vehicle traces. First, vehicles demonstrate dynamic sociality in the network; second, such vehicular sociality has strong temporal correlations. With the knowledge, POST uses Markov chains to infer future vehicular sociality and adopts one greedy heuristic to select the most "centric" vehicles as seeds for mobile advertising. Extensive trace-driven simulation results show that POST can greatly improve the coverage and the intensity of advertising.**

Keywords-vehicular networks; mobile advertising; dynamic sociality; social network analysis

I. INTRODUCTION

Vehicular networks are emerging as a new landscape of mobile ad hoc networks, aiming to provide a wide spectrum of safety and comfort applications to drivers and passengers. In vehicular networks, vehicles equipped with wireless communication devices can transfer data with each other (vehicle-to-vehicle communications) as well as with the roadside infrastructure (vehicle-to-roadside communications). With vehicular networks, a wide range of new Intelligent Transportation System (ITS) applications are enabled, ranging from hazard warning, collision avoidance, and traffic management to navigation based on real-time traffic condition, trip planning and optimal route selection.

Among all the others, *mobile advertising* is an appealing application, where a small number of public vehicles such as taxies and buses, called *seed* vehicles, are chosen to propagate timely digital advertisements to the other vehicles in the network. In such application scenarios, a seed can forward or "post" its advertisements to its neighboring vehicles or other mobile devices (e.g., smart phones) via short-range wireless communications when they approach to each other (called a *contact*). As a result, the advertisement information can be gradually spread out within the network. Considering the limited budget, the goal of the application is to select the best set of seeds paid to post a piece of timely advertisement so that the total number of vehicles seeing the advertisements within a given period of time is maximized. With an effective mobile advertising scheme, a great deal of timely and important information, such as instant municipal announcements, real-time traffic congestion information, and commercial promotion activities, can be fast propagated among mobile devices at very low cost.

To realize the mobile advertising application, however, is very challenging due to three reasons. First, as vehicles move, the topology of the network varies fast over time. It is very hard to know the exact future information of the network. Second, even if the network topology is known, we prove that the problem of choosing a given number of vehicles to post advertisements such that the network coverage is maximal within a given time is NP-hard. Third, since a piece of advertisement may be time-critical, it should be spread as widely as possible while the content of the advertisement is still valid, which makes the problem even harder. One straightforward scheme might be to randomly choose a fix number of vehicles as seeds to flood the network. It is simple but there is no guarantee that those randomly chosen seeds can always have the optimal performance.

In the literature, recent work has studied the influence maximization problem in the area of social network analysis [1] [2] [3]. Based on static social networks, individual influence has been measured using various centrality metrics and utilized to choose good candidate nodes to spread information. Although these studies shed the light on how to select preferable seeds, they cannot be directly applied in the mobile advertising problem as they are based on traditional static social network of which the topology is stable. In the context of data dissemination in vehicular and opportunistic networks, previous work came up with various routing mechanisms [4] [5] [6], originating from the field of conventional mobile ad hoc networks. Those studies mainly focus on relatively short range data coverage and forwarding performance in terms of delivery ratio, delay and network overhead. In the mobile advertising problem, however, the major concern is how to select a best set of seed vehicles so that they can achieve as large coverage as possible within a given period of time. As a result, there is no successful solution, to the best of our knowledge, to addressing the mobile advertising problem in vehicular networks.

In this paper, we first take an empirical study on three real large-scale vehicular traces. By aggregating pairwise contacts, we find that vehicles demonstrate clear sociality in the extracted contact graphs, forming *vehicular social networks*. Considering the inherent dynamics of vehicular networks, we investigate how vehicular sociality, especially on three representative centrality metrics, changes over time by dividing time into slots of equal length under different scales of time. We have two key observations. First, the vehicular social network presents high dynamics which means the position or

978-14799-3360-0/14/$31.00 ©2014 IEEE

the "role" of individual vehicles in the network changes fast when it is observed under a small scale of time. Second, the dynamics of vehicular sociality show strong temporal patterns and correlations. With this knowledge, we propose an innovative scheme, POST, for mobile advertising in vehicular networks. The core idea of POST is to capture and utilize the temporal correlations of vehicular sociality to infer the contact behavior of the whole network in future, which can be further leveraged to improve the performance of mobile advertising. To this end, POST integrates three techniques: *capturing centrality correlations*, *inferring future vehicular centrality* and *selecting preferable seeds*. In particular, as the mobile advertising problem is NP-hard, POST adopts one greedy strategy to select the most "centric" vehicles to serve as seeds for mobile advertising. Results of extensive trace-driven simulations demonstrate the efficacy of POST design.

We highlight our main contributions in this paper as follow:

- We first define the mobile advertising problem in vehicular networks and prove its NP-hardness, where the budget for hiring seed vehicles to post timely advertisements is limited.
- We conduct an extensive measurement study on social centrality characteristics of vehicles in the network. Through our measurements, we observe that vehicular social centrality is highly dynamic and moreover shows strong temporal correlations.
- We propose an innovative scheme, POST, for addressing the mobile advertising problem in vehicular networks, which fully utilizes the temporal correlations of vehicular social centrality to effectively select a small set of seed vehicles to propagate an advertisement with the goal of maximizing the coverage of this advertisement within a short tenancy. We also demonstrate the effectiveness of POST through extensive trace-driven simulations.

The remainder of this paper is organized as following. Section II is the presentation of related work. We describe the system model and the mobile advertising problem in Section III. In Section IV, we study the empirical trace data and analyze vehicular sociality. The findings on the high dynamics and temporal correlations of vehicular centrality are presented in Section V. Section VI is dedicated for the detailed design of POST. Section VII presents the performance evaluation of POST through empirical trace-driven simulations. We conclude and outline the directions for future work in Section VIII.

II. Related Work

The mobile advertising problem is most related to the influence maximization problem in the areas of social network analysis and online marketing. We introduce those studies and compare them with our work in this section.

In the area of social network analysis, the influence of individuals in spread processes is widely studied. Kitsak et al [7] studied various effects of several classic centrality measures for choosing good spreaders in order to obtain the optimal performance in designing dissemination strategies throughout various complex networks. Their work suggests that in order to have the propagation area as large as possible, it is necessary to choose the most important nodes which are not connected directly in the early stage of the dissemination task. The work in [8] studied what is the key factor to influence a node in social contagion processes. Their detailed study shows that the decision made by a node in the network is highly influenced by its connected neighbors. In the work [9], it suggests that the importance of a node depends not only on its popularity but also the similarity between different nodes. In addition, similarity can be utilized to predict new linkage appearance in the future.

In the area of online marketing, as the Internet has expanded to the largest information platform, the influence maximization problem proposed by Domingos and Richardson [10] [11] has been studied as a fundamental algorithmic problem in related applications. Kempe et al [1] [12] prove the NP-hardness of the maximization problem under the classic spread models in their work and have proposed their approximation algorithm based on influential node identification techniques in the social networks. In the following work, Karsai et al [13] found that the information diffusion speed in social networks is usually slower than expected. They suggest that this phenomenon is because of the various correlations, e.g. the community structures embedded in the graph and topological correlations.

Although the above studies shed the light on how to select preferable seeds, they cannot be directly applied in the mobile advertising problem as they are based on traditional static social network where the topology of the network is relatively stable.

In vehicular and opportunistic networks, there are a large number of studies on data dissemination [4] [5] [6]. Conventionally, various protocols originated from the field of traditional mobile ad hoc networks are on the contact-level. The methods utilize local information about the groups of moving nodes, increasing the opportunities of data forwarding. In recent years, the emerging routing and relaying solutions based on social network analysis techniques studied the problem from a new perspective on improving the dissemination efficiency. Daly and Haahr [14] proposed SimBet in their work. SimBet tries to utilize the similarity between nodes in a social network to increase the probability of successfully packet delivery via the most central nodes and community structures in the network. Similarly, Bubble Rap [15] also utilizes the importance of the nodes with their centrality in the relaying. ContentPlace [16] is a method that exploits social behaviors of the users in decentralized interest exchange. Those studies focus on how to improve end-to-end delivery performance in terms of delivery ratio, delay and network overhead. In mobile advertising problem, however, the major concern is how to select a best set of seed vehicles so that they can achieve as large coverage as possible within a given period of time.

In conclusion, there exists no successful solution, to the best of our knowledge, to solving the mobile advertising problem in vehicular networks.

III. System Description and Problem Definition

A. System Description

We consider building the mobile advertising application upon urban vehicular networks, where the initial advertisements can be downloaded to seed vehicles via vehicle-to-roadside communications or other infrastructure-aided communications such as WiFi and 2G/3G networks. The propagation of advertisements relies on opportunistic short-range wireless communications such as Bluetooth, WiFi and DSRC [17].

The main advantage of using vehicles for advertising lies in three folds. First, as there is no need to deploy new billboards built as infrastructure, it can enormously save the deployment cost. Second, it also has much lower system maintenance cost than directly dispatching those advertisements to all vehicles via cellular networks (e.g., 2G/3G) since propagating advertisements via short-range wireless communications is free of charge. Third, it can also achieve extraordinary coverage utilizing the mobility of vehicles comparing to statically posted advertisements in tradition. We further consider using public commuting vehicles like taxies and buses served as seeds because they are public service vehicles and therefore have less privacy issues and they have longer service time and larger areas comparing to normal vehicles.

B. Problem Definition and Its Difficulty

As many advertisements are time-critical, it is preferable to spread out those advertisements before their contents are out-of-date. Furthermore, we also consider the budget to deploy advertisements given a price for employing vehicles. Therefore, we define our mobile advertising problem as follows:

Definition 1. *Given the budget B and price p for employing vehicles to propagate a piece of advertisement in a vehicular network for a given period of time T, how to select the best B/p vehicles in the network so that the total number of vehicles seeing the advertisement in the network is maximized?*

The mobile advertising problem is hard and we have the following theorem,

Theorem 1. *The mobile advertising problem is NP-hard.*

PROOF. Assume that all future movements of all vehicles in the network are known. With this assumption, we know all future communication opportunities between any pair of vehicles within the given period of time. We can construct a graph $G(N, E)$, where N is the set of nodes and E is the set of edges. Each vehicle in the network is a node in the graph and there is an edge between a pair of nodes in the graph if the corresponding vehicles can communicate at least once within the given period of time. Denote the set of each node n_i and all its neighbors as $S_i \subseteq N$ for $i = 1, \cdots, |N|$. With this graph, the problem is equal to finding B/p different S_j for $j = 1, \cdots, B/p$ so that their union contains as many nodes as possible. This is a classic *Max k-cover* problem which is NP-hard [18] and concludes the proof. ■

As it is very hard, if not impossible, to know all future information about the network, the mobile advertising problem can be even harder. We study this problem through an empirical methodology and elaborate the process in the following sections.

IV. Sociality Analysis on Empirical Traces

A. Collecting Vehicular Traces

In order to understand realistic vehicular mobility and conduct informed design of mobile advertising schemes in vehicular networks, it is of great importance to study the empirical data in terms of frequency and temporal distribution of contacts among them. For this purpose, we use three datasets consisting of traces from two metropolises in China and two types of vehicles, i.e., buses and taxies. Key statistics of the traces are listed in Table I.

Shanghai Buses: The trace consists of GPS reports sent from 2,501 buses which serve on 100 routes and cover the whole downtown area in Shanghai between Feb. 19 and Mar. 5, 2007. A commuting bus periodically sends GPS reports back to a backend data center via GPRS channel. The specific information contained in such a report includes: ID, the longitude and latitude coordinates of the current location, timestamp, moving speed, and heading direction. Due to the GPRS communication cost for data transmission, reports are sent at a granularity of around one minute.

Shanghai Taxies: We also collected the GPS trace of taxies in Shanghai collected between Feb. 1 and Mar. 3, 2007. We chose 2,109 taxies in the datasets which have consecutive GPS reports on each day during the 31 days. The information contained in a taxi GPS report is similar to that of bus except that taxies also report whether passengers are onboard. The granularity of reports is one minute for taxies with passengers and about 15 seconds for vacant ones.

Shenzhen Taxies: The trace collection of taxies in Shenzhen is similar to Shanghai taxi trace. We use the whole month trace in October, 2009. We chose 8,291 taxies which continuously send GPS reports during the whole period. Taxies in Shenzhen always send GPS reports on every one minute.

We choose taxies and buses for the study for three reasons. First, taxies and buses are two representative types of vehicles showing two distinct mobility patterns, namely, rather random and well scheduled, respectively. Second, as taxies and buses are public service vehicles, they commute in the city all the time and can cover a wide area, which makes them preferable candidates for mobile advertising. Third, the privacy problem is less concerned since they are public vehicles.

B. Constructing Contact Graph

In order to select the best set of vehicles as seeds to deploy advertisements with respect to the spread coverage, it is of

TABLE I. Main Statistics of Three Data Sets

Data Set	*Shanghai Taxi*	*Shanghai Bus*	*Shenzhen Taxi*
Vehicle number	2,109	2,501	8,291
From date	Feb. 1, 2007	Feb. 19, 2007	Oct. 1, 2009
Duration (day)	31	15	31
Granularity (second)	15*, 60**	60	60
Number of contacts	22,053,178	1,229,380	23,968,860

*Vacant, **passengers onboard

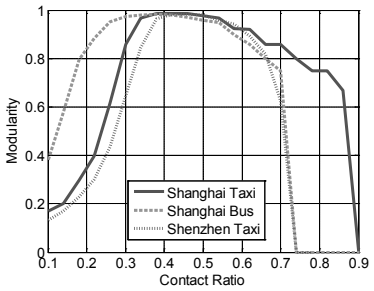

Figure 1. Modularity derived under different contact ratio using Shanghai taxi trace on Feb. 18, 2007

great importance to understand the position of individual vehicles in the network.

For this reason, we first construct a *contact graph* $\mathcal{G}(N, E)$ for each trace by aggregating the pairwise contacts. Each vehicle i is a node of the graph, $n_i \in N$, and the edge $e_{ij} \in E$ represents node i and j have certain acquaintance between them. The key to establishing a meaningful contact graph is the metric used to aggregate contacts, which determines whether two nodes share a link and the strength of this connection if exists. Various metrics, such as the number of total contacts observed [15], the age of last contact [19], and the contact frequency and total duration [15], have been used to derive edge strengths. In our study, we use contact ratio [20] to extract edges so that there is an edge between two nodes in the contact graph if the ratio of time with contacts observed to the total period of a trace is higher than a threshold and the weight on this edge takes the contact frequency value.

In order to determine the appropriate contact ratio to extract contact graph, we apply the Louvain algorithm [21] to find the community structure embedded in an established contact graph and evaluate the partition result using *modularity* [22] defined as

$$Q = \frac{1}{2m}\sum_{ij}\sum_{r}\left(A_{ij} - \frac{k_i k_j}{2m}\right)S_{ir}S_{jr}, \qquad (1)$$

where m is the total number of edges, A_{ij} is the element of the adjacency matrix (if there is an edge between node i and j, A_{ij} is the contact ratio between node i and j; otherwise, $A_{ij} = 0$), k_i and k_j are the degree of node i and j, respectively, and $S_{ir} = 1$ if node i belongs to group r and zero otherwise. We then examine the modularity of contact graphs derived under different contact ratios (shown in Fig. 1). It can be seen that the modularity first increases and then drops as the contact ratio increases. The reason is that on one hand, if a small threshold is used, extra edges derived from random or "unexpected" contacts would be added in the established graph, which may blur the network structures; on the other hand, if a large threshold is used, then more "regular" relationships would be abandoned from the graph, which also causes the loss of valuable topology information. Modularity higher than 0.3 implies there are strong social structures in the graph [22]. In order to reduce the influence of random contacts and while being able to preserve the essential topology information, we use the minimum contact ratio which results the established contact graph with a modularity greater than 0.6. Fig. 2 illustrates a contact graph extracted with a contact ratio of 0.26 using Shanghai taxi trace on Feb. 18, 2007, which contains 1,802 vehicle nodes. It can be seen that the community structure appears in this contact graph is very clear. With a corresponding modularity of 0.62, all the vehicles are divided into 32 communities. As relationships in the contact graph are extracted from pairwise contacts, it implies that a vehicle would have more chance to meet another vehicle within the same community than those outside the community.

C. Centrality Analysis on Contact Graph

With extracted contact graphs, we study the relative importance of individual vehicles within the network with respect to their *centrality* [23] as a more "centric" vehicle has higher probability to meet more vehicles and therefore can cover a larger area of the network. As there are many classic centrality metrics introduced in the field of social network analysis [23] [7], we study three most-related metrics:

1) The degree centrality [23]*:* is a natural way of measuring the importance of a node, which presents the number of other nodes in the graph that the node shares edges with. It presents the basic local structural property of nodes. In a contact graph, a vehicle with higher degree has greater probability to meet more vehicles, reflexing a higher popularity. The degree centrality of a given node in a contact graph is defined as the number of edges it is associated with.

2) *The closeness centrality* [23]*:* measures the reciprocal of the average distance of a node to all the other nodes in the network. A node with a higher closeness centrality implies it is closer to the other nodes in the network.

3) The coreness centrality [24]*:* measures the "depth" of a node within the network. The coreness of a node is k if it belongs to the k-core but not to the (k+1)-core, where the k-core of a graph is defined as a maximal sub-graph where each node has at least degree of k. A node with higher coreness is considered to be a better individual spreader in large-scale complex networks [7].

We examine the degree, closeness and coreness centrality in contact graphs derived from all available traces and plot their complementary cumulative density functions (CCDF) in Fig. 3. It can be seen that the CCDF of degree and coreness on all traces have exponential tails (i.e., linear under semi-logarithmic scale), which have been seen with different networks such as the power grid and railway networks [25]. The closeness, however, does not show any obvious

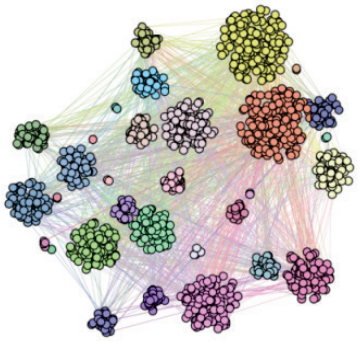

Figure 2. Contact graph extracted from Shanghai taxi trace on Feb. 18, 2007 containing 1,802 vehicles and 32 communities with a modularity of 0.62.

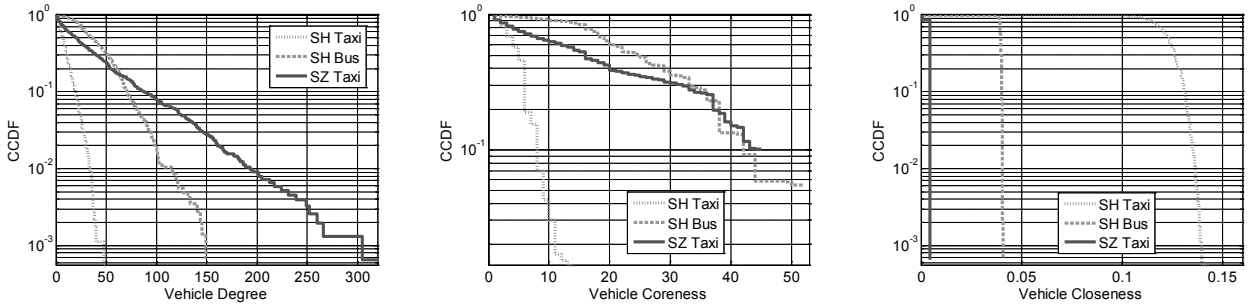

Figure 3. CCDFs of degree, coreness and closeness in all contact graphs constructed from traces of Shanghai taxies, Shanghai buses and Shenzhen taxies; exponential tails have been found in plots of degree and coreness.

distribution but is rather centralized. More specifically, we have two following key observations:

1) It shows that larger degree or coreness values do not mean larger closeness values, and vice versa, although the degree and coreness have similar distributions. Furthermore, the degree and coreness are good metrics to evaluate the centrality of vehicles with our traces while the closeness metric can hardly distinguish those "centric" vehicles from the rest as all vehicles have similar closeness values;

2) The exponential decay of both the degree and coreness metrics indicates that the portion of vehicles that have the largest degree or coreness values of all three types of vehicles appears small in both cities. This inspires us that it is possible to select only a small number of vehicles with the largest degree or coreness values to serve as candidate vehicles to perform mobile advertising.

V. Characterizing the Dynamic Sociality

Comparing to traditional social networks where edges are usually stable social relationships between people, links in contact graphs are extracted from highly dynamic contact information and therefore may vary significantly over time. In this section, we examine how contact graphs evolve with time.

A. Observing High Dynamics of Contact Graphs

In order to understand whether and how the network topology and the centrality properties of vehicles change over time, we divide time into slots of same length and construct a contact graph for each slot for all traces by aggregating all the contact events in that period using the graph extraction method as introduced in above section. We then compute all concerned centrality metrics for all vehicles in each contact graph.

Fig. 4(a) plots the time series of degrees of one hundred randomly-chosen vehicles from the Shanghai taxi trace. Contact graphs are constructed using time slots of one hour from Feb. 1 to Feb. 4 (96 hours in total). Note that all experiment vehicles are ordered according to their degrees in the contact graph in the first hour (the first column shown in Fig. 4(a)). Three key observations should be pointed out. First, by checking each row, it can be seen that the degree of every vehicle varies enormously over time; second, by checking each column, it can also be seen that the relative importance (rank) among vehicles also changes between consecutive time slots; last, although the degrees of vehicles demonstrate high dynamics over time, there are obvious temporal patterns embedded in the time series. For example, a clear periodicity of one day can be seen in the figure. Similar observations can be found in Fig. 4(b) when examining the coreness centrality of vehicles. In contrast, although clear temporal patterns can also found in the plot of closeness in Fig. 4(c), the absolute value and the relative rank of closeness among vehicles are rather stable which also explains the vertical drops in the CCDF shown in Fig. 3.

With these observations, we make statements as follows. First, in order to gain the maximum coverage of advertising within a given period of time, seed vehicles for mobile advertising should be chosen according to the current or even future status of the network instead of using some particular static set of vehicles. Second, the dynamics of node sociality especially centrality are possible to be captured and utilized for selecting better seed vehicles as they have clear temporal patterns. Third, the degree and coreness centralities are good metrics to study the dynamics of vehicular sociality as they show both temporal patterns and wide diversity among vehicles. In the remainder of this paper, we select the degree centrality to study for demonstration.

B. Revealing the Temporal Correlations of Centrality

In order to capture the dynamics observed, we examine the entropy and conditional entropy of degree centrality measures.

Specifically, let X be the random variable representing the degree measures of a vehicle. If we have observed M measures, these measures can be presented by a vector $D = (d_0, d_1, \cdots, d_{M-1})$ where d_i, $0 \le i \le M-1$denotes the i^{th} degree measure during the i^{th} time slot. The probability of the measure being j can be computed as x_j/M, where x_j represents the number of measure being j. Therefore, the entropy of D is:

$$H(X) = \sum_{j=0}^{\infty}(x_j/M)\log_2\frac{1}{x_j/M}. \quad (2)$$

When $K = 1$, let X' be the random variable for the last measure of this vehicle given the measure X. X' and X have the same distribution when M is large enough. The vector D can be written as $Q = \{(d_i, d_{i+1}): 0 \le i \le M-2\}$. Therefore, the joint entropy of $X'$and X can be computed as:

$$H(X', X) = \sum_{(x',x)\in Q} P(x', x)\log_2\frac{1}{P(x',x)}, \quad (3)$$

where $P(x', x)$is the number of times (x', x)appearing in Q divided by the total number of elements in Q. With $H(X)$ and $H(X', X)$, the conditional entropy of X given X' is:

$$H(X|X') = H(X', X) - H(X') = H(X', X) - H(X). \quad (4)$$

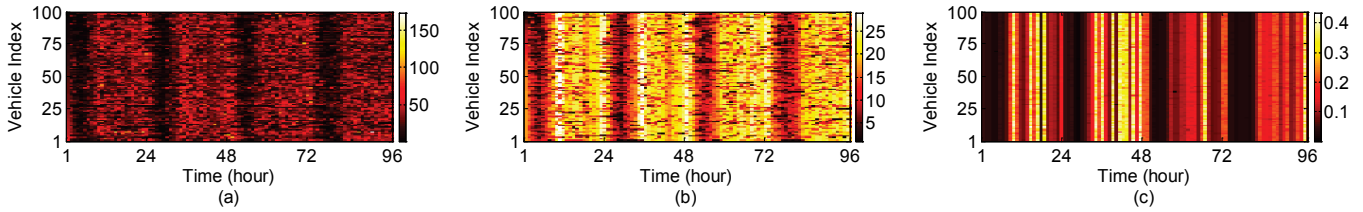

Figure 4. Time series of (a) degrees, (b) coreness and (c) closeness of one hundred randomly-chosen vehicles using Shanghai taxi trace; first column in each plot is sorted and placed in the descending order

When $K = 2$, let X'' denote the random variable representing the distribution of the previous two measures given X. Similarly, we can compute the conditional entropy $H(X|X'')$ as:

$$H(X|X'') = H(X'', X) - H(X'') = H(X'', X) - H(X', X). \quad (5)$$

The cumulative distribution functions (CDF) of the entropy and the conditional entropy of degrees for $K = 1, 2$ and 3 over all vehicles in Shanghai taxi trace is shown in Fig. 5, using one hour to divide the trace and construct contact graphs. It can be seen that the conditional entropy when $K = 1$ is much smaller than the marginal entropy and the conditional entropy when $K = 2$ is much smaller than when $K = 1$, which implies that the uncertainty about the current degree decreases when knowing the last degree of the same vehicle. We also show the mean entropy and the mean conditional entropy of all centrality metrics in Fig. 6 and have similar results. In summary, we conclude that the centrality of vehicles has strong temporal correlation. Generally, the more historical centrality information we know, the less uncertainty the current measure has.

C. Characterizing the Evolution of Vehicular Sociality

In order to characterize how vehicular sociality evolves along time, for a contact graph extracted in time slot t, denoted as $\mathcal{G}_t$, we study the distribution of contacts with other vehicles of each vehicle, and examine the correlation between the distribution in time slot t and that in time slot $t - n$, increasing n from one to a large number. We use *redundancy* to quantify the correlation.

In specific, the contacts between a vehicle v_i and all the other vehicles in time slot t forms a contact vector $C_t = (c_1, c_2, \cdots, c_{|N|})$, where $|N|$ is the total number of vehicles in $\mathcal{G}_t$ and c_j is the number of contacts that vehicle v_i has met with vehicle v_j in time slot t for $j = 1, \cdots, |N|$, where $c_j = 0$ if $i = j$. We also have the contact vector in time slot $t - n$, C_{t-n}. We compute the mutual information of C_t and C_{t-n}, $I(C_t, C_{t-n})$ via the joint entropy $H(C_t, C_{t-n})$ and the marginal entropy $H(C_t)$ and $H(C_{t-n})$ as follows:

$$I(C_t, C_{t-n}) = H(C_t) + H(C_{t-n}) - H(C_t, C_{t-n}). \quad (6)$$

We define the redundancy of C_t and C_{t-n} by

$$R(C_t, C_{t-n}) = \frac{I(C_t, C_{t-n})}{H(C_t) + H(C_{t-n})}. \quad (7)$$

We compute the mean redundancy averaged over all vehicles in Shanghai taxi and bus data sets. Time is divided into time slots of four hours. Fig. 7 shows the result for n = 1 to 120 (a period of one month). It can be seen that the layout of the contact relationship of vehicles has clear periodicities, i.e., a period of one day for buses and a period of two days for taxies. This might reflect the different shift rules of buses and taxies. In Shanghai, each bus is assigned to commute on a fixed route on every day. In contrast, taxi drivers usually shift every 24 hours so a taxi behaves very differently on every day but very similarly on every other day. In addition, buses have higher redundancy than taxies, which implies that contact relationship between buses is more predictable.

To better understand how much history data should be considered in capturing the vehicular sociality dynamics, we examine the redundancy between the layout of the contact vector in time slot t and the aggregated historical contact information from $t - 1$ to $t - n$, i.e., $\sum_{i=1}^{n} C_{t-i}$. The average redundancy over Shanghai taxies and buses is also shown in Fig. 7. It is clear that the redundancy increases as the amount of history data increases and tends to stabilize. This implies that history information of four weeks should be sufficient for capturing vehicular sociality temporal patterns.

VI. Design of Post

A. Overview

From our analysis above, we learn the knowledge that vehicle networks show clear social structures when aggregating pairwise contacts and the vehicular sociality (particularly, the degree and coreness centrality) is highly dynamic but also demonstrates strong temporal patterns and correlations. This inspires our POST scheme to utilize the temporal correlations of vehicular sociality to infer the contact behavior of the whole network in future, which can be further leveraged to improve the performance of mobile advertising. More specifically, POST first uses Markov chains to capture historical temporal correlations of vehicular centrality and predict the expected centrality of each vehicle in the short future. With this information, POST adopts a greedy strategy to select the most "centric" or important vehicles to serve as seeds for mobile advertising.

We elaborate the three integrated components of POST in the rest of this section.

B. Capturing the Temporal Correlations of Centrality

As the class of finite-state Markov processes (Markov chain models) is rich enough to capture a large variety of temporal dependencies, we adopt Markov chains of k^{th} order for capturing the temporal correlations of vehicular centrality. In Markov chain models, the current state of the process depends only on a certain number of previous values of the process, which is the order of the process.

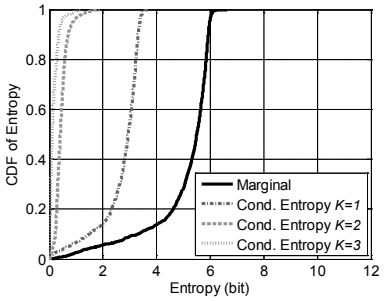

Figure 5. CDF of entropies of degree centrality

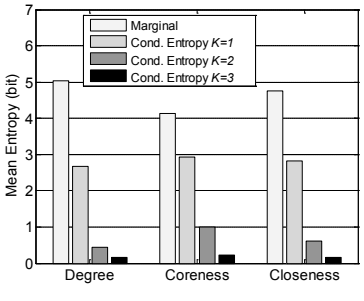

Figure 6. Average entropies of centrality measures

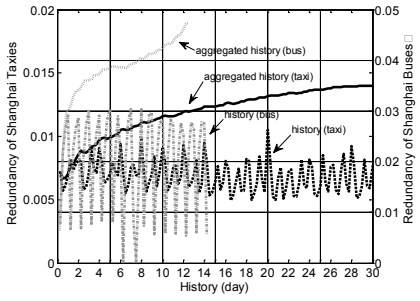

Figure 7. Mean redundancy of contact layout

Specifically, given the time requirement of a mobile advertising problem T, in order to capture the centrality dynamics at the time scale of the problem in the network, we divide time into slot of equal length T. Clearly, when T is relatively short, then more network dynamics can be seen between consecutive time slots but at the same time more random factors would involve in the observations which makes it hard to capture the temporal correlations of vehicular sociality and infer future network behavior. In contrast, when T is long, the topologies of the extracted graphs tend to be stable and easy to estimate but also lose most of the vehicular dynamics. We will extensively study the impact of the problem scale to the performance of mobile advertising in the performance evaluation section.

Given a time series of contact graphs, for each vehicle v_i, we measure the centrality of each vehicle in each contact graph and get a sequence of centrality measures, denoted as $\{x_i\}_{i=1}^{n}$. By discretizing continuous measures, we can get a finite state space, denoted as $\mathcal{S}$. The k-order state transition probabilities of the Markov chain can be estimated for all $a \in \mathcal{S}$ and $\underline{b} \in \mathcal{S}^k$, $\underline{b} = (b_1, b_2, \cdots, b_k)$ as follows. Let $n_{\underline{b}a}$ be the number of times that state $\underline{b}$ is followed by value a in the sequence. Let $n_{\underline{b}}$ be the number of times that state $\underline{b}$ is seen and let $p_{\underline{b};a}$ denote the estimation of the state transition probability from state $\underline{b}$ to state $(b_2, \cdots, b_k, a)$. The maximum likelihood estimators of the state transition probabilities of the k^{th} order Markov chain are:

$$p_{\underline{b};a} = \begin{cases} n_{\underline{b}a}/n_{\underline{b}}, & if\ n_{\underline{b}} > 0 \\ 0, & otherwise \end{cases}. \quad (8)$$

C. Inferring Future Vehicular Centrality

As the network is studied at the time scale of the mobile advertising problem, we only need to estimate the centrality of all vehicles in the next time slot with the established Markov chains.

In specific, let $\underline{b_i}$ denote the current state in the k^{th} order Markov chain built for v_i. The estimated centrality of the vehicle in the next time slot $\mathcal{E}^i_{centrality}$ can be calculated as:

$$\mathcal{E}^i_{centrality} = \sum_{a \in \mathcal{S}} p_{\underline{b_i};a} \cdot a. \quad (9)$$

In the process of inferring the future centrality for a node, two key parameters, i.e. the order of a Markov model and the length of historical data used to train the model, are critical to the accuracy of estimations. For a Markov chain model with a set of known states, simply increasing k can not necessarily fit for the temporal correlation included in a time series. The order of Markov chain can be evaluated via information content tests such as AIC and BIC [26].

D. Selecting Preferable Seeds for Mobile Advertising

As the mobile advertising problem with known future information is still NP-hard as proved in Section III, we adopt greedy heuristics to select preferable seeds.

The greedy heuristic of POST is to rank all the vehicles according to their predicted centrality values and take the top B/p "centric" vehicles as seeds, where B is the budget and p is the price for employing one vehicle.

VII. Performance Evaluation

A. Methodology

In this section, we evaluate POST scheme and compare with several alternative schemes. As the primary concern of mobile advertising is the information propagation speed and coverage, we assume that vehicles have infinitely large memory and bandwidth, and messages can always be successfully transferred between vehicles. In all the schemes, only seeds actively send advertisements to other vehicles via vehicle-to-vehicle communications.

We compare POST with several alternative seed selection schemes as follows:

- **Random.** In this scheme, seed vehicles are randomly chosen for advertising. This scheme is simple and needs no extra information in selecting seeds.
- **Static.** In this scheme [15] [16], seeds are chosen based on the rank of network position of all the vehicles in a static contact graph, which is extracted using traces of a long period of time, e.g., all available traces.

We consider the following two metrics to evaluate the effect of our solution and the above schemes.

1) ***Coverage ratio.*** It refers to the ratio of the number of successfully posted vehicles (i.e., who have already seen the advertisement) to the total number of vehicles in the network at the end of the advertising period T. This metric is used to measure the gained coverage of an advertisement by a given set of seeds.

2) ***Average impact.*** It refers to the average number of times a vehicle has seen the advertisement. The purpose of this metric is to reflect the gained intensity of an advertisement by a given set of seeds. It can be calculated as the ratio of the total number of times that one of the seeds has

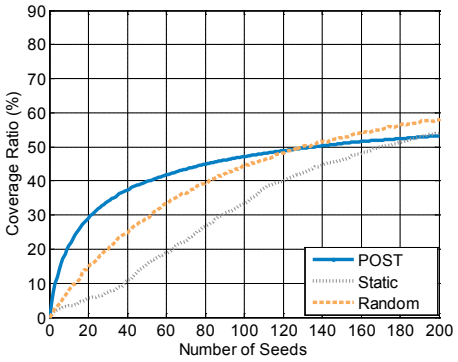

Figure 8. Coverage ratio vs. the number of seeds under $T = 1$h.

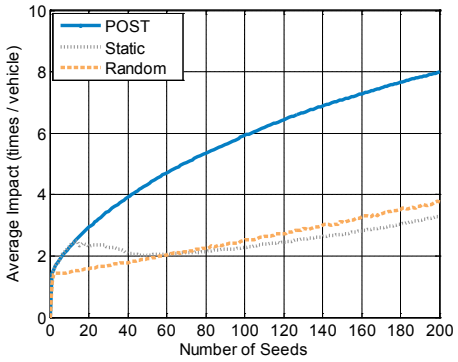

Figure 9. Average impact vs. the number of seeds under $T = 1$h.

posted the advertisement to another vehicle to the number of posted vehicles at the end of the advertising period T.

In the following simulations, we evaluate the above metrics of POST and all alternative schemes, using real trace data of Shanghai taxies for demonstration. We use the contact records in the whole February of 2007 as the learning stage for all alternative schemes and use the trace of three days in March for testing. In order to investigate different time scales of the mobile advertising problem, we examine the performance of all schemes under three different time scales, i.e., $T = 1$ hour, $T = 6$ hours and $T = 12$ hours. Under each time scale, we divide the trace and construct contact graphs accordingly. The order of Markov chains is set to five when $T = 1$ hour and set to two when $T = 6$ hours and $T = 12$ hours based on the results of AIC and BIC tests.

B. Performance Comparison

In this experiment, we compare POST with all the other alternative seed selection schemes under the time scale of one hour. We change the quota of seeds from one to two hundred and run all schemes over all available testing trace and get the average.

Fig. 8 plots the average coverage ratio as a function of the number of seeds. It can be seen that POST outwits the corresponding static scheme and the random scheme. In particular, when the budget is limited (i.e., only a small number of vehicles can be employed), POST has the best performance comparing with the Static scheme and the Random scheme. Interestingly, it can also be seen that when the number of hired vehicles increases, the coverage performance of all three schemes also increases but the speed of such increment of POST tends to decrease. For example, recruiting the first 20 seeds can achieve a coverage ratio of about 30% but adding another 20 seeds can only bring an extra 8% coverage ratio. In contrast, both Random and Static have better coverage than POST when the number of seed vehicles exceeds 120 and 190, respectively. The reason is because POST always selects those top vehicles with respect to centrality metric values as seed vehicles. As the number of seed vehicles increases, the probability of two selected vehicles having similar network positions also increases, which means they are redundant and may not be able to gain as large coverage as possible. On the contrary, the Random scheme has larger probability to uniformly scatter seed vehicles among all communities and therefore can achieve better coverage when the number of seed vehicles is large. The Static scheme have similar results as the static network structure is quite different from the current network topology.

Fig. 9 plots the average impact as a function of the number of seeds. It can be seen that POST extraordinarily outperforms all the other schemes. This implies that POST can achieve very good intensity of advertising, which may be more preferable under certain circumstances (e.g., expecting the advertisement to be remembered longer). The reason is that POST tends to select "centric" vehicles with more coverage overlap as seeds so that the intensity of an advertisement is enhanced as explained above.

C. Impact of Time Scale of Mobile Advertising

We further examine the impact of time scale of mobile advertising to the performance of POST. We set the seed quota to 40 and run all schemes using testing trace and get the average.

Fig. 10 and Fig. 11 plots the coverage ratio and average impact as a function of time scale of the mobile advertising problem, respectively. It can be seen that in both figures, as the time scale increases, POST always achieves best performance but the performance differences between POST and Static schemes also decrease. This verifies our discussion about the relation between network dynamics and the time scale of the problem. In general, as the time scale increases, the network dynamics tend to vanish and the network structure also tends to stable even though the underlying vehicles are mobile.

VIII. Conclusion and Future Work

In this paper, we have studied the mobile advertising problem in vehicular networks and proved its NP-hardness. By analyzing three large-scale vehicular traces, we have found that the vehicles show clear sociality within the network and the vehicular sociality is highly dynamic and has strong temporal correlations. Based on the observations, we have proposed to use Markov chains to capture the patterned vehicular centrality and infer the expected future. We have also employed one greedy heuristic to utilize the estimated centrality information to improve the performance of mobile advertising. We have demonstrated the efficacy of our method via extensive trace-driven simulations. Overall, POST can achieve best performance against the state-of-art algorithms. For our future, we intend to involve more types of vehicle traces into the

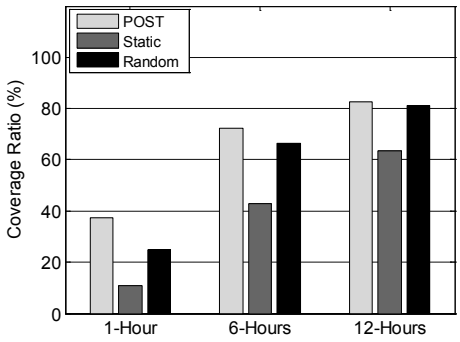

Figure 10. Coverage ratios vs. time scales of the problem

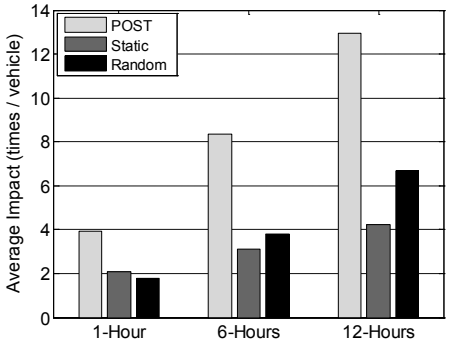

Figure 11. Average impact vs. time scales of the problem

network. Furthermore, we will further study the best time scale to observe the vehicular networks so that the inherent dynamics would not be under- or over-estimated. The result can be used to further improve POST.

ACKNOWLEDGMENT

This research was supported in part by National Natural Science Foundation of China (Grants No. 61202375, 61373157, 61173171, 61170237, 61170238, 60903190), the National High Technology Research and Development Program (2013AA01A601), the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2012J072), Program for Changjiang Scholars and Innovative Research Team in Universities of China (IRT1158, PCSIRT), Science and Technology Commission of Shanghai Municipality (Grant No.12ZR1414900).

REFERENCES

[1] Kempe, David, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network," in Proceeding of ACM SIGKDD, Washington, DC, USA, 2003.

[2] Chen, Wei, Yajun Wang, and Siyu Yang. "Efficient influence maximization in social networks," in Proceeding of ACM KDD, Paris, France, 2009.

[3] Chen, Wei, Wei Lu, and Ning Zhang. "Time-critical influence maximization in social networks with time-delayed diffusion process," in Proceeding of AAAI, 2012.

[4] Wischhof, Lars, André Ebner, and Hermann Rohling. "Information dissemination in self-organizing intervehicle networks," *IEEE Transactions on Intelligent Transportation Systems,* vol. 6(1), pp. 90-101, 2005.

[5] Kosch, Timo, Christian Schwingenschlogl, and Li Ai. "Information dissemination in multihop inter-vehicle networks," in Proceeding of IEEE ITSC, 2002.

[6] Naumov, Valery, and Thomas R. Gross. "Connectivity-aware routing (CAR) in vehicular ad-hoc networks," in Proceeding of IEEE INFOCOM, Anchorage, Alaska, USA, 2007.

[7] Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6(11), pp. 888-893, 2010.

[8] Ugander, J., Backstrom, L., Marlow, C., & Kleinberg, J. "Structural diversity in social contagion," *PNAS*, vol. 109(16), pp. 5962-5966, 2012.

[9] Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguná, M., & Krioukov, D. "Popularity versus similarity in growing networks," *Nature*, vol. 489(7417), pp. 537-540, 2012.

[10] Richardson, Matthew, and Pedro Domingos. "Mining knowledge-sharing sites for viral marketing," in Proceeding of ACM SIGKDD, Edmonton, Alberta, Canada, 2002.

[11] Domingos, Pedro, and Matt Richardson. "Mining the network value of customers," in Proceeding of ACM SIGKDD, San Francisco, CA, USA, 2001.

[12] Kempe, David, Jon Kleinberg, and Éva Tardos. "Influential nodes in a diffusion model for social networks," *in Proceedings of ICALP,* Springer Berlin Heidelberg, pp. 1127-1138, 2005.

[13] Karsai, M., Kivelä, M., Pan, R. K., Kaski, K., Kertész, J., Barabási, A. L., & Saramäki, J. "Small but slow world: How network topology and burstiness slow down spreading," *Physical Review E*, vol. 83(2), 2011.

[14] Daly, Elizabeth M., and Mads Haahr. "Social network analysis for routing in disconnected delay-tolerant manets," in Proceeding of ACM MOBIHOC, Montréal, Québec, Canada, 2007.

[15] P. Hui, J. Crowcroft, and E. Yoneki. "Bubble Rap: social-based forwarding in delay tolerant networks," in Proceeding of ACM MOBIHOC, Hong Kong, China, 2008.

[16] Boldrini, Chiara, Marco Conti, and Andrea Passarella. "ContentPlace: social-aware data dissemination in opportunistic networks," in Proceeding of ACM MSWiM, Vancouver, BC, Canada, 2008.

[17] Micek, Juraj, and Ján Kapitulík. "Car-to-car communication system," in Proceeding of IEEE IMCSIT, 2009.

[18] Feige, Uriel. "A threshold of ln n for approximating set cover," *Journal of the ACM (JACM)*, vol. 45(4), pp. 634-652, 1998.

[19] H. Dubois-Ferriere, M. Grossglauser, and M. Vetterli, "Age matters: efficient route discovery in mobile ad hoc networks using encounter ages," in Proceeding of ACM MOBIHOC, 2003.

[20] H. Zhu, M. Dong, S. Chang, Y. Zhu, M. Li, and X. Shen, "ZOOM: scaling the mobility for fast opportunistic forwarding in vehicular networks," in Proceeding of IEEE INFOCOM, Turin, Italy, 2013.

[21] V. D. Blondel, J. L. Guillaune, R. Lanbiotte, and E. Lefebvre, "Fast unfolding the communities in large networks," *J. STAT. MECH.*, 2008.

[22] M. E. J. Newman, "Modularity and community structure in networks," *PNAS*, 2006.

[23] Freeman, Linton C. "Centrality in social networks conceptual clarification," *Social networks*, vol. 1(3), pp. 215-239, 1979.

[24] Miorandi, Daniele, and Francesco De Pellegrini. "K-shell decomposition for dynamic complex networks," in Proceeding of IEEE WiOpt, Avignon, France, 2010.

[25] Amaral, L. A. N., Scala, A., Barthélémy, M., & Stanley, H. E. "Classes of small-world networks," *PNAS,* 2000.

[26] Burnham, Kenneth P., and David R. Anderson. "Multimodel inference understanding AIC and BIC in model selection," *Sociological methods & research*, vol. 33(2), pp. 261-304, 2004.