A Tale of Two Cities - Characterizing Social Community Structures of Fleet Vehicles for Modeling V2V Information Dissemination

Fan Bai†, Keyvan Rezaei Moghadam‡, Bhaskar Krishnamachari‡
† General Motors Research Center, ‡ University of Southern California

***Abstract*—We study the presence of social communities in mobility traces from vehicular fleets. By analyzing publicly available sets of fleet vehicle mobility traces obtained from two real-world deployments – consisting of more than 2000 taxis in Shanghai and Beijing respectively, we confirm the existence of small numbers of distinct social communities in vehicular networks, which is in direct contrast to the general belief that vehicular networks are best modeled as a relatively homogeneous system. We examine the spatio-temporal characteristics of social communities, gaining the insight that they are driven primarily by social proximity induced by geographic locality. We then develop a parsimonious multi-community ordinary differential equation (ODE) model, which uses the heterogeneous structure introduced by social communities to model information dissemination. We show through simulations that this approach dramatically outperforms the conventional homogeneous ODE model in capturing the dynamics of the dissemination process. We further demonstrate that the use of the ODE model to optimize seeding of an initial set of vehicles results in improved utility for information dissemination compared to seed-optimization using a homogeneous model.**

I. Introduction

The U.S. National Highway Traffic Safety Administration recently announced that it is to begin taking steps to enable the deployment of vehicle-to-vehicle (V2V) technology [1]. It is of great interest to be able to model, tractably and with high-fidelity, the dissemination of information about traffic and road-conditions across such a vehicular network.

The prevailing methodology for modeling information dissemination in vehicular networks has been generally to consider the entire system as a homogeneous network, assuming that any two vehicles are equally likely to meet each other. On the other hand, it is possible to conceive an alternative approach in which each pair of nodes is considered to have a different encounter rate. However, modeling the dissemination process over such a extremely fine-grained model becomes rapidly intractable as the number of possible information states (*i.e.*, which cars have a given piece of information at some time) would scale exponentially with the number of cars. Thus, we seek a *parsimonious* model that can realistically capture the heterogeneity of interactions but with relatively few parameters so that it remains tractable. We show in this paper that such a parsimonious model of information dissemination for vehicular networks can be built using social communities.

A number of studies in recent years have applied social network analysis to general opportunistic mobile networks to understand how various characteristics such as the presence of communities, hubs, and bridges can be used to design efficient protocols; however, virtually all of these have focused primarily on human mobility traces [5] [6]. Somewhat to our surprise, there are very few studies which investigate the social community structure of vehicular networks. We analyze two publicly available sets of fleet vehicle mobility traces – more than 2000 taxis in Shanghai and Beijing respectively. We confirm the existence of a relatively small number of distinct social communities in vehicular networks, which is in direct contrast to a widely-held, albeit implicit, belief that encounters in vehicular fleets such as taxis are likely to be always homogeneous in nature. We further examine the spatio-temporal characteristics of social communities to analyze their root causes. Finally, we consider how to incorporate social communities to enhance the modeling of information dissemination in vehicular fleets.

The contributions of this paper are as follows:

1) Using large-scale empirical measurement traces, our study is the first to confirm the existence of social community structure in the context of vehicular fleets. Our study reveals two unique characteristics of social communities in vehicular networks: (a) the community structure arises primarily due to geographic correlations; (b) We find that the social communities in the taxi fleets are relatively stable over time.
2) Building on the discovery of a relatively small number of distinct yet stable social communities and their unique characteristics in vehicular networks, we develop a multi-community Ordinary Differential Equation (ODE) model, which takes advantage of heterogeneous structure of social communities to model the information dissemination process in vehicular fleets. This model is parsimonious because of the small number of communities detected.
3) Via extensive simulation, we show that the multi-community ODE model offers a high-fidelity modeling of the information dissemination process, particularly when it comes to sufficiently large communities, significantly outperforming the conventional homogeneous ODE model which does not take community structure into account.

4) We further use our proposed multi-community ODE model to optimize the number of initial seeds and show that it improves the utility of information dissemination compared to the optimal seeding based on a homogeneous model.

II. Related Works

This piece of work is inspired and informed by a number of other pioneering works.

Pairwise Contact Analysis in DTN. By studying empirical human mobility traces, a number of research works analyze several metrics of human contacts, including node degree, contact duration and inter-encounter interval [2]. In particular, the probabilistic distribution of inter-encounter interval receives a great deal of attention. The distribution of inter-contact time between walking pedestrians was reported to follow a power-law distribution, in direct contrast to the commonly used assumption of exponential decay [4]. Unlike these prior works which focus on pairwise relationship between mobile nodes, our study is more interested in understanding the behaviors of social communities in vehicular networks.

Social Network Analysis. Departing from pairwise contact analysis, social network analysis has gained momentum in the past few years. Exploiting often ignored social network structure of delay-tolerant networks, SimBet [5] and Bubble Rap [6], utilize a number of social-based metrics (e.g., centrality, betweenness, and similarity) to guide the opportunistic forwarding decisions. For instance, SimBet [5] utilizes the metric of similarity and betweenness centrality to determine the role of nodes in the hidden architecture of social networks; Bubble Rap [6] uses between centrality to identify the bridging nodes between different social communities and uses the metric of centrality to gradually find better nodes to relay information within a given community.

Apart from leveraging metrics of social behaviors, several studies further investigate the behaviors of social communities so that the structure of social community could be used in the context of delay-tolerant networks. Using Kalman filter as a forecasting technique, CAR [8] predicts the future evolution of nodal mobility based on social behaviors (e.g., co-location of nodes in the same communities) in order to guide the decision of packet forwarding. Social community structure is explicitly used in Socio-aware Publish/Subscribe framework [7] to establish a two-tier overlay network architecture. To our best knowledge, our work, which focuses on larger taxi fleets from two big cities, is the first of its kind to demonstrate the existence of social community structure in vehicular network fleets. However, not surprisingly, we do find that the community structure in vehicular networks is different from their counterpart in pedestrian network, because of the higher speeds and greater spatial coverage of vehicles.

Community Detection. Community detection algorithms help to identify the local community structure inherent to a networked system. Literature in theoretical network science reveals that finding the optimal allocation of nodes to social communities is a computationally challenging task. A rich set of methods were developed to detect social communities in a cost-efficient manner. In specific, among other solutions, several mainstream methodologies commonly used all adopt heuristic approach – Louvain [9] and k-clique [10], mainly because these heuristic approaches are easy to implement. In our study, we leverage these works (in particular, Louvain algorithm) to understand the structure of social communities in vehicular networks and their behaviors.

Information Dissemination. Virus propagation and contamination, in wired networks, could be modeled as spreading process of infectious disease [11]. Recent studies shows that, in mobile networks, disease modeling could be applied to information dissemination process as well [12]. These efforts use stochastic models (i.e., Markov Chain) or their deterministic approximation (i.e., fluid model) to analyze the performance of encounter-based information propagation process [12]. However, all these studies assume that nodal encounter process is of homogeneous nature, and do not take social community structure into account. Our work differs from these works as follows: Realizing that inherent social community structure indeed exists in vehicular networks, we propose a multi-community Ordinary Differentiation Equation (ODE) model to study fairly complicated information dissemination process; our extensive simulations show that our model taking social community structure into account significantly outperforms other models that do not.

III. Data Sets

Data Collection. We choose a set of taxi traces as our first step to understand the community structure of encounter patterns among fleet vehicles. Our primary set was collected from Shanghai, China on January 31, 2007 - February 27, 2007 (1 month), and composed of over 2,439 taxis covering 6,340 km^2 area providing regular GPS data. A secondary set we use in our analysis and validation is of Taxis from Beijing, China, and consists of 2,721 nodes examined over two weeks, May 1 - May 14, 2009.

Data Processing. The logged data in the data sets includes (1) the vehicle's ID, (2) the current time-stamp, (3) the longitude and latitude coordinates of the vehicle's current position, (4) the current speed and heading of the vehicle, and (5) the occupancy status of the taxi. Due to the cost associated with cellular communication, each taxi could only afford to report its mobility trajectory every 15-60 seconds. Using a linear interpolation technique, we convert a set of coarse-granularity mobility trace into a set of fine-granularity mobility trace (e.g., an update of every second). Using this fine-granularity trace, we could assume that two vehicles are in direct contact if their distance is less than or equal to a parameter r (we use $r = 300m$ here in order to match the typical range of a DSRC/WiFi radio).

We acknowledge that some of the assumptions that we have applied here may introduce certain inaccuracies. On one hand, using a simple linear interpolation technique (without

TABLE I
DETECTED SOCIAL COMMUNITIES IN SHANGHAI TAXI TRACE (JAN 31– FEB 6, 2007 (MON - SUN), 1 WEEK)

	Mon	Tue	Wed	Thur	Fri	Sat	Sun
Num. of Vehicles	2292	2308	2319	2309	2299	2294	2286
Num. of Edges	2293726	2374790	2461384	2285386	1981036	2223864	2284842
Q_{max}	0.08984	0.08623	0.08016	0.09425	0.1107	0.09062	0.08791
Num. of Communities	4	7	10	5	8	7	4
Size of Major Communities	924, 811, 556	886, 885, 583	925, 813, 574	942, 813, 552	890, 849, 555	940, 790, 560	899, 809, 577

referring to a map), mobility trajectory of a vehicle is not fully correctly interpolated. In addition, we acknowledge that a disk range model that ignores realistic radio propagation simplifies the vehicle encounter process. Despite these inaccuracies, we believe our analysis of these traces, which are a valuable source of data covering thousands of vehicles over a long time period, are able to reveal the first-order characteristics of vehicle encounter process and its social community structure in real-world environments.

IV. SOCIAL COMMUNITY STRUCTURE IN VEHICULAR NETWORKS

Several sets of human mobility traces had been examined to uncover the community structure among people. To our best knowledge, nonetheless, knowledge about social community structure in vehicular networks is fairly limited. Our study aims to bridge this gap.

A. Social Community Detection Algorithm

Contact Graph. The sequence of actual contacts over time is mapped into a conceptual contact graph, in which the weight of link indicates the strength of relationship between two nodes. Mathematically, the entire vehicle encounter sequence is aggregated into a static contact graph $G(N, V)$, where N is number of vehicles and weighted matrix $V = \{v_{i,j}\}$ represents the strength of the relationship between vehicle i and vehicle j. In our study, we focus on unweighted contact graph such that $v_{i,j} = 1$ if vehicle i and vehicle j ever encounter once in the mobility trace; otherwise, $v_{i,j} = 0$.

Louvian Algorithm. To identify the existence of social community structure in vehicular networks, we apply a well-known Louvian community detection algorithm to the contact graph.

Neuman Modularity is a metric to measure the fitness of detected communities [3]. Neuman Modularity directly compares the fraction of links in a graph that connect nodes within particular social communities with the fraction of links in a graph whose links follow a random distribution, while the node degree of both scenarios are kept the same. The mathematical representation of Neuman Modularity is given as

$$Q = \frac{1}{2m} \sum_{i,j} (A_{i,j} - \frac{d_i d_j}{2m}) \delta(C_i, C_j) \quad (1)$$

where $A_{i,j}$ is the weight of the link between node i and node j if this link does exist (otherwise, $A_{i,j} = 0$), $d_i = \sum_k A_{i,k}$ is the node degree of node i (similarly, d_j is the node degree of node j), and $m = \frac{1}{2} \sum_i d_i$ is the total weight of the entire network. C_i (or C_j) represents the community that node i (or node j) belongs to, and Kronecker delta function $\delta(C_i, C_j)$ is 1 if node i and node j belong to the same community, and 0 otherwise. $Q = 0$ indicates that network graph is a perfect random graph, and a nonzero Q value indicates the existence of social community.

Louvain algorithm [9] offers a heuristic solution that approximates the optimal allocation of nodes to different communities, so that the Neuman modularity of the whole graph could be maximized. The simple approach of Louvain algorithm reduces the computational complexity of finding the theoretically optimal solution on one hand, but still satisfies the accuracy of community detection on the other hand. Results obtained from Louvian algorithm are shown to be as good as those from other community partition algorithms [9]. Because of these merits, we adopt Louvain algorithm in our study.

B. Social Communities in Vehicular Networks

Table I summarizes the community structures detected by Louvain algorithm in the Shanghai taxi trace. We observe that the Q value of Shanghai taxi trace is in the range between 0.080 and 0.110 (similar values are observed for the Beijing trace). Albeit weak, this shows that there is inherent social community structure as vehicles encounter each other.

We further look into the number of social communities. We find that the number of social communities in Shanghai taxi traces varies from 4 (on Monday) to 10 (on Wednesday) over a week. Interestingly, among them, there are 3 constantly well-connected communities that have about 900, 800 and 600 vehicles across the entire week, while the remaining communities only have a single vehicle. The analysis of the Beijing trace similarly yields four major communities. It is striking that the taxi fleet traces naturally decompose into such small numbers of distinct communities, indicating that these communities may be useful from a parsimonious modeling perspective, as we shall see is indeed the case.

V. CHARACTERISTICS OF SOCIAL COMMUNITY IN VEHICULAR NETWORKS

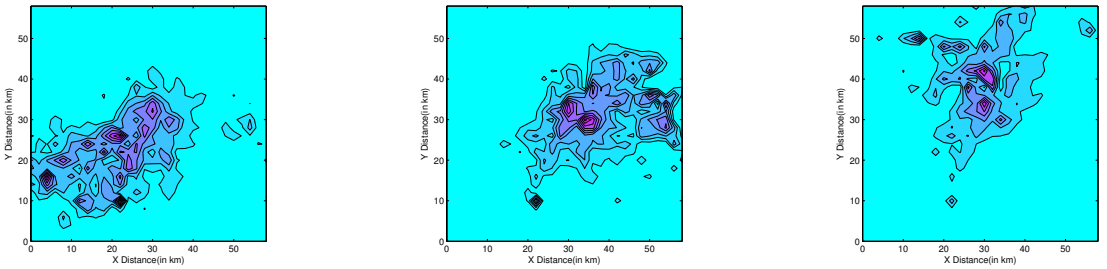
In this section, we study several characteristics of social communities in vehicular networks. Moreover, we identify that their root cause is the *social proximity behavior* of vehicles.

A. Temporal Correlation of Social Communities

We observe that there are always 3 communities in Shanghai taxi trace across different days; this observation motivates us

TABLE II
CORRELATION BETWEEN BEST-MATCHED SOCIAL COMMUNITIES ACROSS A WEEK IN SHANGHAI TAXI TRACE

	Mon	Tue	Wed	Thur	Fri	Sat	Sun
Mon	1, 1, 1	0.72, 0.68, 0.72	0.75, 0.71, 0.74	0.74, 0.71, 0.73	0.75, 0.75, 0.75	0.69, 0.70, 0.71	0.75, 0.74, 0.71
Tue		1, 1, 1	0.69, 0.69, 0.72	0.74, 0.71, 0.75	0.72, 0.70, 0.73	0.73, 0.73, 0.72	0.72, 0.71, 0.71
Wed			1, 1, 1	0.73, 0.72, 0.71	0.75, 0.73, 0.74	0.69, 0.69, 0.69	0.72, 0.73, 0.74
Thur				1, 1, 1	0.74, 0.73, 0.72	0.77, 0.74, 0.75	0.70, 0.72, 0.68
Fri					1, 1, 1	0.70, 0.72, 0.71	0.73, 0.72, 0.74
Sat						1, 1, 1	0.69, 0.70, 0.71
Sun							1, 1, 1



(a) Social Community 1 (b) Social Community 2 (c) Social Community 3

Fig. 1. The Geographic Coverage of 3 Social Communities Detected in Shanghai Taxi Trace (over 1 Week).

to study the temporal correlation between corresponding social communities on different days. We define a correlation metric

$$Sim(C_i(m), C_j(n)) = \frac{|C_i(m) \cap C_j(n)|}{|C_i(m) \cup C_j(n)|} \quad (2)$$

where $C_i(m)$ and $C_j(n)$ represent the member set of social community C_i on m-th day and the member set of social community C_j on n-th day, respectively. $|C_i(m) \cap C_j(n)|$ is the size of overlapping member set that belongs to both social communities, and $|C_i(m) \cup C_j(n)|$ is the size of overall member set which is the union of both social communities. This metric reflects the portion of overlapping members of these two communities to all the members in these two communities; the higher the value is, the stronger similarity these two social communities exhibit.

In Shanghai taxi case, Table II shows the correlation of corresponding social communities across different days. Each entry of this table represents a pair of different days (m-th day and n-th day), for each social community $C_i(m)$ on m-th day, among all other social communities on n-th day, we search and find its best matched social community $C_j(n)$ on that day. We find that all these 3 social communities have correlation value higher than 0.7 with their respective counterparts, suggesting that the membership composition of each social community is fairly stable and the members of social community does not change much over a week.

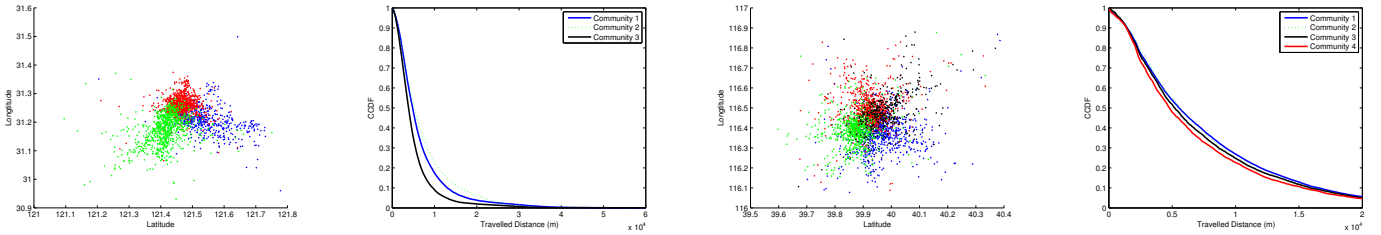
B. Geographic Concentration of Social Communities

We also look into the spatial distribution of social communities, by examining the geographic "coverage heatmap" of member vehicles belonging to a particular social community. First, we look into the mobility trajectory of a single vehicle in a given duration, and then assign weight to each geographic zone according to the dwell time of this vehicle at this geographic zone. By aggregating all the member vehicles of the same social community, we plot the geographic contour graph ("heatmap") of each social community. This geographic contour graph illustrates if social community tends to concentrate within specific geographic areas. Fig. 1 shows the geographic heatmap of social communities in Shanghai taxi trace within a week. It is observed that 3 social communities are distributed in 3 geographically disjointed areas, with each corresponding to one of urban activity centers of Shanghai city separated by the Huangpu River and the Wusong River.

C. Social Proximity Behaviors

It is clear from the above that the social communities of the Shanghai taxi fleet tend to travel and dwell in distinct geographic zones; we made very similar observations with the Beijing trace as well. We speculate that is rooted in *social proximity* behavior unique to humans and vehicles: a vehicle typically moves within a bounded region only, which relates to the social life of the driver (e.g., home and work), and is unlikely to move everywhere in the entire network equally likely.

We find that the mobility trajectory of Shanghai taxis is typically around certain social spots of Shanghai city. We plot the center point (i.e., home spot) of each taxi in Fig. 2(a), with vehicles from different communities labeled with different colors. It is clearly observed that the three social communities tend to have their own geographic hot spots, corresponding to different urban activity centers of Shanghai. These 3 community centers are 11 km, 14 km and 12 km away from each other. In Fig. 2(b), we plot the CCDF of taxi's traveled distance away from its home spot. We find that the probability of traveling more than 10 km away from a taxi's home spot is only about 10%-20% for all these three communities. Very



(a) Home Spot of Taxis in color (Shanghai)

(b) The CCDF of Travelled Distance (Shanghai)

(c) Home Spot of Taxis in color (Beijing)

(d) The CCDF of Traveled Distance (Beijing)

Fig. 2. Social Proximity Behaviors of Shanghai Taxis and Beijing Taxis. For both traces, we first plot the community-level geographic distribution of each taxi's home spot, and then we examine the CCDF of traveled distance away for each taxi's home spot.

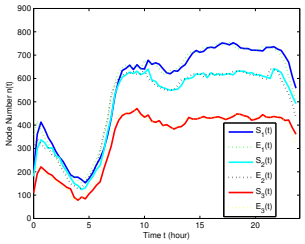


Fig. 3. The Number of Operating Vehicles in Each Community.

similar observations can be made from the results presented for the Beijing trace in Fig. 2(c) and Fig. 2(d), with the only difference being that there are four communities to consider in this city. We thus believe that this *social proximity* behavior is the fundamental root cause of the community structure we have observed using the social network analysis.

D. Vital Dynamics of Social Communities

Via our study, we also find that the size of social communities varies over time within a day, since different vehicles are switched on and off at different times. We call this as *vital dynamics*. We capture the vital system dynamic of 3 social communities collected from Shanghai taxi trace (Jan 31, 2007) in Fig. 3. For each time slot t, we plot the number of operating vehicles at its beginning ($S(t)$) and at its end ($E(t)$). We observe that, for all these 3 social communities, except a small portion of taxis serving the night shift, the majority of taxis were operating during normal working hours and early night (i.e., between 8am and 11pm). As later shown in Sec. VI, the system vital dynamics of social communities plays an important role in determining the information propagation speed in the vehicular delay-tolerant networks.

VI. Information Dissemination Process in Vehicular Networks

In this section, by taking the structure of social community into consideration, we develop a community-based Ordinary Differential Equation (ODE) model for information dissemination in vehicular networks. Our proposed community-based model is able to better capture the real-world heterogeneous nature of vehicular encounter process than conventional model.

A. Compartmental Model

Our theoretical framework is built upon the foundation of compartmental models in epidemiology; however, we take a further step to enhance generic compartmental models by taking into account the social community structure inherent to vehicular mobility patterns.

In this paper, we focus on two compartments: *Infected (I)* compartment (vehicles that have received a copy) and *Susceptible (S)* compartment (vehicles that have not received the copy yet). We only focus on a SI model in this study, but our analytical framework could also cover other sophisticated compartmental models such as SIR (Susceptible-Infected-Recovered) and SIS (Susceptible-Infected-Susceptible) models. Without loss of generality, we assume that, at time $t = 0$, there is only one infected node in the entire system ($S(0) = 1$ for homogeneous ODE model) or in each community ($s_i(0) = 1$ for heterogeneous ODE model).

We first start from a simple scenario in which no system vital dynamic is assumed. At time t, the state transition probability from $i(t)$ to $i(t+1)$ (or the state transition probability from $s(t)$ to $s(t+1)$) is determined by the $i(t)$, $s(t)$ and contact rate $\alpha(t)$ as follows:

$$\frac{di(t)}{dt} = \alpha(t)i(t)s(t) \quad (3)$$

$$\frac{ds(t)}{dt} = -\alpha(t)i(t)s(t) \quad (4)$$

This simple ODE model has been proposed [12] to model the performance of the Epidemic Routing protocol.

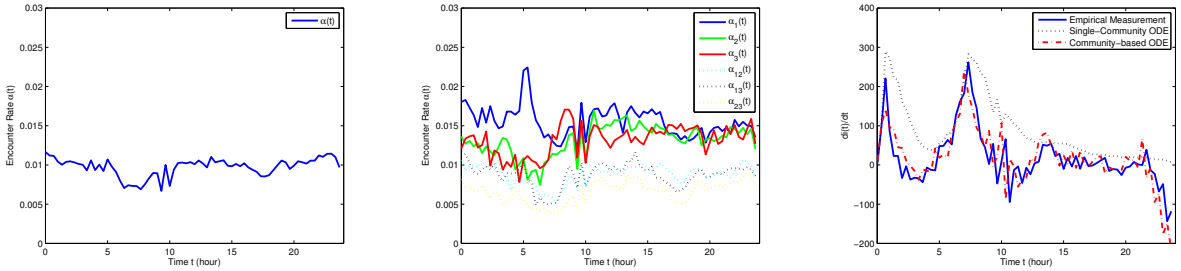
B. Social-Community Compartmental Model with Vital Dynamics

The generic compartmental model (Eqn.3 and Eqn.4) could only be applied to homogeneous mobility models. In addition, in generic compartmental model, $\frac{di(t)}{dt} + \frac{ds(t)}{dt} = 0$ since this model does not support vital system dynamics (i.e., birth-death process). Realizing these shortcomings, in this paper, our proposed model includes two aspects – *system vital dynamics* and *social community structure*, which had long been ignored in the literatures.

Compartmental Model with Vital Dynamics. Vital dynamics is often ignored in epidemiology. This is because (1) epidemic

TABLE III
SYMBOLS USED IN ANALYSIS

System-Level Variable	Community-Level Variable	Definition
C		The number of social communities in the system.
N	N_k	The total number of vehicles in the system (or in the k-th community).
$n(t)$	$n_k(t)$	The number of vehicles that are operating at time t.
$N-n(t)$	$N_k-n_k(t)$	The number of vehicles that are not operating at time t.
$I(t)$	$I_k(t)$	The total number of "infected" vehicles at time t.
$i(t)$	$i_k(t)$	The number of "infected" vehicles that are operating at time t
$I(t)-i(t)$	$I_k(t)-i_k(t)$	The number of "infected" vehicles that are not operating at time t.
$S(t)$	$S_k(t)$	The number of "susceptible" vehicles at time t.
$s(t)$	$s_k(t)$	The number of "susceptible" vehicles that are operating at time t.
$S(t)-s(t)$	$S(t)-s_k(t)$	The number of "susceptible" vehicles that are not operating at time t.
$B(t)$	$B_k(t)$	The number of "newly born" vehicles (operating at time t but not at time $t-1$)
$D(t)$	$D_k(t)$	The number of "newly dead" vehicles (those operating at time $t-1$ but not at t) at time t in the system (or in the k-th community) .
$\alpha(t)$		The average contact rate between any given pair of vehicles at time t.
	$\alpha_k(t)$	The average contact rate between vehicles both from k-th community.
	$\alpha_{(k,l)}(t)$	The average contact rate between vehicles from communities k and l



(a) Encounter Rate of Overall Population

(b) Intra- and Inter-Community Encounter Rate

(c) The Number of Operating Rate of Infected Nodes $di(t)/dt$

Fig. 4. The Encounter Rate Between Vehicles and the Rate of Change of Number of Operating Infected Nodes $i(t)$ Over 24 Hours in Shanghai Taxi Trace (Jan 31, 2007). For the case of $\frac{di(t)}{dt}$, empirical measurement from simulation and 2 prediction values using two different ODE models are plotted.

outbreak (in a few weeks) is usually far more rapid than the vital dynamics of human community (human population does not change significantly over a couple of years); and (2) the birth rate $B(t)$ and death rate $D(t)$ of human population tend to be equal ($|B(t)-D(t)| \approx 0$), leading to a constant population. As shown in Sec. V-D, nonetheless, these two observations made from human society become luxury assumptions when applied to our case in vehicular networks.

Lemma 1: We assume that there is significant system vital dynamic in the vehicle network ($|B(t)-D(t)| \gg 0$). At time $t=0$, as one message is propagated through epidemic routing protocols, the information contamination process only happens between both infected vehicles and susceptible vehicles that are operating (*i.e., alive*). The number of infected vehicles $I(t)$ and susceptible vehicles $S(t)$ evolves as follows:

$$\frac{dI(t)}{dt} = \alpha(t)i(t)[n(t)-i(t)] \quad (5)$$

$$\frac{dS(t)}{dt} = -\alpha(t)i(t)[n(t)-i(t)] \quad (6)$$

where $n(t)$ is the number of operating vehicles at time t. The number of infected operating vehicles $i(t)$ and susceptible operating vehicles $s(t)$ evolves as follows:

$$\frac{di(t)}{dt} = \alpha(t)i(t)[n(t)-i(t)] + \frac{I(t)-i(t)}{N-n(t)}B(t) - \frac{i(t)}{n(t)}D(t) \quad (7)$$

$$\frac{ds(t)}{dt} = -\alpha(t)i(t)(n(t)-i(t)) + [1-\frac{I(t)-i(t)}{N-n(t)}]B(t) - [1-\frac{i(t)}{n(t)}]D(t) \quad (8)$$

Social-Community Compartmental Model. A compartmental model built upon single community accurately characterizes the performance of Epidemic Routing in a homogeneous setting [12]. By taking more sophisticated social community structure into account, we extend compartmental model to incorporate the heterogeneous nature of vehicle networks.

Lemma 2: We assume that there are C social communities in a city and we do not assume vital system dynamic. At time $t=0$, one message starts from a randomly selected vehicle belonging to social community m and we propagate this message through epidemic routing process. For the social community k, the number of infected vehicles $I_k(t)$ and susceptible vehicles $S_k(t)$ evolves as follows:

$$\frac{dI_k(t)}{dt} = \sum_{j=1}^{C} \alpha_{(j,k)}(t) I_j(t) S_k(t) \quad (9)$$

$$\frac{dS_k(t)}{dt} = \sum_{j=1}^{C} -\alpha_{(j,k)}(t) I_j(t) S_k(t) \quad (10)$$

Social-Community Compartmental Model With Vital Dynamics. After taking vital dynamics and social community into consideration, we have theorem as follows.

Theorem 1: We assume that there are C different social communities in a city and we also assume vital system dynamic. For the social community k, the number of infected vehicles $I_k(t)$ and susceptible vehicles $S_k(t)$ evolves as follows:

$$\frac{dI_k(t)}{dt} = \sum_{j=1}^{C} \alpha_{(j,k)}(t) i_j(t)[n_k(t) - i_k(t)] \quad (11)$$

$$\frac{dS_k(t)}{dt} = \sum_{j=1}^{C} -\alpha_{(j,k)}(t) i_j(t)[n_k(t) - i_k(t)] \quad (12)$$

and the number of infected operating vehicles $i_k(t)$ and susceptible operating vehicles $s_k(t)$ evolves as follows:

$$\frac{di_k(t)}{dt} = \sum_{j=1}^{C} \alpha_{(j,k)}(t) i_j(t)[n_k(t) - i_k(t)] + \frac{I_k(t) - i_k(t)}{N_k - n_k(t)} B_k(t) - \frac{i_k(t)}{n_k(t)} D_k(t) \quad (13)$$

$$\frac{ds_k(t)}{dt} = -\sum_{j=1}^{C} \alpha_{(j,k)}(t) i_j(t)[n_k(t) - i_k(t)] + [1 - \frac{I_k(t) - i_k(t)}{N_k - n_k(t)}] B_k(t) - [1 - \frac{i_k(t)}{n_k(t)}] D_k(t) \quad (14)$$

C. Simulation Validation on Shanghai Trace

We evaluate our proposed model against conventional ODE model through simulations.

Encounter Rate. Fig. 4(a) illustrates the encounter rate of entire taxi population, indicating that the encounter rate $\alpha(t)$ among all operating vehicles does not vary significantly over 24 hours on that particular day. Fig. 4(b) shows the intra- and inter-community encounter rate of operating vehicles. Intra-community encounter rate $\alpha_i(t)$ is at least 1/3 higher than inter-community encounter rate $\alpha_{i,j}(t)$, suggesting that a vehicle has higher chance to encounter with vehicles from the same community than with those from other communities.

Model Validation. We develop a customized C++ simulator to emulate how information propagates in the delay-tolerant networks, in which Shanghai taxi trace on Jan. 31, 2007 is used as mobility profiles of these emulated vehicles. Since $i(t)$ and $s(t)$ are two aspects of the same problem, to save limited space, we only present the measurement results of $i(t)$ in our simulations. As shown in Fig. 4(c), we plot measurement results $\frac{di(t)}{dt}$ obtained from simulation studies, together with prediction values derived from two different compartmental models – (1) conventional model (Eqn.3-4), and (2) our proposed model taking both social community and vital dynamics into account (Eqn.13-14). It is observed that our proposed model drastically outperforms the conventional model, and it closely matches measurement results obtained from simulations. To take a deep dive, in Fig. 5, we plot $\frac{di_n(t)}{dt}$ in each of the three major communities, in which both measurement results from simulations and prediction values derived from our model were plotted. For all communities, our proposed community-based compartmental model accurately matches the empirical measurements. The observation made above motivates us to believe that the hidden social community structure and system vital dynamics, which have been long ignored, in fact turn out to be critical factors determining the information propagation process in vehicular networks.

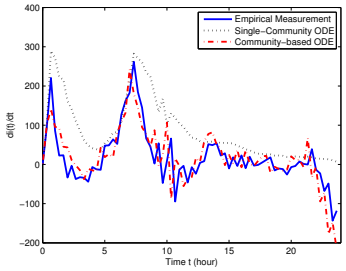


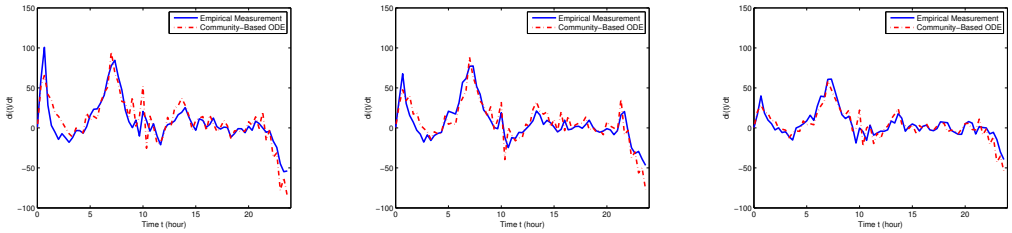
Fig. 7. The Rate of Change of the Total Number of Operating Infected Nodes $i(t)$ in the Beijing Taxi Trace. Comparison of empirical measurement from simulation with predictions from the homogeneous and social-community-based ODE models is presented.

D. Simulation Validation on Beijing Trace

Figure 6 shows that the community ODE model also predicts the number of alive infected nodes in each community for the Beijing trace very well for the first three communities. The one community where this prediction is not so good is the smallest of the four communities, which contains only about 20 vehicles. This is primarily due to the sparsity of contacts in this small community (first-order ODE models generally do not fit small populations well due to this very problem of small samples and high variance). Figure 7 shows that the overall number of operating infected nodes in the whole city is modeled very well by the community approach, decisively showing the efficacy of this approach in modeling information dissemination in large vehicular fleets in real cities.

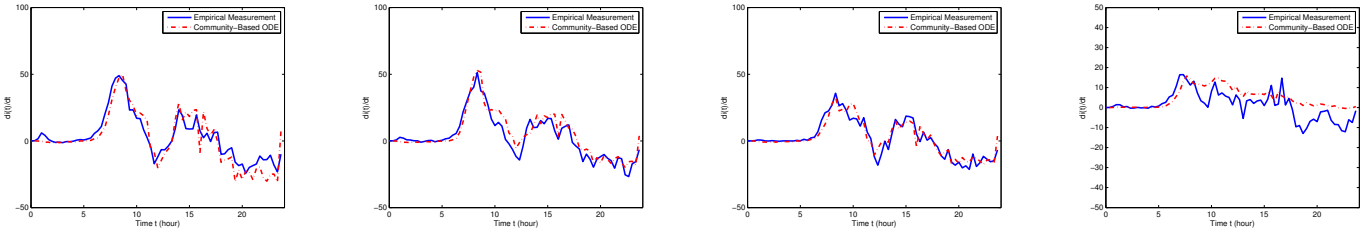
VII. Optimal Seeding Mechanism in Vehicular Networks

In this section, we tackle an optimization problem in vehicular delay-tolerant networks using the novel community-based ODE model developed in the previous section and show that it outperforms the state of the art solution obtained with a simple homogeneous ODE model.



(a) Social Community 1 (b) Social Community 2 (c) Social Community 3

Fig. 5. The Rate of Change of Number of Alive Infected Nodes $i_n(t)$ in Each Community over 24 Hours in Shanghai Taxi Trace (Jan 31, 2007). Both empirical measurement from simulation and prediction based on community-based ODE model are plotted.



(a) Community 1 (b) Community 2 (c) Community 3 (d) Community 4

Fig. 6. The Rate of Change of Alive Infected Nodes $i_k(t)$ in Each Community over 24 Hours in Beijing Taxi Trace. Both empirical measurement from simulation and prediction based on community-based ODE model are plotted.

TABLE IV
OPTIMUM SEED ALLOCATION VIA SIMPLE ODE MODEL AND COMMUNITY-BASED ODE MODEL (SHANGHAI TRACES)

Settings	Simple ODE		Community-based ODE	
	Allocated Seeds	Utility by Deadline	Allocated Seeds	Utility by Deadline
$T = 5, w = 2$	98	386	[40 0 48]	426
$T = 10, w = 0.9$	85	780	[10 2 68]	779
$T = 10, w = 2$	53	704	[0 0 48]	743
$T = 10, w = 5$	30	619	[0 0 26]	645
$T = 15, w = 10$	9	613	[0 0 10]	708

TABLE V
OPTIMUM SEED ALLOCATION VIA SIMPLE ODE MODEL AND COMMUNITY-BASED ODE MODEL (BEIJING TRACES)

Settings	Simple ODE		Community-based ODE	
	Allocated Seeds	Utility by Deadline	Allocated Seeds	Utility by Deadline
$T = 10, w = 0.9$	190	-30	[10 50 80 50]	-19
$T = 10, w = 2$	36	35	[0 0 0 42]	64
$T = 15, w = 0.9$	159	59.9	[0 30 80 40]	62
$T = 15, w = 2$	64	48	[0 0 0 50]	75
$T = 15, w = 5$	2	-2	[0 0 0 12]	62
$T = 20, w = 2$	65	104	[6 0 0 40]	140

A. Problem Formulation

The cellular networks are still feeling the strain of rapidly increasing data traffic because of new mobile platforms and applications. By applying the concept of WiFi offloading to the context of vehicular networks, hybrid protocols that synergistically combine direct cellular access along with store-carry-forward routing through peer-to-peer vehicular communication will provide a bandwidth-efficient and cost-effective way for dissemination in vehicular networks.

One extreme way for the dissemination is to send the contents to each one of vehicles in interest through cellular radio only, which incurs significant access fees although the delay would be small. The other extreme is to send the message to only a small number of seed vehicles in each interested group through cellular radio, and let it spread to other vehicles through V2V communications. The authors of [13] formulate this as an optimization problem from the perspective of a content provider, with the goal of maximizing the number of vehicles that obtain the content within a given deadline while minimizing the expense of using the cellular infrastructure. Mathematically, one can define a utility function as follows:

$$U(\mathbf{k}) = \sum_{m=1}^{c} i_m(T) - w \sum_{n=1}^{c} k_n \qquad (15)$$

in which $\mathbf{k}$ is the vector of seeds; each of its elements, k_n, represents number of seeds allocated to the data chunk in each cluster, n. w is the normalized cost of planting each seed and T is the deadline by which we count the number of infected nodes. Furthermore, $i_m(T)$ represent the number of satisfied nodes by the deadline T in cluster m and is a function of vector $\mathbf{k}$ which can be computed numerically by solving the linear system of ODE. As a result, the optimization problem

can be formulated as follows:

$$\underset{\mathbf{k}}{\text{maximize}} \quad U(\mathbf{k}) = \sum_{m=1}^{c} i_m(T) - w \sum_{n=1}^{c} k_n \quad (16)$$

subject to: (17)

$$0 \leq k_n \leq N_n, \quad n \in \{1 \dots c\} \quad (18)$$

$$\sum_{n=1}^{c} k_n \leq C, \mathbf{k} \in \mathbb{N} \quad (19)$$

where N_n is the total number of nodes in cluster n.

Although one might think that solving the above linear program might be costly for a large network of multiple clusters, we can prove that the computational cost is linear in number of nodes and in number of clusters by referring to the following theorem, which allows the use of a gradient descent algorithm. We omit the proof due to space constraints.

Theorem 2: The number of infected nodes in the community based model of the network is a concave function of the vector of seeds.

B. Simulation Validation

Through extensive simulations, we study if the structure of social communities could be used to improve the performance of optimal seeding process. We consider two different approaches. In the first approach, by assuming a homogeneous contact pattern among all vehicles, the conventional simple ODE model is applied to find an optimal seeding scheme (this is what was done by the authors of [13]). In the second approach, we instead take social community structure into consideration (and only assume mobility homogeneity within each social community), and then apply the social community-based ODE model to derive the number of optimal seeds for each individual community. For our optimization problem formulation, we consider varying two key system parameters: (1) the deadline for interested node to receive content (T); and (2) normalized costs for seeding (w). By using the simple ODE model and community-based ODE model, the optimum number of seeds requirements under a given (w, T) for the two different traces of Shanghai and Beijing are listed in Tables IV and V respectively.

Under the simple homogeneous ODE model, for each setting, we choose the predicted number of seeds randomly from the set of all active nodes. And for the social community-based ODE model in each setting we choose the predicted number of seeds for each community randomly from the active nodes in that community. Then we trace the propagation of files through the network using the real traces and the customized C++ simulator. The ultimate utilities gained by the deadline in each setting under the above mentioned two models are compared in the utility columns of Tables IV and V.

It can be seen from the tables that the community ODE modeling always improves the utility of dissemination, with less amount of seeding cost. When the resources become scarce (i.e., the deadline is tight, or the seeding cost is high), the performance gap between the community ODE model and the simple ODE model is much more noticeable in both Beijing and Shanghai traces. However, if there is enough time or sufficient seeds to flood the network, the advantage of community ODE model is not obvious (such scenarios can be spotted with the normalized seeding cost $w = 0.9$). Meanwhile, we also observe that the network average contact rate is a critical factor which determines what set of resources are tight. For example, in our two different traces, the Beijing average contact rate is almost 2.6 times less than Shanghai average contact rate. As a result, the advantage of our community ODE model over the simple ODE model could be clearly observed in all the settings with $T \leq 15$ or $w \geq 2$; in contrast, in Shanghai trace, due to its high contact rate, our community ODE model shows its vast advantage over the simple ODE model only if $w \geq 10$.

VIII. Conclusion

We have presented a detailed analysis of taxi traces showing that such vehicular fleets are often characterized by a small number of distinct communities. We have further shown that these communities essentially come about due to social proximity or geographic locality of vehicle movements. We have shown that these communities can be used as the basis for developing a parsimonious multi-community ODE model, which is much better at predicting the dissemination process than a standard homogeneous ODE approach. We have further shown that use of this model also improves the utility of seed-based vehicular information dissemination.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1217260

References

[1] "U.S. Department of Transportation Announces Decision to Move Forward with Vehicle-to-Vehicle Communication Technology for Light Vehicles," http://www.nhtsa.gov/

[2] W.-j. Hsu, A. Helmy, "On Nodal Encounter Patterns in Wireless LAN Traces," IEEE Trans. Mob. Comput. 9(11): 1563-1577 (2010).

[3] M.E.J. Newman, "Analysis of weighted networks," Physical Review E 70, 056131 (2004).

[4] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of Human Mobility on the Design of Opportunistic Forwarding Algorithms," Proceeding of IEEE Infocom 2006.

[5] E. Daly, M. Haahr, "Social Network Analysis for Routing in Disconnected Delay-Tolerant MANETS," Proceeding of Mobihoc 2007.

[6] P. Hui, J. Crowcroft, E. Yoneiki, "Bubble rap: Social-based Forwarding in Delay Tolerant Networks," Proceeding of Mobihoc 2008.

[7] P. Costa, C. Mascolo, M. Musolesi, G. P. Picco, "Social-aware Routing for Publish-Subscribe in Delay-Tolerant Mobile Ad Hoc Networks," IEEE JSAC June 2008.

[8] M. M. Musolesi, C. Mascolo, "CAR: Context-Aware Adapative Routing for Delay Tolerant Mobile Networks," IEEE TMC 2009.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, "Fast unfolding of Communities in Large Networks," Journal of Statistical Mechanics: Theory and Experiments 2008.

[10] N. O. Nguyen, T. N. Dinh, S. Tokala, M. T. Thai, "Overlapping Communities in Dynamic Networks: Their Detection and Mobile Applications," Proceeding of Mobicom 2011.

[11] A. Ganesh, L. Massoulie, D. Towsley, "The Effect of Network Topology on the Spread of Epidemics," Proceeding of IEEE Infocom 2006.

[12] X. Zhang, G. Neglia, J. Kurose, D. Towsley. "Performance Modeling of Epidemic Routing," Elsevier Computer Networks journal, 2007

[13] J. Ahn, M. Sathiamoorthy, B. Krishnamachari, F. Bai, L. Zhang, "Optimizing Content Dissemination in Vehicular Networks with Radio Heterogeneity," *IEEE Tran. on Mobile Computing*, 2014.