This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

On the Opportunistic Topology of Taxi Networks in Urban Mobility Environment

Ran Xu, Yong Li, *Member, IEEE,* and Sheng Chen, *Fellow, IEEE*

Abstract—Understanding and characterizing the connectivity of vehicular networks has become increasingly important because of their wide applications and fast development. To address the dynamical links in vehicular networks, time-varying graph (TVG) is one of the most important models. Nowadays, due to the fact that lots of vehicular applications can tolerate a small amount of latency in communication, opportunistic reachability graph (ORG) characterizes the connectivity better by introducing delay tolerance to the model. However, people still do not have a high-level summarization, i.e. the topology, of the vehicular network on how nodes are clustered and isolated.

In this paper, based on ORG model, we analyze the opportunistic topology of taxi networks in urban mobility environment by mainly focusing on the number, location and evolution of connected components and the size of the largest components to reveal the unique properties of the taxi networks instead of just links and hops. Our analysis is based on the real taxi traces of big cities and reflects the real urban mobility environment. We find that the opportunistic topology of the networks with delay tolerance is substantially different from the instantaneous topology without considering the delay. Moreover, we unveil the fundamental relationships and trade-offs between the dynamical topology and the key network parameters related to mobility, e.g., delay tolerance, transmission distance, etc. To the best of our knowledge, our study is the first work to reveal the characteristics of opportunistic topology models in the large-scale urban mobility environment with real traces.

Index Terms—Vehicular networks, taxi networks, mobility, dynamical connectivity, opportunistic topology, delay tolerance

1 Introduction

Nowadays, due to the rapid development in communication technology and traffic systems, vehicular networks, which provide a basic model for many vehicular applications, have received considerable attention. The newly emerged vehicular communication networks are seen as a key technology for improving road safety and building intelligent transportation system (ITS) [1]. Many applications of vehicular networks are also emerging, including automatic collision warning, remote vehicle diagnostics, emergency management and assistance for safe driving, vehicle tracking, automobile high speed Internet access, and multimedia content sharing [2]. Since most of these applications are inherently designed for the information exchange of vehicle-based data, related to the positions, speeds, and locations among vehicles in a restricted region, the dynamical connectivity of the vehicular networks is well worth of researching.

Time-varying graphs and temporal reachability graphs [3] are the two basic graph models to analyze the dynamical connectivity of the vehicular networks considering the multi-hop connections and the delay-tolerance property. To be more specific, an edge (i, j) in temporal reachability graph means that a message can be sent from vehicle node i at the moment t and delivered to node j by the moment $t + \delta$ via multi-hop connections, where δ is the maximum delay that can be tolerated.

Beyond the opportunistic connectivity that considers connected pairs and average density [3], the *connected component* in the graph models is a basic unit to study the opportunistic topology, which focuses on the graph model of the vehicular networks and analyzes how nodes are clustered and isolated dynamically. How the snapshot of the connected components looks like [3] and how it evolves [16] are all interesting topics to understand the opportunistic topology of the vehicular networks.

In vehicular networks, transmission distance and delay tolerance are two important factors that affect the network topology. The transmission distance is defined as the maximum distance of two vehicles with reliable communication of certain vehicle-to-vehicle communication technology. Generally speaking, the longer the transmission distance is supported, the better the connectivity of the whole network, which also offers better data transmission between any two nodes.

On the other hand, delay tolerance is defined as the maximum network transmission latency that an application can tolerate. It also affects the network topology considerably, since the connection between two nodes no longer depends solely on the instantaneous topology but the dynamical topology evolved within a period of time. Some existing literatures for opportunistic networks or delay tolerant network (DTN) [4], [5], [6] have set the delay tolerance to 10 minutes to 1 hour or even longer. The study [4] quested for killer applications in DTN and summarized several useful applications in specific scenarios, e.g., short message service,

R. Xu and Y. Li are with Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (E-mails: xu943@purdue.edu, liyong07@tsinghua.edu.cn). R. Xu is also affiliated with Purdue University.
S. Chen is with Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK (E-mail: sqc@ecs.soton.ac.uk), and also with King Abdulaziz University, Jeddah 21589, Saudi Arabia.
This work was partially supported by the National Natural Science Foundation of China (NNSFC) under grants No. 61201189 and No. 61132002, the Creative Research Groups of NNSFC under grant No. 61021001, the National S&T Major Project of China under grant No. 2011ZX03004-001-01, National High Tech (863) Projects of China under Grant No. 2011AA010202, and Research Fund of Tsinghua University under grants No. 2011Z05117 and No. 20121087985.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

file sharing and bulk data transfer, etc. The work [5] studied a message ferrying scheme in a DTN, where the message deliver delay can be 1 hour to several hours. The study [6] presented a prototype of sharing location-based information by a peer-to-peer (P2P) synchronization mechanism. Their results show that it usually takes 10 minutes to 1 hour for information to spread to 90% of the nodes. Thus, we set the range of delay tolerance from 6 minutes to 1 hour in our study, and typically using 10, 20, 40 and 60 minutes in most of the evaluations. Our targeted applications are not safety-related or latency-sensitive ones that requires millisecond or second level latencies.

In this paper, we discuss the dynamical topology of taxi networks affected by the two key network parameters – transmission distance and delay tolerance. Our contributions are briefly summarized as follows.

- We propose to use the number, location and evolution of connected components to characterize the opportunistic topology of the taxi networks. Considering delay tolerance, the opportunistic topology provides a high-level summarization on the dynamical connectivity of the taxi networks. Thus, our model is beneficial to the future inter-vehicle applications.
- We provide insights into the dynamical topology affected by the key network parameters, i.e. transmission distance and delay tolerance.
- We perform the in-depth analysis on the real taxi traces, which reflects the real opportunistic topology of large-scale urban taxi mobility environment.
- We propose a detailed algorithm to compute ORG with a large amount of nodes in the graph.

The rest of the paper is structured as follows. Section 2 introduces the concepts of time-varying graphs and opportunistic reachability graphs with the related key metrics, while Section 3 presents the method of processing and the algorithm in details together with the associated definitions. In Section 4, we use the key metrics to analyze the opportunistic topology and dynamical connectivity of taxi networks. In Section 5, we introduce the related works to highlight our differences from these existing works, and our conclusions are drawn in Section 6.

2 Graph Models, Key Metrics and Channel Link Models

2.1 Graph Models

Since the graph is a natural model to represent static networks, the time-varying graph (TVG) offers a natural approach to represent the highly dynamical vehicular networks. In particular, TVGs offer a useful high-level abstraction for investigating the instantaneous connectivity and reachability of vehicular networks. While the ORGs reveal the connectivity and reachability of delay-tolerant vehicular networks with multi-hop connections. In this work, we utilize definitions and notations proposed by Casteigts et al. [7], who present the TVG and ORG formalism with dedicated notations and integrate the existing models, concepts, and results into a unified framework. We then introduce our key metrics of taxi networks based on ORG. These metrics are able to reveal the crucial properties of dynamical connectivity in the taxi networks.

Definition 1 (Time-varying Graphs, TVGs). Let V be a set of vertices (vehicle nodes) where $|V| = N$ is the number of vertices. Let $E^G \subseteq V \times V$ be the set of edges among the vertices V. Assume that the dynamical events take place over a time span $\mathcal{T} \subseteq \mathbb{R}^+$ in a positive real-valued temporal domain. A general TVG is defined by a tuple $\mathcal{G} = (V, E^G, \mathcal{T}, G)$, where

- $G : E^G \times \mathcal{T} \rightarrow \{0, 1\}$, called *presence* function, indicating whether a given edge $e \in E^G$ exists at a given time $t \in \mathcal{T}$, or whether the two vehicle nodes are connected through direct V2V link at t;

Definition 2 (Discrete Time-varying Graphs, DTVGs). To describe the TVG with a limited amount of data, we divide the time into slices and each slice has a length of η. Within each time slice, the TVG is assumed to be constant. That is, $\forall k \in \mathbb{N}$, $t \in \mathbb{R}^+$, $k\eta \leq t < (k+1)\eta \Rightarrow \mathcal{G}(t) = \mathcal{G}(k\eta)$. In this work, we investigate the discrete TVG, i.e., $\mathcal{T} = \mathbb{N}\eta$.

Definition 3 (Journey). Given an edge $e = (u, v)$, we define $\texttt{from}(e) = u$ and $\texttt{to}(e) = v$. A journey in $\mathcal{G}$ is a sequence of couples $\mathcal{J} = \{(e_1, t_1), (e_2, t_2), \cdots, (e_k, t_k)\}$ such that $\{e_1, e_2, \cdots, e_k\}$ is a walk in $\mathcal{G}$ satisfying

$$\begin{cases} \texttt{from}(e_i) = \texttt{to}(e_{i-1}), 2 \leq i \leq k \\ t_i \in \mathbb{N}\eta,\ 1 \leq i \leq k \\ \exists t, s.t. t_i \leq t < t_{i+1}, G(t)_{e_i} = 1 \end{cases} \quad (1)$$

We denote $\mathcal{J}_{(u,v)}$ as the journey from u to v, i.e. $from(e_1) = u$ and $to(e_k) = v$. Obviously, the existence of a journey is not symmetrical, which means that u can reach v does not imply v can reach u. Further, let $\texttt{departure}(\mathcal{J})$ be t_1, $\texttt{arrival}(\mathcal{J})$ be the time when e_k is present, and $\mathcal{J}^{tl} = \texttt{arrival}(\mathcal{J}) - \texttt{departure}(\mathcal{J})$, named *temporal length*, define the end-to-end transmission latency.

It can be seen that a journey represents a sequence of hops that data follow through the direct links in DTVGs from a source node to a destination one. Thus, it can be used to derive the reachability and topology properties of delay-tolerant vehicular networks with multi-hop connections. Also, whether the journey exists depends largely on the maximum delay tolerance δ of the message transmission.

Taking the DTVGs with five nodes shown in Fig. 1(a) as an example of vehicular networks, where the edges mean the direct bi-directional V2V links between nodes and time slice is set $\eta = 10$ s. We observe that the links in the networks are varying with time – node 1 and node 2 are connected from 10 s to 40 s, and node 1 and node 5 are connected from 20 s to 30 s, etc. Obviously, a DTVG has different topologies at different timestamps $t \in \mathbb{N}\eta$, and an instantaneous graph only indicates the network topology at a particular discrete timestamp t.

In Fig. 1(b), we consider the delay tolerance in the vehicular networks, For example, assume that at time $t = 10$ s, node 1 has a message to node 3. If the delay tolerance of this message is $\delta = 20$ s, this message cannot be delivered since there exists no journey $\mathcal{J}_{(1,3)} = \{(e_1, t_1), (e_2, t_2), \cdots, (e_k, t_k)\}$ with $from(e_1) =$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

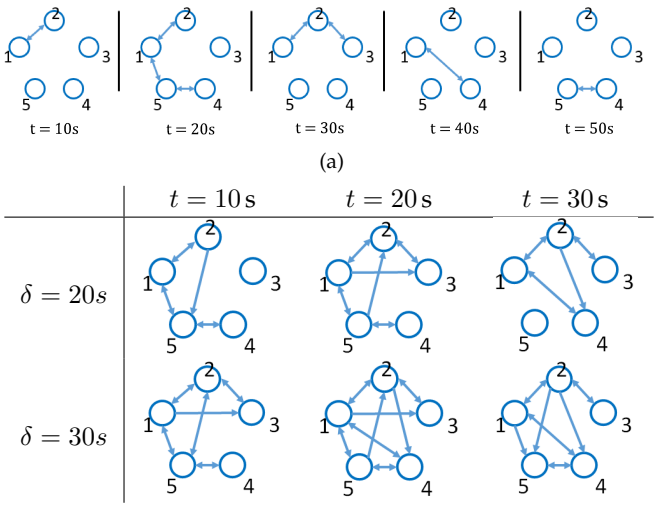


Fig. 1. An example of graph models with five nodes: (a) instantaneous time-varying graphs (TVGs), and (b) opportunistic reachability graphs considering delay tolerance (ORGs).

node 1, $to(e_k)$ = node 3, $\texttt{departure}(\mathcal{J}) = 10\,\mathrm{s}$ and $\texttt{arrival}(\mathcal{J}) \leq 30\,\mathrm{s}$. However, if the delay tolerance is $\delta = 30\,\mathrm{s}$, then there exists a journey $\mathcal{J}_{(1,3)}$ with $\texttt{departure}(\mathcal{J}) = 10\,\mathrm{s}$ and $\texttt{arrival}(\mathcal{J}) = 40\,\mathrm{s}$, which indicates the message can be delivered via this journey.

It can be seen that the concept of journey is important to the dynamic network topology. Based on the above discussions for DTVG and journey, we introduce the definition of ORG.

Definition 4 (Opportunistic Reachability Graphs, ORGs). Let V be the set of vertices and $E_\delta^R \subseteq V \times V$ be the set of edges between the vertices in V given the maximum delay tolerance δ. The dynamical events take place in the temporal domain $\mathcal{T} = \mathbb{N}\eta$ and then we consider discrete network delay tolerance δ, taking values from the set $\{\eta, 2\eta, ...\}$. Given different delay tolerances, the graphs are different. In particular, ORG is same as TVG if the delay tolerance is set to $\delta = \eta$, assuming that the time to transmit the data is within one time slice η. An ORG is then defined by a tuple $\mathcal{R}_\delta = (V, E_\delta^R, \mathcal{T}, R_\delta)$, where

- R_δ: $E_\delta^R \times \mathcal{T} \rightarrow \{0, 1\}$, called *presence* function, indicates whether a given edge $e = (u, v) \in E_\delta^R$ or a journey $\mathcal{J}_{(u,v)}$ exists which satisfies $\mathcal{J}^{tl} \leq \delta$, at a given starting time $t \in \mathcal{T}$.

For example, let us denote the presence function of edge (i, j) at t by $R_\delta(t)_{(i,j)}$. In Fig. 1(b), where $\eta = 10\,\mathrm{s}$, $R_{2\eta}(\eta)_{(1,3)} = 0$ since there exists no journey $\mathcal{J}_{(1,3)}$ with $\texttt{departure}(\mathcal{J}) = 10\,\mathrm{s}$ and $\texttt{arrival}(\mathcal{J}) \leq 30\,\mathrm{s}$, under the delay tolerance of $\delta = 2\eta$. However, if the delay tolerance is $\delta = 3\eta$, $R_{3\eta}(\eta)_{(1,3)} = 1$ because there exists a journey $\mathcal{J}_{(1,3)}$ with $\texttt{departure}(\mathcal{J}) = 10\,\mathrm{s}$ and $\texttt{arrival}(\mathcal{J}) = 40\,\mathrm{s}$.

It can be seen that delay tolerance alters the topology, and a consequence is that the connection between two nodes may be asymmetric. For example, under the delay tolerance of $\delta = 20\,\mathrm{s}$, node 2 can reach node 5 at time $t = 10\,\mathrm{s}$ but the reverse connection is disabled, as indicated by the arrow in a single direction from node 2 to node 5. This implies that at $t = 10\,\mathrm{s}$, node 2 may transmit messages to node 5, but node 5 cannot transmit messages to node 2. This is very different from the instantaneous TVG without considering delay tolerance, where every edge is bi-directional. If the delay tolerance is loosened, for example, to $\delta = 30\,\mathrm{s}$, then node 2 and node 5 become connected by the bi-directional arrows at time $t = 10\,\mathrm{s}$, as shown in Fig. 1(b), implying that node 2 may transmit messages to node 5 and/or node 5 may transmit messages to node 2. Therefore, in an ORG, there exist some single-hop/multi-hop paths or connections which are directed only, while other connections are bi-directional.

2.2 Key Metrics and notations

Based on ORG, we now introduce the key metric we use in this paper to characterize the dynamical topology of vehicular networks.

Definition 5 (Component). Considering the opportunistic topology graph (ORG) at a discrete time-point t, we have a graph $\mathcal{R}_\delta(t)$ consisting of its node-set V and edge-set E_δ^R at t. To study the partitioned sets of the ORG, we define component to be a set of nodes where each node can link to at least one other node in the component through a multi-hop journey. Formally, the i-th component $C_{\delta,i}(t)$ contains a subset of the node-set V, denoted by $V_{\delta,i}(t) = \{v_j | \exists v_k \in V_{\delta,i}(t), R_\delta(t)_{(j,k)} = 1 || R_\delta(t)_{(k,j)} = 1\}$, and a subset of the edge-set E_δ^R, denoted by $E_{\delta,i}^R(t) = \{e_{j,k} | v_j, v_k \in V_{\delta,i}(t) \&\& e_{j,k} \in E_\delta^R\}$. The size of a component is the number of nodes belonging to it, i.e., $S_{\delta,i}(t) = |V_{\delta,i}(t)|$.
We can represent a vehicular network by a set of components $\{C_{\delta,i}(t)\}$. A vehicle may reach other vehicles in the same component via multi-hop communication at time t.

Definition 6 (Degree). We consider the adjacent node set of vertex v_i in ORG $V_{\delta,i}^{'}(t) = \{v_j | R_\delta(t)_{(i,j)} = 1 || R_\delta(t)_{(j,i)} = 1\}$, which represents all the nodes that can communicate with vertex v_i and vice versa via multi-hop transmissions from t to $t + \delta$. The degree of vertex v_i is simply defined as $D_{\delta,i}(t) = \left|V_{\delta,i}^{'}(t)\right|$.

To study the dynamical topology of vehicular networks, we mainly consider the number, geographical distribution and evolution of connected components, the size of the largest component and node degree. Component number is used to characterize how heavily the network is partitioned. The higher the number, the heavier the network is partitioned. This metric is significant in providing a whole picture of the network topology, while common metrics used in other works, such as the probability of a link, may fail to do. Geographical distribution and evolution of connected components is used to clearly depict where the components are located in the urban area and how the components merge and break over time. Largest component is where the majority of communications take place and its size shows the maximum number of vehicles in a connected component. Last but not least, the node degree characterizes the connectivity from an individual node to show how many vehicles one node can reach at most. These metrics cover the network topology and connectivity in different aspects and are of great importance to the research community and vehicular industry.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

TABLE 1
Notations of key metrics and network parameters.

Abbreviation	Notation	Description
Component number	C	The count of connected components
Component size	S	The size of a connected component
Largest component size	S_{max}	The size of the largest connected component
Degree	D	Count of connected nodes for a given node
Distance	d	Within d, nodes are considered connected in TVG
Delay	δ	The maximum tolerated latency for the networks
Network size	N	Count of vehicle nodes considered, $N = \|V\|$

In the following sections, we simply denote the number of components by C, the size of each component by S_i, where $1 \leq i \leq C$, and the size of the largest component S_{max}, as well as the degree of each node by D_i, where $1 \leq i \leq N$. It is notable that these key metrics are still affected by delay tolerance δ, although the notation is dropped for simplicity. It is also affected by the maximum transmission distance as introduced in Sec. 2.3 and Sec. 3.3. In Table 1, we summarize the notations of the key metrics and network parameters.

2.3 Channel Link Models

Channel link models determine whether two vehicles are directly connected without rely nodes. We consider all the vehicle pairs within transmission distance d defined in Sec. 1 as bi-directional connected and not connected otherwise. This model is simplified from a log-normal path loss model, which is discussed in Sec. 2.3.1. Although obstacles from other vehicles (Sec. 2.3.2), buildings (Sec. 2.3.3) and real packet loss ratio (Sec. 2.3.4) are all the real factors to the model, we actually consider it in a high level that the two vehicles are by all means connected with a certain V2V communication technology that supports a maximum transmission distance d. It is true that the real connectivity may be better, since the vehicles can still stay connected beyond the distance d. We admit the limitation on the conservative analysis where the real connectivity can be better than our results.

2.3.1 Path loss model

To determine whether an edge in the TVG exists, we first study the path loss model for the received signal strength (RSS) at distance d away from the source node. A widely used log-normal model is given as follows [8],

$$\mathrm{RSS}(d) \;=\; \mathrm{RSS}(d_0) + \alpha \log\left(\frac{d}{d_0}\right) + X \tag{2}$$

where $\mathrm{RSS}(d)$ is the RSS at distance d, $\mathrm{RSS}(d_0)$ is the measured RSS at a reference distance d_0 and α is the path loss exponent, and X is a random variable which counts for small-scale or fast fading effects and it is well-known to follow a Rice distribution if there exists the line of sight (LOS) between the communicating vehicles or a Rayleigh distribution if there exists no LOS. In mobile networks design, the effects of X is taken into consideration by assigning a so-called fast-fading margin F_M which ensures that the probability of X exceeding F_M is less than a threshold, e.g. 1%. Measurement studies and investigations for appropriate values of path loss exponent α and fast-fading margin F_M under various mobile communication environments can readily be found in the literature. For example, a measurement study is given in [9] for dedicated short range communication (DSRC) in vehicular networks.

For the receiver to correctly detect the transmitted signal, the RSS should be above a minimum threshold, denoted $\mathrm{RSS}_{\min}$. Thus, the maximum transmission distance d between the transmitter and receiver can be derived by,

$$d \;=\; d_0 \cdot \exp\left(\frac{\mathrm{RSS}_{\min} - \mathrm{RSS}(d_0) - F_M}{\alpha}\right) \tag{3}$$

More specifically, when the distance between two vehicles is smaller than d, a communication link between them can be established. Obviously, the maximum transmission distance d depends on the propagation environment (reflected in α and F_M), e.g., urban or rural, as well as the transmit power (reflected in $\mathrm{RSS}(d_0)$).

In this paper, we simplify the maximum transmission distance in the real scenario by considering a single maximum transmission distance d for all cases, which corresponds to a certain V2V communication technology that by all means supports the communication within distance d.

From the above discussion on the link model, it is clear that the edges in our TVG are undirected, since an edge represents a bi-directional communication between the two vehicles.

2.3.2 Obstacles of other vehicles

In the studies [10], [11], an obstacle-based channel model is used to characterize the effects from an obstructing vehicle on the LOS. To be specific, consider that vehicle i is communicating with vehicle j, where vehicle k obstructs the LOS between i and j. Let the distance between vehicles i and j be $d_{i,j}$, and the distance between i and k be d_{obs}. The heights of vehicles i and j are h_i and h_j, respectively, while the height of their antennas is h_a. Further denote μ and σ as the mean and standard deviation of the height of the obstructing vehicle. The radius for the first Fresnel zone ellipsoid r_f is given by

$$r_f = \sqrt{\frac{\lambda d_{obs}(d_{i,j} - d_{obs})}{d_{i,j}}} \tag{4}$$

where λ is the wavelength of the signal. The effective height of the LOS line is given by

$$h = (h_j - h_i)\,\frac{d_{obs}}{d_{i,j}} + h_i - 0.6 r_f + h_a \tag{5}$$

The probability of the link between node i and j is given by

$$\Pr(\mathrm{LOS}|h_i, h_j) = 1 - Q\left(\frac{h - \mu}{\sigma}\right) \tag{6}$$

where $Q(\cdot)$ is the standard Gaussian error function.

2.3.3 Obstacles of buildings in urban area

The studies [12], [13] use an empirical model for radio shadowing when considering the obstacles of buildings, by taking into account the walls of building outlines and the length of the path through the buildings. Specifically, the additional attenuation of a transmission due to the obstructing buildings is given by

$$L_{obs}\,[dB] = \beta n + \gamma d_m \tag{7}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

where β is given in dB per wall which represents the attenuation due to any additional exterior walls of the buildings standing on the LOS line, and n is the number of walls that standing on the LOS, while γ is given in dB per meter to represent the attenuation of the internal structure of the buildings, and d_m is the total length of the transmission inside the buildings.

Parameters $\beta \approx 9\,\text{dB}$ and $\gamma \approx 0.4\,\text{dB}/m$ in their model fit well into their experimental results.

2.3.4 Packet loss ratio

Study [13] considers the packet loss ratio in V2V communications. Specifically, the packet loss probability $\text{loss}(P_r)$ (in dbm) is given by

$$\text{loss}(P_r) = \begin{cases} 0, & \text{if } P_r > P_r^{\max} \\ \min\left\{1, \left(\dfrac{P_r^{\max} - P_r}{P_r^{\max} - P_r^{\min}}\right)^{\theta}\right\}, & \text{if } P_r \le P_r^{\max} \end{cases} \tag{8}$$

with $P_r^{\max} = -78\,\text{dbm}$, $P_r^{\min} = -91\,\text{dbm}$ and $\theta = 3.6$.

The probability that a communication link can be established is then given by $1 - \text{loss}(P_r)$.

2.3.5 Summary of our model in terms of these factors

It is unrealistic to explicitly count for all the last three factors in our link model for a large-scale urban vehicular environment for the following reason. The obstacle-based model needs extremely fine-grained data, including the height of every vehicle and the details of every building, to calculate the additional attenuation. This information is unavailable in any public dataset and would be extremely costly and hard to acquire in practice. Just image the situation that a vehicular network protocol adopts this model to establish link. In a particular location, the vehicle would need to know the details of the surrounding area, which is completely impossible. Even assuming one can obtain all the parameters required, the model is only valid for this very small specific area of V2V communication, that is, the model would be different from one small area to another. Thus, it is impractical to establish such a model for large-scale vehicular networks.

By contrast, the log-normal model in equation (2) is sufficiently general and can be applied anywhere in a large-scale vehicular network. This model is reasonably accurate, as it represents the 'typical' or 'average' channel propagation environment encountered in the large-scale vehicular network. Note that the average obstacle effect of vehicles and buildings is implicitly been taken into account by the fast fading component X, and the fast-fading margin F_M is chosen according to the distribution of X. Furthermore, the 1% F_M indicates that 99% of the RSS will be higher than the required threshold for correctly detecting data, which corresponds to an equivalent small packet loss probability. It can be seen that the packet loss ratio is also implicitly counted for in the log-normal model.

In order to concentrate on studying the dynamical topology in a large-scale urban environment, we finally choose the log-normal model with the maximum transmission distances d as a threshold to determine whether an edge exists to balance the accuracy and the practicability.

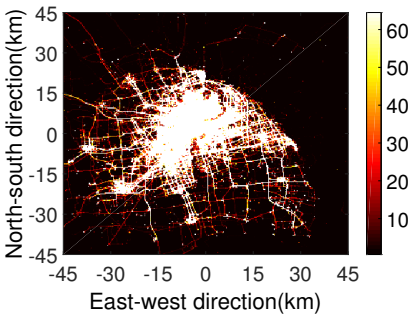


Fig. 2. The geographical distribution of GPS-reported taxis superimposed on Shanghai map. A colored dot presents the average number of appearance per taxi per day in the corresponding 300 m×300 m square area on the map.

3 Datasets, Preprocessing, and Algorithm

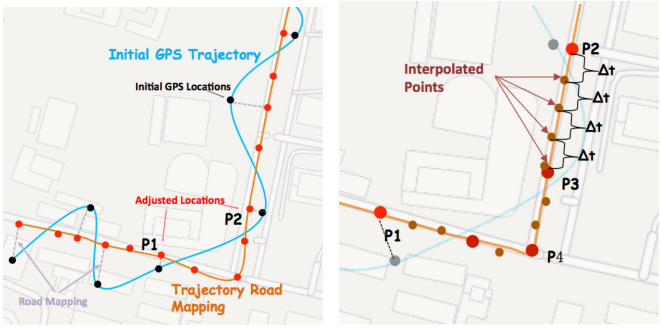
We first provide a brief description of the taxi mobility datasets used in our study and introduce the preprocessing carried out on them. Then, we discuss how to obtain the ORG and related metrics from the datasets. Due to the fact that we have taxi-only traces, we restrict our contribution and evaluation within taxi networks, although all the graph models and channel link models apply to general vehicle datasets.

3.1 Datasets

To provide realistic vehicular mobility and connectivity in urban scenarios, ideally large-scale vehicular datasets are needed which involve all types of vehicles, including taxis, buses and private cars. However, such vehicular trace records do not exist and are unlikely to be available soon. By contrast, real-world taxi datasets are available. In this paper, we employ large-scale taxi mobility trace data to study opportunistic topology of taxi networks in urban environment. Our TVG and ORG models are also useful to characterize vehicular networks if vehicular datasets including all types of vehicles are available.

We employ three large-scale taxi mobility-trace datasets of Shanghai, Beijing and Nanjing. Shanghai trace [14] was collected by SG project [15], in which the mobility trace data from over 4,000 taxis were collected during the whole month of February 2007 in Shanghai. Beijing trace was collected during the whole month of May 2012, including more than 28,000 taxis. Nanjing trace on the other hand was collected over a longer time period from July 2013 to December 2013, involving more than 7,000 taxis. In all these three traces, reports were continuously sent back to the data center by GPRS. Specifically, the frequency of reports was every 15 seconds in most of the time. The information of reports included the taxi's ID, the longitude and latitude coordinates of the taxi's location, the instant speed and other factors like heading angle as well as the status of the taxi. The original Beijing trace covers a very large area, including many suburban areas of Beijing. We limit the use of Beijing trace to the part covering only the downtown region of Beijing, so that the geographic scale of the Beijing trace used is not vastly different from those of Shanghai and Nanjing traces. A distribution map of Shanghai trace is depicted in Fig. 2, which includes the trajectories of 1,500 taxis during

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data



(a) (b)

Fig. 3. Illustration of taxi GPS data preprocessing: (a) location adjustment, and (b) time frequency adjustment.

one day and is superimposed on the downtown region of Shanghai that covers an area of over 2000 square kilometers.

Among thousands of taxis, we randomly sample 1,500 taxis from the trace of each city for further processing. Obviously, 1,500 taxis only represent a small portion of all taxis in the city. However, since taxis operate long service time and have much longer trajectories, they cover the geographical topology of the city well, as shown in Fig. 2. We also show in Section 4.3 that a larger sample size does not necessarily contribute to a more precise model. On the other hand, as discussed in Section 3.4, the time complexity of the algorithm is $O(N^3)$, where N is the number of taxis. Thus considering the tradeoff between the precision and the timing budget, we believe that a subset of 1,500 taxis in Shanghai is a good choice for us to investigate the network topology and connectivity. The same number also applies to Beijing and Nanjing trace, because the geographic scale of Nanjing and downtown region of Beijing are smaller or similar.

3.2 Data Preprocessing

We will use Shanghai trace as an example. Data processing for the other two traces is similar. We need to obtain the taxis' locations varying with time from the taxis' moving traces. The taxis' coordinates reported by GPS devices are longitudes and latitudes with a precision of 0.00001 degrees. For convenience, we convert the coordinates to meters with a precision of 1 m and set the origin point $(0, 0)$ at $(31.2°\text{N}, 121.5°\text{E})$ near the center of Shanghai. We also choose the center of the city as the origin point in the other traces. Since the location data are measured by GPS devices, the noise may exist in the collected data owing to the inaccuracy of GPS devices. Also since the taxis may not report their locations at the same time slots with the same fixed frequency, we need to process the data trace to obtain the accurate locations of all the taxis in the same time slots and at the same frequency. Thus we first use the city map to correct the taxis' locations so that they are all in the city's roads, and this location adjustment is illustrated in Fig. 3 (a), where the measured locations in sequence by the GPS devices shown as black dots. Here, the map matching, or named road matching problem is to find the locations of which roads that the taxi is on. Due to measurement noise, it is prone to error if we match each point with the nearest road. Here, we utilized the widely used Hidden Markov Model (HMM) [37] to find the most likely road route that are represented by a time-stamped sequence of latitude/longitude pairs recorded in our dataset. Note that there is no ambiguity in mapping a point to the correct road, because there maybe many reference points of the same taxi trajectory that are on a road, which is suitable utilized in the HMM model [37]. As tested by authors in [37], HMM elegantly accounts for measurement noise and the layout of the road networks [37]. With the road matching process, we obtain the points/locations of the taxi in the road, and in each road, we also obtain a more density points, even their intervals are still large. We then use the method of interpolation to insert the location points at the time slots we need so that all the taxis have location information at every ten-second interval. Referring to Fig. 3 (b), we now explain how to carry out this interpolation.

Assume that we have the location information samples (x_1, y_1) and (x_2, y_2) of a taxi at time points t_1 and t_2, respectively, where $t_2 > t_1$. If $t_2 - t_1 \leq 10\,\text{s}$, we do not need to carry any interpolation. If $t_2 - t_1 > 10\,\text{s}$, then in order to get the location of the taxi at any time $t \in (t_1,\ t_2)$, we estimate the location (x_t, y_t) by the following interpolation

$$l_t = l_1 + \frac{t - t_1}{t_2 - t_1}(l_2 - l_1),\ \text{with } l = x \text{ or } y \tag{9}$$

After obtaining (x_t, y_t), we do not need to adjust it to be in a city's road since all the neighboring points obtain my the HMM road matching algorithms are straight line in the road. Since the obtained road matching locations are in discrete time points at the time interval of 15 s in most of the time, we choose to calculate the position of the taxis every 10 s using this interpolation.

We now analyze the precision limit of the location estimation (9). In Fig. 4, the instantaneous velocity values of taxis in Shanghai trace are depicted over the 24 hours of a day. The results of Fig. 4 show that the taxis move slower than 7 m/s (15 mph) over 75% of the cases during the day. We may infer an upper bound of maximum error in 75% of the cases as follows. Assume that during the interval $[t_1, t]$, where $t - t_1 \leq 10s$, the taxi travels at the maximum constant velocity of 7 m/s, and hence it will travel the maximum distance of $l_t - l_1 = 70\,\text{m}$. Therefore, the maximum error in (9) cannot exceed this value in 75% of the cases. We do have the limitation that we cannot locate the taxi precisely between the two timestamps and thus we do not know precisely the distance between two taxis between the two timestamps. However, the real maximum error is much smaller than this upper bound, since although we assume that taxis travel at a constant velocity of 7 m/s, in reality they are often slowed down by traffic or stopped at traffic lights. Moreover, the position error due to the interpolation (9) will not accumulate, as each interpolated position is based on two true measurement samples.

After the data preprocessing, we obtain the instantaneous two-dimensional distribution maps of the taxis' positions for every 10 s over the duration of a month, which then become our data for the taxi networks in Shanghai trace. In order to realize the time-consuming algorithm within a reasonable time, we re-sample the data with a period of 20 s. To study how a sampling period of 20 s affects V2V links less than 20 s, we gather the statistics from our Shanghai trace and depict the results in Fig. 5. It can be found that more than 75% of the V2V links last more than 20 s given the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

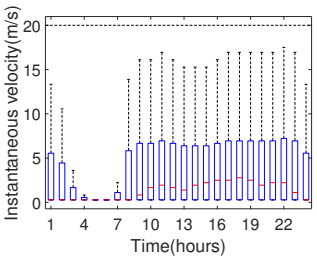


Fig. 4. Instantaneous velocity of taxis in Shanghai trace over the 24 hours of a day.

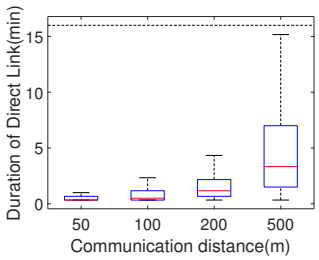


Fig. 5. Duration of direct vehicle-to-vehicle links over one day.

maximum transmission distance of 200 m, while more than 75% of the V2V links last more than 2s minutes given the maximum transmission distance of 500 m. Clearly, by using a sampling period of 20 s, some short (10 seconds) links are neglected and thus our model cannot exactly characterize those one-time-unit (10 seconds) links. Future work may better characterize the network connectivity by considering trajectories with a higher temporal resolution with those short-lived link.

3.3 Calculating ORG

We now discuss how we derive ORGs from TVGs. Based on an urban taxi mobility trace, we have a large number of instantaneous topologies by sampling the data with a fixed frequency $1/\eta$. We model each instantaneous topology at each timestamp t by the TVG, i.e. $\mathcal{G}(t)$. We define the edges $E^G(t)$ as weighted edges, where the weights represent the distances between pairs of nodes. The presence function $G(t)$ of TVG represents the successfully established communication links in the graph and is given by

$$\forall (i,j) \in E^G(t):\ G(t)_{(i,j)} = \begin{cases} 1, & \text{if } E^G(t)_{(i,j)} \leq d \\ 0, & \text{if } E^G(t)_{(i,j)} > d \end{cases} \tag{10}$$

where d is the maximum transmission distance introduced in Sec. 2.3. Next, we can compute the opportunistic topology of the network at time t, i.e. the ORG $\mathcal{R}_\delta(t)$, step by step. We also extend the definition of edges $E^R_\delta(t)$ to weighted edges, where the weights represent the so-called opportunistic distances between the two vertices.

Definition 7 (Summed-Duration Direct Link Graphs). For delay tolerance $\delta \in \mathbb{N}^+\eta$, we define a graph $L_\delta(t)$, called the summed-duration direct link graph at time t, where the weight of the edge $e = (i,j) \in L_\delta(t)$ represents the total duration of direct links between nodes i and j, from time t to $t+\delta$.

For $\delta = \eta$, the presence function $L_\eta(t)$ is the same as the presence function of the TVG, i.e.

$$L_\eta(t) = G(t) \tag{11}$$

For $\delta = n\eta$, we can add the weight of edges of the summed-duration direct link graph between $t+(n-1)\eta$ and $t+n\eta$, i.e., $G(t+(n-1)\eta)$, to those of $L_{(n-1)\eta}(t)$ to obtain the summed-duration direct link graph $L_{n\eta}(t)$, which is symbolically denoted by

$$L_{n\eta}(t) = L_{(n-1)\eta}(t) + G(t+(n-1)\eta) \tag{12}$$

Definition 8 (Summed-Distance Direct Link Graphs). For delay tolerance $\delta \in \mathbb{N}^+\eta$, we define a graph $I_\delta(t)$, called the summed-distance direct link graph at time t, where the weight of the edge $e = (i,j) \in I_\delta(t)$ represents the summed distance of the direct links, i.e., one-hop journeys, between nodes i and j, from time t to $t+\delta$.

For $\delta = \eta$, the summed distance of direct link or the weight of edge $e = (i,j) \in I_\eta(t)$ is equal to that of edge $e = (i,j) \in E^G(t)$ with the distance less than d. Therefore, symbolically we can obtain $I_\eta(t)$ as

$$I_\eta(t) = E^G(t) \times G(t) \tag{13}$$

where the operator $\times$ is the element-wise multiplication.

For $\delta = n\eta$, we can obtain $I_{n\eta}(t)$ by adding the weights of edges in the TVG between $t+(n-1)\eta$ and $t+n\eta$ which are less than d, i.e., $E^G(t+(n-1)\eta) \times G(t+(n-1)\eta)$, to those of $I_{(n-1)\eta}(t)$, which symbolically is

$$I_{n\eta}(t) = I_{(n-1)\eta}(t) + E^G(t+(n-1)\eta) \times G(t+(n-1)\eta) \tag{14}$$

Definition 9 (Averaged-Distance Direct Link Graphs). For delay tolerance $\delta \in \mathbb{N}^+\eta$, we define a graph $D_\delta(t)$, called the averaged-distance direct link graph at time t, where the weight of edge $e = (i,j) \in D_\delta(t)$ represents the averaged distance of the direct links, i.e., one-hop journeys, between nodes i and j, from time t to $t+\delta$.

For each pair of nodes (i,j), within the delay tolerance δ, the averaged distance of direct link or the weight of edge $e = (i,j) \in D_\delta(t)$ is equal to the summed distance of direct links between the pair divided by the corresponding total or summed duration of links. Symbolically, we have

$$D_\delta(t) = I_\delta(t) \div L_\delta(t) \tag{15}$$

where the operator $\div$ is the element-wise division. It is notable that if an element (i,j) in $L_\delta(t)$ is 0 ($I_\delta(t)_{(i,j)}$ should also be 0), we define $D_\delta(t)_{(i,j)} = \infty$.

Definition 10 (Indirect Distance Graphs). For delay tolerance $\delta \in \mathbb{N}^+\eta$, we define a graph $M_\delta(t)$, called the indirect distance graph at time t, where the weight of the edge $e = (i,j) \in M_\delta(t)$ represents the minimum summed-distance of the multi-hop links from nodes i to j through all possible journeys, from time t to $t+\delta$.

Definition 11 (Minimum Distance Graphs). For delay tolerance $\delta \in \mathbb{N}^+\eta$, we define a graph $B_\delta(t)$, called the minimum distance graph at time t, where the weight of the edge $e = (i,j) \in B_\delta(t)$ represents the minimum distance to transmit between nodes i and j among all the possible paths, both direct and multi-hop links, from time t to $t+\delta$.

For each pair of nodes, the minimum distance is the smaller one of the averaged distance of direct link and the indirect distance, since the nodes always choose the path with the lowest cost. Therefore, symbolically we have

$$B_\delta(t)_{(i,j)} = \min\left\{D_\delta(t)_{(i,j)}, M_\delta(t)_{(i,j)}\right\} \tag{16}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

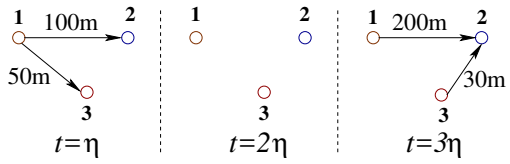


Fig. 6. Illustrative example for the relationships among various graphs, assuming $\delta = 3\eta$.

Compute Indirect Distance Graphs $M_\delta(t)$

For delay tolerance $\delta = n\eta$, there are a total of $n-1$ types of indirect transmission from node i to node j. Supposing that data are relayed at time $t + k\eta$, where $1 \le k \le n-1$, data can be transmitted from node i to an intermediate node w between time t and $t + k\eta$, and then be transmitted from node w to j between time $t + k\eta$ and $t + \delta$. These two transmission processes can be either through direct links or via multi-hop links, and we simply define the indirect distance as the sum of the distances in these two processes. Thus we need to consider all possible intermediate timestamps $t + \eta \le t + k\eta \le t + (n-1)\eta = t + \delta - \eta$ and also all possible intermediate nodes $1 \le w \le N$ to obtain the indirect path that is the shortest. Hence, symbolically, the indirect distance graph can be calculated from $M_\delta(t)_{(i,j)} =$

$$\min_{1\le k\le n-1}\left\{\min_{1\le w\le N}\left\{B_{k\eta}(t)_{(i,w)} + B_{\delta-k\eta}(t+k\eta)_{(w,j)}\right\}\right\} \tag{17}$$

We use the simple example depicted in Fig. 6 to illustrate the relationships among the various graphs defined in Definitions 7 to 11. Consider the edge $e = (1, 2)$ for example, we have

- The total duration of links or weight of $L_{3\eta}(\eta)_{(1,2)}$: $1 + 0 + 1 = 2$ (Definition 7).
- The summed distance of direct link or weight of $I_{3\eta}(\eta)_{(1,2)}$: $100\text{m} + 200\text{m} = 300\text{m}$ (Definition 8).
- The averaged distance of direct link or weight of $D_{3\eta}(\eta)_{(1,2)}$: $\frac{\text{weight of } I_{3\eta}(\eta)_{(1,2)}}{\text{weight of } L_{3\eta}(\eta)_{(1,2)}} = \frac{300\text{m}}{2} = 150\text{m}$ (Definition 9).
- There exists one journey from node 1 to node 3 at $t = \eta$ and then from node 3 to node 2 at $t = 3\eta$. Thus, the indirect minimum distance or weight of $M_{3\eta}(\eta)_{(1,2)}$: $50\text{m} + 30\text{m} = 80\text{m}$ (Definition 10).
- The minimum distance for all possible links or weight of $B_{3\eta}(\eta)_{(1,2)}$: $\min\{150\text{m}, 80\text{m}\} = 80\text{m}$ (Definition 11).

Compute ORG

We use the minimum distance graphs, which take both direct transmission and indirect transmission into account, to represent the opportunistic reachability graphs (ORGs), i.e.

$$E^R_\delta(t) = B_\delta(t) \tag{18}$$

Finally, the presence function $R_\delta(t)$ can be acquired from $E^R_\delta(t)$ according to

$$\forall (i,j) \in E^R_\delta(t):\ R_\delta(t)_{(i,j)} = \begin{cases} 1, & \text{if } E^R_\delta(t)_{(i,j)} < \infty \\ 0, & \text{if } E^R_\delta(t)_{(i,j)} = \infty \end{cases} \tag{19}$$

Note that $E^R_\delta(t)_{(i,j)} = \infty$ in (19) is symbolically used to indicate that the weight of $E^R_\delta(t)_{(i,j)}$ is infinitely large, i.e. no journey between i and j exists. Also we do not test the weight of the edge $E^R_\delta(t)_{(i,j)}$ against d, because d is the maximum transmission distance for direct link between two nodes. As long as the weight of $E^R_\delta(t)_{(i,j)}$ is finite, there must exist a journey from node i to j within the time period from t to $t+\delta$. Thus, we test by checking whether the weight of $E^R_\delta(t)_{(i,j)}$ is finite.

Algorithm 1 The algorithm to compute ORG.

Require: The weighed edges of instantaneous topology graph $E^G(t)$, number of nodes N, delay tolerance $\delta = n\eta$, maximum transmission distance d, and time span $\mathcal{T}$.

$L_0(t) = I_0(t) = 0$;
compute $G(t)$ according to (10).
for $(t = \eta; t < T; t += \eta)$ **do**
  $L_\eta(t) = G(t)$;
  $I_\eta(t) = E^G(t) \times G(t)$;
  $E^R_\eta(t) = I_\eta(t) \div L_\eta(t)$;
  for $(\delta = 2\eta; \delta \le n\eta; \delta += \eta)$ **do**
    $L_\delta(t) = L_{\delta-\eta}(t) + G(t+\delta-\eta)$;
    $I_\delta(t) = I_{\delta-\eta}(t) + E^G(t+\delta-\eta) \times G(t+\delta-\eta)$;
    $D_\delta(t) = I_\delta(t) \div L_\delta(t)$;
    for $(i = 1; i \le N; i++)$ **do**
      for $(j = 1; j \le N; j++)$ **do**
        $M_\delta(t)_{(i,j)} = \min_{m\le k\le n-m}\left\{\min_{1\le w\le N}\left\{B_{k\eta}(t)_{(i,w)} + B_{\delta-k\eta}(t+k\eta)_{(w,j)}\right\}\right\}$;
        $B_\delta(t)_{(i,j)} = \min\left\{D_\delta(t)_{(i,j)}, M_\delta(t)_{(i,j)}\right\}$;
      end for
    end for
    $E^R_\delta(t)_{(i,j)} = B_\delta(t)_{(i,j)}$;
  end for
end for
compute $R_\delta(t)$ according to (19).
return $E^R_\delta(t), R_\delta(t)$.

3.4 Algorithm

Given the instantaneous topology $\mathcal{G}(t)$, i.e., $E^G(t)$ and $G(t)$, and the time span $\mathcal{T}$, the delay tolerance $\delta = n\eta$, the maximum transmission distance d and the number of nodes N, we present the algorithm to calculate the ORG $\mathcal{R}_\delta(t)$ in Algorithm 1. In the algorithm, as can be mapped to Eq. (10) to (19), for each time t and each pair of nodes i and j, we calculate the minimum distances of direct transmission and indirect transmission, respectively, and choose the minimum distance as the smaller of the two. For indirect transmission, in particular, we need to consider all the possible intermediate nodes from 1 to N and all the possible divisions of time from $k = 1$ to $n-1$, in order to find the smallest case. The complexity of Algorithm 1 is on the order of $\mathsf{O}(n^2TN^3)$. Although the computation cost is cubic in the number of taxis, the real runtime is not so high due to the sparse TVG and ORG where each taxi is only able to connect a few nearby taxis.

4 Opportunistic Topology Modeling

We mainly use Shanghai trace in our modeling, but Beijing and Nanjing traces are also employed in evaluation. Hence,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

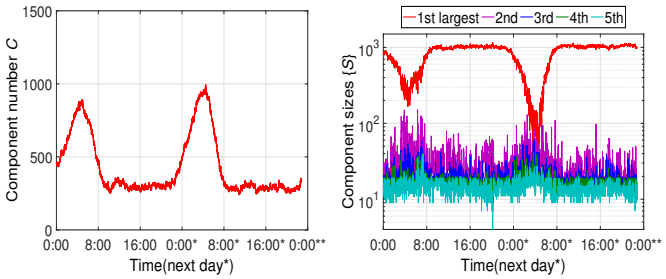


Fig. 7. Evolution of (a) the component number and (b) the component sizes of the top 5 largest components, over 2 days.

unless otherwise specifically stated, Shanghai trace is used. To show our results, we randomly pick 2 days for Shanghai(Feb 5 and 6, 2007) and Beijing(May 5 and 6, 2012) trace and 1 day for Nanjing(Jul 3, 2013) trace. For Nanjing trace, we only use the data in one single day just to verify quickly on our models.

4.1 Component Number

The connectivity of a taxi network is mainly characterized by the two metrics: the number of components, denoted by C, which reflects the level of network fragmentation, and the component sizes, denoted simply by the set $\{S\}$, which describes the heterogeneity of the fragmentation.

For simplicity, we limit the size of the messages so that they could be able to transmitted within one unit period through V2V direct link. We also assume that the temporal resolution η is $20s$, which equals to the sampling frequency after data preprocessing as mentioned in Section. 3.2.

We initially consider the case of the network parameters $d = 100\,\text{m}$ and $\delta = 10\,\text{min}$, as 100 m is the distance found by field tests as a typical distance for reliable V2V DSRC, which ensures a packet delivery ratio of around 80% in urban environments, under common power levels of 15-20 dBm as well as for the BPSK modulation of 3 Mbps and the QPSK modulation of 6 Mbps [16], [17], [18]. Then, we extend our study to $d = 50\,\text{m}$, which is identified as the largest distance at which V2V communication allows almost 100% of the packets to be correctly received [16], [17], [18], $d = 200\,\text{m}$, which is the maximum distance with a reception ratio above 0.5 [16], [18], and $d = 500\,\text{m}$, the maximum distance for vehicular communication, as well as changing the network delay tolerance from 10 minutes to 1 hour, to analyze its impact on the network connectivity.

Analysis for $d = 100\,m$ and $\delta = 10\,min$

The evolutions of the component number C and the component sizes $\{S\}$ of the top five largest components are depicted in Fig. 7 (a) and (b), respectively, which are aggregated over 2 days and extracted from 1500 active nodes of road traffic. It can be seen from Fig. 7 (a) that the component number C takes a value of around 300 during the daytime and has higher varying values at night. The results of Fig. 7 clearly show the variation of the network connectivity over the time. In particular, we observe that the number of components is stable during daytime and very dynamic at night. Also, the network is highly heterogeneous, and the largest component makes up the two thirds of the whole network during the daytime.

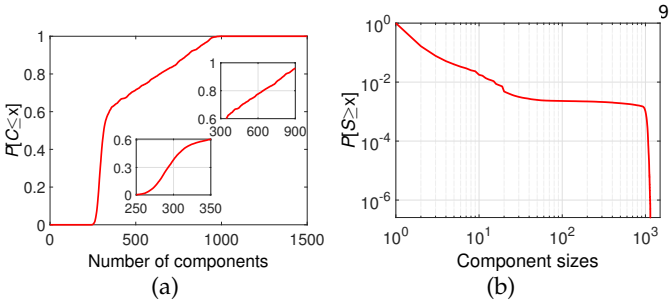


Fig. 8. Distribution of (a) the number of components and (b) the component sizes, aggregating all the samples over two days.

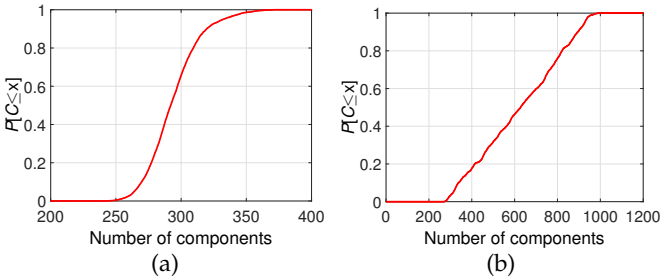


Fig. 9. Distribution of the number of components when aggregating all the samples: (a) over daytime, and (b) over nighttime.

The cumulative distribution function (CDF) of C is portrayed in Fig. 8 (a), where we observe that the CDF curve has two growing parts. In two thirds of all the cases, the taxi networks have 250 to 350 components with a relatively fast growth. This part of the curve shows the relatively stable connectivity of the network during the daytime with relatively fewer components, implying that more nodes aggregate together to form a component. The slow-growing part of the CDF curve represents the connectivity of the network at night, suggesting that the network connectivity is varying and nodes are more separated than during the daytime. Fig. 8 (b) shows the complementary cumulative distribution function (CCDF) of $\{S\}$, where we find that the network largely consists of singletons with a percentage of over 80%. Moreover, nearly 99% of the components have only 15 taxis or less, while the larger components only account for 1% but they make up almost the whole network.

We now focus on the intriguing differences in the taxi networks, in terms of network fragmentation, between the daytime and the nighttime. We consider 8:30 to 22:00 as daytime and the rest as nighttime, and we plot the CDFs of C during the daytime and the nighttime in Fig. 9 (a) and (b), respectively. It can be seen that the CDF of the component number C is almost linear. Furthermore, the slope of the CDF during the daytime is larger than that during the nighttime. This again implies that the networks have a relatively stable connectivity during the daytime, while the networks are more fragmented at night.

Analysis for different d and different δ

We next study how different transmission distances d and delay tolerances δ impact on the network connectivity. These two factors are important since they are the range constraint and time constraint related to the mobility of taxis. The main plot in Fig. 10 (a) portrays the CDFs of the number of components C during the daytime when the transmission

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

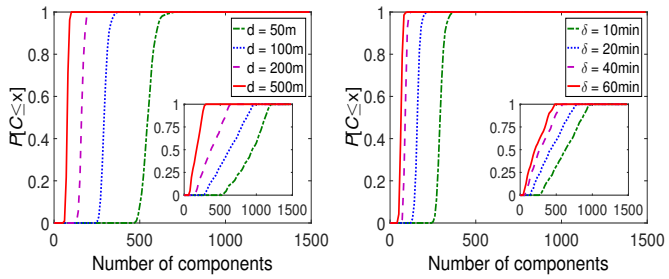


Fig. 10. Distribution of the number of components C when aggregating over all the samples of two days: (a) the CDFs for different transmission distances d with the delay tolerance $\delta = 10$ min, and (b) the CDFs for different delay tolerance δ with the transmission distance $d = 100$ m. The main plot is for the daytime, while the subplot inside is for the nighttime.

distance d varies from 50 m to 500 m with the delay tolerance $\delta = 10$ min, while the subplot inside Fig. 10 (a) shows the CDFs of C during the nighttime with varying d from 50 m to 500 m and a fixed $\delta = 10$ min. Similarly, the main plot in Fig. 10 (b) depicts the CDFs of the number of components C during the daytime when the delay tolerance δ changes from 10 min to 60 min with the transmission distance $d = 100$ m, while the the subplot inside Fig. 10 (b) shows the CDFs of C at during the nighttime under the same network parameters δ and d. It can be seen from Fig. 10 that the slope of the CDF for the daytime is large than the slope of the CDF for the nighttime given d and δ, which again confirms the result of Fig. 7 (a).

Moreover, from Fig. 10 (a), we observe that the slope of the CDF increases, while the intercept point of the CDF with the x-axis becomes smaller, as the transmission distance increases when fixing the δ value, and this is true for both the daytime and nighttime cases. Since the slope or derivative of the CDF is related to the probability density function (PDF) of C while the intercept point of the CDF with the x-axis is the minimum component number $C_{\min}$, we can see that with the fixed delay tolerance δ, the number of components decreases with the increase of the transmission distance. Additionally, the maximum component number $C_{\max}$, which is the point that the CDF reaches the maximum value of 1, also decreases with the increase of the transmission distance, and moreover the range $C_{\max} - C_{\min}$ gets smaller with the increase of d. The implication is that if devices support communication in longer distance, they are more likely to merge into a component in which devices can communicate with each other and, therefore, the number of components is smaller and more stable. Similarly, from Fig. 10 (b), we observe that with the fixed transmission distance d, the number of components decreases with the increase of the delay tolerance. Again this is true for the both daytime and nighttime cases.

We also generalize our models to Beijing trace with the results shown in Fig. 11. It can be seen from Fig. 11 that the number of components varies, from about 500, given $d = 500$ m and $\delta = 10$ min, to about 1,300, given $d = 50$ m and $\delta = 10$ min. Additionally, given $d = 100$ m, the number of components increases from 700 to about 1,100, when δ increases from 10 min to 60 min. Therefore, the results based on both Shanghai and Beijing traces confirm that d and δ are the two factors that impact on the CDF of C significantly.

The above results agree well with our intuition for

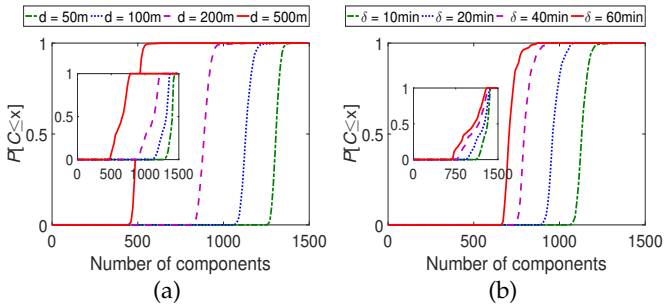


Fig. 11. Distribution of the number of components C when aggregating over all the samples of two days in Beijing trace: (a) the CDFs for different transmission distances d with the delay tolerance $\delta = 10$ min, and (b) the CDFs for different delay tolerance δ with the transmission distance $d = 100$ m. The main plot is for the daytime, while the subplot inside is for the nighttime.

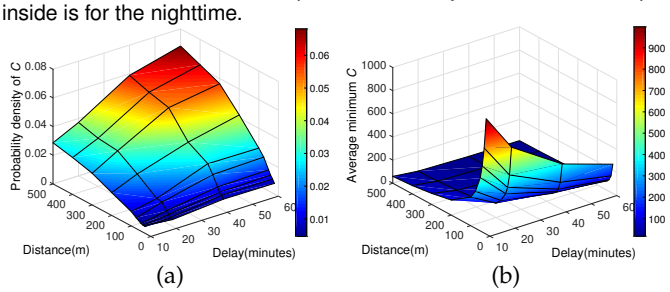


Fig. 12. (a) Probability density function of C and (b) minimum component number $C_{\min}$, aggregating over all the samples of daytime.

opportunistic topology. In particular, they offer us useful insights on the relationship between the CDF of C and the two key network parameters, d and δ. Base on these measurements and extracted knowledge, we now build a model to predict the CDF of C for given d and δ. Such a model is extremely valuable in analyzing the ORG and in further studying envisaged vehicular-based communication networks.

Uniform distribution approximation

The above empirical results extracted from both Shanghai and Beijing trace measurements clearly indicate that the slope of the CDF of C, i.e., the PDF of C, is an increasing function of d and δ, while the intercept point of the CDF with the x-axis, i.e., the minimum component number $C_{\min}$, is a decreasing function of d and δ. For Shanghai trace, we additionally plot the PDF of C and the minimum component number $C_{\min}$ in the case of daytime in Fig. 12 (a) and (b), respectively, as the functions of d and δ. Also from both Figs. 10 and 11, we observe that the CDF of C in the case of nighttime can be accurately represented by a linear function between $C_{\min}$ and $C_{\max}$. In the case of daytime, the CDF is more nonlinear but nevertheless can also be reasonably approximated by a linear function.

To know the CDF of component number C for any arbitrary d and δ, we propose the uniform distribution approximation to avoid computing ORG and the connected component again. This approximation can further reveal the relation between the CDF of component number C and the network parameters like d and δ.

The PDF of C is modeled as follows,

$$\mathrm{PDF}(C; d, \delta) = \begin{cases} S_{\mathrm{CDF}}(d, \delta), & C_{\min}(d, \delta) \le C \le C_{\max}(d, \delta) \\ 0, & \text{otherwise} \end{cases} \quad (20)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

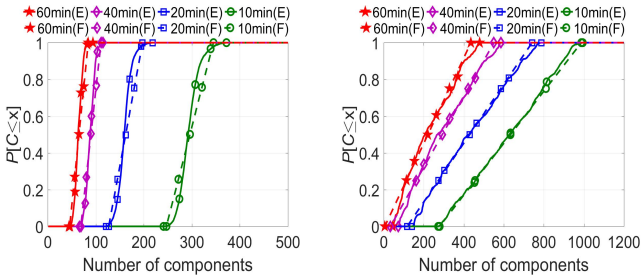


Fig. 13. Comparison of the empirical CDF(E) of C with the fitted CDF(F) based on the uniform distribution model (20) for different δ and $d = 100$ m: (a) the daytime case, and (b) the nighttime case.

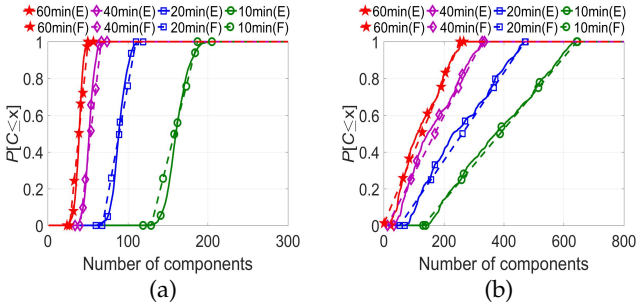


Fig. 14. Comparison of the empirical CDF(E) of C with the fitted CDF(F) based on the uniform distribution model (20) for different δ and $d = 200$ m: (a) the daytime case, and (b) the nighttime case.

In the uniform distribution (20), there are only two independent parameters $S_{\text{CDF}}(d,\delta)$, which is the derivative or slope of the CDF, and $C_{\min}(d,\delta)$, since

$$C_{\max}(d,\delta) = C_{\min}(d,\delta) + \frac{1}{S_{\text{CDF}}(d,\delta)} \tag{21}$$

Both $S_{\text{CDF}}(d,\delta)$ and $C_{\min}(d,\delta)$ are clearly the functions of d and δ. We use the polynomial models to fit $S_{\text{CDF}}(d,\delta)$ and $C_{\min}(d,\delta)$ according to

$$\begin{aligned} S_{\text{CDF}}(d,\delta) =& a_1 + a_2 d + a_3\delta + a_4 d^2 + a_5 d\,\delta + a_6\delta^2 \\ &+ a_7 d^3 + a_8 d^2\delta + a_9 d\,\delta^2 + a_{10}\delta^3 \end{aligned} \tag{22}$$

$$\begin{aligned} C_{\min}(d,\delta) =& b_1 + \frac{b_2}{d} + \frac{b_3}{\delta} + \frac{b_4}{d^2} + \frac{b_5}{d\,\delta} + \frac{b_6}{\delta^2} + \frac{b_7}{d^3} \\ &+ \frac{b_8}{d^2\delta} + \frac{b_9}{d\,\delta^2} + \frac{b_{10}}{\delta^3} \end{aligned} \tag{23}$$

Specifically, we use the data to fit the polynomial coefficients a_i and b_i. From the above empirical distribution results, we note that when the product $d \cdot \delta$ is small, the CDF of C in the daytime exhibits notable nonlinearity, see for example Fig. 9 (a). In such a case, the uniform distribution will not be an accurate approximation to the true empirical distribution. Therefore, we only use the data with $d \cdot \delta > 1000\,\text{m} \cdot \text{min}$ to fit the uniform distribution model.

With $d = 100$ m and various delay tolerance values δ, Fig. 13 (a) and (b) compare the empirical CDFs(E) of C from the data with the fitted CDFs(F) based on our uniform distribution model for the cases of daytime and nighttime, respectively. Similarly, Fig. 14 (a) and (b) compare the data based empirical CDFs(E) with our model fitted CDFs(F) for the daytime and nighttime cases, respectively, with $d = 200$ m and various delay tolerance values δ. From Fig. 13 (b) and Fig. 14 (b), it can be seen that our uniform distribution model (20) is an accurate model for the data

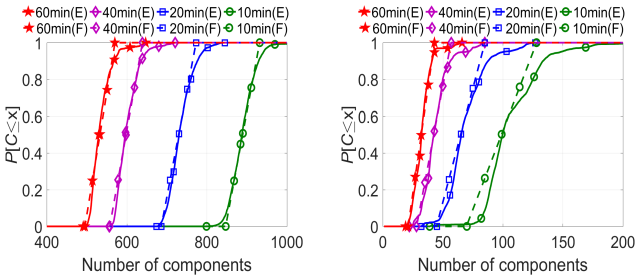


Fig. 15. Comparison of the empirical CDF(E) of C with the fitted CDF(F) based on the uniform distribution model (20) for different δ and $d = 200$ m, in (a) Beijing trace, and (b) Nanjing trace.

based empirical distribution, for the case of nighttime. As for the daytime situation, with the exception of $d = 100$ m and $\delta = 10$ min, our uniform distribution model (20) fits the empirical distribution reasonably well, particular for large value of the network parameter product $d \cdot \delta$.

We further apply the uniform distribution approximation on the Beijing and Nanjing traces during the daytime. Because of the different environments, we cannot use the same coefficients, i.e. as and bs, to approximate. Instead, for each city, we trained the approximation model with some ds and δs and tested them with more other configurations. In Fig. 15 (a) and (b), almost every CDF of component number in the two cities is approximated really well based on our model, as it is in Shanghai trace.

We also use the correlation-based similarity measure between the empirical CDF and the fitted CDF of the uniform distribution model to evaluate the goodness of the fitted CDF model. For notational convenience, denote the empirical CDF as $P^{(\text{e})}(c)$ and the fitted CDF as $\widehat{P}(c;d,\delta)$. Given a test data set of $\{c_1, c_2, \cdots, c_{K_\text{t}}\}$, first compute the means and variances of $P^{(\text{e})}(c)$ and $\widehat{P}(c;d,\delta)$, respectively, over the test data set as

$$\bar{P}^{(\text{e})} = \frac{1}{K_\text{t}} \sum_{k=1}^{K_\text{t}} P^{(\text{e})}(c_k) \tag{24}$$

$$\bar{\widehat{P}} = \frac{1}{K_\text{t}} \sum_{k=1}^{K_\text{t}} \widehat{P}(c_k;d,\delta) \tag{25}$$

$$\sigma^2_{P^{(\text{e})}} = \frac{1}{K_\text{t}} \sum_{k=1}^{K_\text{t}} \left(P^{(\text{e})}(c_k) - \bar{P}^{(\text{e})}\right)^2 \tag{26}$$

$$\sigma^2_{\widehat{P}} = \frac{1}{K_\text{t}} \sum_{k=1}^{K_\text{t}} \left(\widehat{P}(c_k;d,\delta) - \bar{\widehat{P}}\right)^2 \tag{27}$$

Then the correlation-based similarity measure between $P^{(\text{e})}(c)$ and $\widehat{P}(c;d,\delta)$ is computed according to

$$\text{SM}_{\text{cor}} = \frac{1}{K_\text{t}} \sum_{k=1}^{K_\text{t}} \frac{\left(\widehat{P}(c_k;d,\delta) - \bar{\widehat{P}}\right)\left(P^{(\text{e})}(c_k) - \bar{P}^{(\text{e})}\right)}{\sigma_{P^{(\text{e})}}\sigma_{\widehat{P}}} \tag{28}$$

$\text{SM}_{\text{cor}} = 1$ indicates that the two CDFs are completely similar, while $\text{SM}_{\text{cor}} = 0$ indicates that the two CDFs are completely dissimilar. More specifically, let $C^{(\text{e})}_{\min}$ and $C^{(\text{e})}_{\max}$ be the minimum and maximum component numbers specified by the empirical CDF, respectively, while $C^{(\text{f})}_{\min}$ and $C^{(\text{f})}_{\max}$ are the minimum and maximum component numbers defined by the fitted CDF, respectively. Further define

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

TABLE 2
The correlation-based similarity measure (daytime case/nighttime case) between the empirical CDF and the fitted CDF for various network parameters d and δ.

δ/d	50 m	100 m	200 m	500 m
10 min	/	/	0.99/1.00	0.95/1.00
15 min	/	0.98/1.00	0.98/1.00	0.98/1.00
20 min	/	0.98/1.00	0.99/1.00	0.98/1.00
25 min	0.96/1.00	0.98/1.00	0.99/0.99	0.98/1.00
30 min	0.97/1.00	0.98/1.00	0.99/0.99	0.98/1.00
35 min	0.98/1.00	0.98/1.00	0.98/0.99	0.99/0.99
40 min	0.99/1.00	0.98/1.00	0.98/0.99	0.99/0.99
45 min	0.99/1.00	0.99/1.00	0.99/0.99	0.98/0.99
50 min	0.99/1.00	0.99/0.99	0.99/0.99	0.99/0.99
55 min	0.99/1.00	0.99/0.99	0.99/0.99	0.99/0.99
60 min	0.98/1.00	0.98/0.99	0.98/0.99	0.99/0.99

TABLE 3
The Kullback-Leibler divergence measure (daytime case/nighttime case) between the empirical PDF and the fitted PDF for various network parameters d and δ.

δ/d	50 m	100 m	200 m	500 m
10 min	/	/	0.06/0.07	0.10/0.04
15 min	/	0.13/0.08	0.10/0.09	0.09/0.05
20 min	/	0.12/0.09	0.09/0.10	0.08/0.07
25 min	0.14/0.09	0.10/0.08	0.08/0.09	0.08/0.08
30 min	0.12/0.08	0.10/0.09	0.08/0.10	0.08/0.09
35 min	0.10/0.08	0.09/0.09	0.08/0.10	0.05/0.09
40 min	0.08/0.09	0.08/0.08	0.07/0.10	0.05/0.10
45 min	0.07/0.09	0.08/0.09	0.07/0.12	0.07/0.12
50 min	0.08/0.10	0.08/0.10	0.07/0.11	0.06/0.13
55 min	0.12/0.11	0.08/0.10	0.07/0.12	0.06/0.14
60 min	0.16/0.12	0.09/0.10	0.09/0.12	0.06/0.13

$C_{\min} = \min\{C^{(\mathrm{e})}_{\min}, C^{(\mathrm{f})}_{\min}\}$ and $C_{\max} = \max\{C^{(\mathrm{e})}_{\max}, C^{(\mathrm{f})}_{\max}\}$. Then our test data set is given by $C \in \{C_{\min}, C_{\min}+1, \cdots, C_{\max}\}$. Table 2 lists the correlation-based similarity measure values between the empirical CDF and the fitted CDF of the uniform distribution model for different network parameters d and δ. It can be seen that for the daytime situation, the correlation-based similarity measure between the empirical CDF and our fitted CDF is at least 0.96 or higher, and this indicates the accuracy of our fitted CDF model. For the nighttime situation, our fitted CDF model is even more accurate, as we have the correlation-based similarity measure values of over 0.99 in all the cases.

The Kullback-Leibler divergence (KLD) measures the dissimilarity of two PDFs, and it can also be used to validate the goodness of the fitted PDF distribution. More specifically, the KLD between the empirical PDF $p^{(\mathrm{e})}(c)$ and the fitted PDF $\widehat{p}(c; d, \delta)$ can be approximated by

$$\mathrm{DM}_{\mathrm{KL}} = \sum_{k=2}^{K_{\mathrm{t}}} p^{(\mathrm{e})}(c_k) \log \frac{p^{(\mathrm{e})}(c_k)}{\widehat{p}(c_k; \boldsymbol{\alpha})} (c_k - c_{k-1}) \qquad (29)$$

where we have to obtain the empirical PDF samples $p^{(\mathrm{e})}(c_k)$ by differentiating the empirical CDF samples $P^{(\mathrm{e})}(c_k)$. The smaller the value of $\mathrm{DM}_{\mathrm{KL}}$ computed in (29), the closer the two distributions are. Clearly, the KLD measure (29) and the correlation based similarity measure (28) are equivalent. A drawback of the KLD measure (29) is that differentiating the empirical CDF samples amplifies the noise in the data. In Table 3, we list the KLD measure values between the empirical PDF and our fitted PDF for different network parameters d and δ. The results of Table 3 agree with the results of Table 2.

In summary, the component number C is steady around 300 during daytime but higher and changing a lot at nighttime. We also successfully use the same uniform model (with different coefficients) to characterize the connectivity both in daytime and nighttime – the CDF of the component number can be accurately approximated by the uniform distribution.

This distribution model provides a very convenient and powerful tool to simulate the opportunistic topology of taxi networks and to investigate the impact of the network parameters, i.e., transmission distance and delay tolerance, on the network connectivity. Together with our TVG and ORG, the developers with our introduced characteristics can have a better design on the delay tolerant network, like estimating the delay of messages, tunning the critical paths and relay nodes, etc.

TABLE 4
Percentage of different sizes of components and the percentage of component in urban area (in parentheses), which is corresponding to Fig. 16.

Configurations	Mini	Medium	Large
(a)	100.0%(38.9%)	0.0%(/)	0.0%(/)
(b)	97.3%(35.6%)	2.7%(51.9%)	0.0%(/)
(c)	95.1%(28.7%)	4.6%(40.6%)	0.3%(100.0%)
(d)	96.1%(22.5%)	3.7%(22.2%)	0.2%(100.0%)
(e)	97.0%(41.8%)	2.9%(25.0%)	0.1%(100.0%)
(f)	97.3%(38.8%)	2.7%(51.9%)	0.0%(/)
(g)	96.6%(34.9%)	3.3%(44.8%)	0.1%(100.0%)

4.2 Geographical Distribution of Components

We now study the geographical distribution of the components in taxi networks based on Shanghai trace to further investigate the impact of the network parameters, transmission distance and delay tolerance, on the number of components and component sizes. In Fig. 16(a)-(d), we depict the snapshots of the taxi networks fragmentation calculated from Shanghai trace at 8:00 AM, a typical heavy-traffic time, under different d and δ, where each chromatic circle corresponds to a component whose size is represented by the size and color of the circle. If delay is not permitted, almost all the nodes are isolated, as can be seen from Fig. 16 (a). In Fig. 16 (b) and (c), the delay tolerances are both 6 minutes, but they have different transmission distances. Longer transmission distance offers more opportunities of communication, and therefore there are larger components in the network in Fig. 16 (c) for $d = 100$ m than in Fig. 16 (b) with $d = 50$ m. On the other hand, Fig. 16 (c) and (d) show the geographical distribution snapshots for two different δ given the same transmission distance. When the delay tolerance is increased, big components swallow up small ones, and in particular, the largest component becomes even larger.

We also pick up another time in the same day and the same time in the second day with $\delta = 6min$ and $d = 50m$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

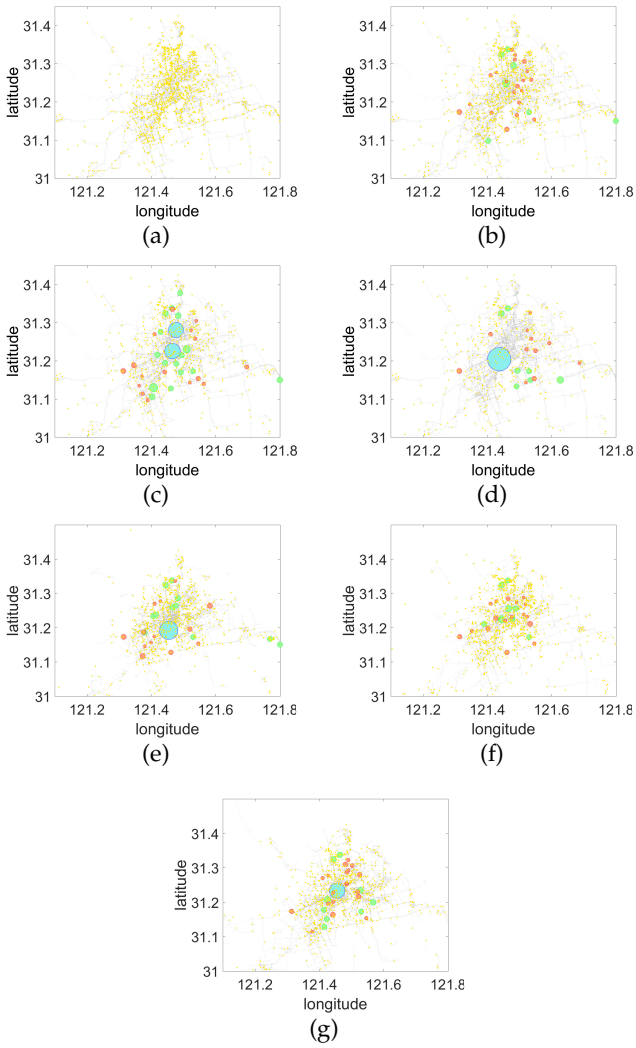


Fig. 16. Geographical distribution snapshot of the network components: (a) instantaneous topology at 8:00 AM, day 1, (b) $\delta = 6$ min, $d = 50$ m at 8:00 AM, day 1, (c) $\delta = 6$ min, $d = 100$ m at 8:00 AM, day 1, (d) $\delta = 10$ min, $d = 100$ m at 8:00 AM, day 1, (e) $\delta = 6$ min, $d = 50$ m at 12:00 PM(noon), day 1, (f) $\delta = 6$ min, $d = 50$ m at 0:00 AM(midnight), day 1 and (g) $\delta = 6$ min, $d = 50$ m at 8:00 AM, day 2. Each chromatic circle corresponds to a component whose size is represented by the size and color of the circle.

The results, as shown in Fig. 16(e)-(g), demonstrates that the distribution and size of the components varies from time to time, but the taxi networks are always composed of several really large components in urban area, some surrounding medium component and lots of isolated mini component. To be mentioned, no large components exist in the midnight, as shown in Fig. 16(f), and there are some smaller components instead in the urban area.

We classify the components into three classes according to their sizes. The components with no more than 5 nodes are deemed mini components, and the large components have more than 100 nodes, while the rest components with the sizes between 6 to 99 are called medium components. Table 4 shows the percentages of the three classes of components, under different d and δ. It can be seen that the mini components predominate the networks, but they may not form the main part of the taxi networks, particularly for the taxi networks with the transmission distance and

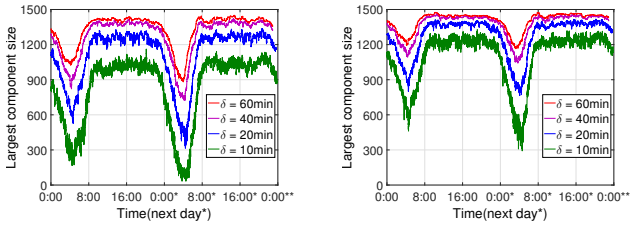


Fig. 17. Evolution of the largest component size $S_{\max}$ over 2 days for different delay tolerances and given the transmission distance: (a) $d = 100$ m, and (b) $d = 200$ m.

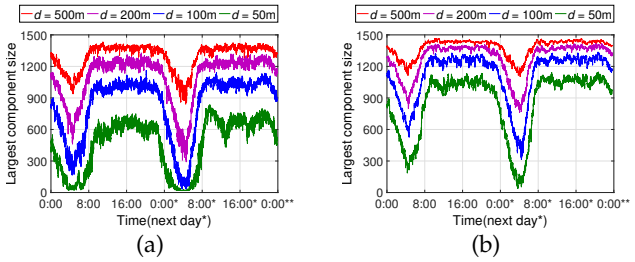


Fig. 18. Evolution of the largest component size $S_{\max}$ over 2 days for different transmission distances and given the delay tolerance: (a) $\delta = 10$ min, and (b) $\delta = 20$ min.

delay tolerance constraints in Table 4. The large components actually take up the majority of the nodes in these situations.

In each class of component size, we show the percentage of components who are located mainly in urban area in the parentheses in Table 4. Large components, if exists in the taxi networks, are by no doubt located in urban area. However, only 40% to 50% of the medium components and 20% to 40% of the mini components are located in urban area when $\delta = 6min$. After the delay tolerance increases to 10 minutes. The percentage of mini and medium components located in the urban area drops to 22%. This significant drop cross-validate our previous conclusion that the largest component in urban area swallows up the mini and medium component.

Thus from the results in Figs. 16 and Table 4, we observe that when the transmission distance and delay tolerance increase, first the numbers of medium components and large components will increase, and then the largest component begins to swallow up many smaller ones which leads to the decrease of the total number of components. The reasons behind these observations can be inferred as follows. Firstly and obviously, when transmission distance and delay tolerance are small, nodes are largely isolated and components are really small. Secondly, when the transmission distance and delay tolerance are enlarged to $d = 100$ m and $\delta = 6$ min, isolated nodes are able to link to the near-by medium and large components, making them larger. Finally, after the delay tolerance is further increased to $\delta = 10$ min, some of the nodes in the medium components are able to link to the largest component. Thus, they merge together, greatly increasing the size of the largest component and reducing the number of components.

4.3 Properties of Largest Component

To study the intriguing structures and properties of the highly heterogeneous network topology, we cast our eyes on the largest component, as it takes up the main part of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

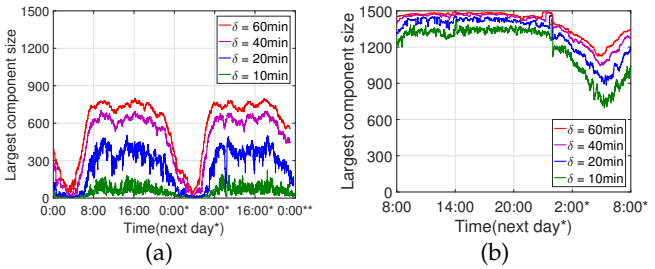


Fig. 19. Evolution of the largest component size $S_{\max}$ over 2 days for different delay tolerances and given the transmission distance $d = 100$ m: (a)for Beijing trace, and (b)for Nanjing trace.

whole network, and most multi-hop communications take place in it.

Component dynamics versus delay tolerance and transmission distance

Fig. 17 (a) and (b) portray the evolutions of the largest component size $S_{\max}$ for different delay tolerances and with the transmission distance $d = 100\,\text{m}$ and $d = 200\,\text{m}$, respectively, while Fig. 18 (a) and (b) depict the evolutions of $S_{\max}$ for different transmission distances and with the delay tolerance $\delta = 10\,\text{min}$ and $\delta = 20\,\text{min}$, respectively. As expected, $S_{\max}$ is an increasing function of δ and d. In general, given d, the largest component shrinks heavily when decreasing δ. Similarly, given δ, $S_{\max}$ decreases considerably when reducing d. Additionally, the largest component size during the daytime is more stable and much larger than the largest component size during the nighttime. For example, with $d = 500\,\text{m}$ and $\delta = 10\,\text{min}$, during the daytime the largest component is really big, taking up above 90% of the whole network, while during the nighttime, $S_{\max}$ shrinks heavily to a merely 60% of the whole network, as confirmed in Fig. 18 (a). With $d = 100\,\text{m}$ and $\delta = 20\,\text{min}$, the largest component takes up 80% of the whole network during the daytime, and $S_{\max}$ reduces to 40% of the network at night, as can be seen from Fig. 18 (b).

We also study on the largest component in taxis networks in other cities. Fig. 19 (a) and (b) show the evolutions of $S_{\max}$ obtained from Beijing and Nanjing traces, respectively, using different delay tolerances and with the transmission distance $d = 100\,\text{m}$. In addition to confirm the general observations based on Shanghai trace, the results of Fig. 19 also indicate that there are differences among the results extracted from Shanghai, Beijing and Nanjing traces. To be specific, given $d = 100\,\text{m}$ and $\delta = 20\,\text{min}$, for Beijing trace, the size of the largest component can only reach to about 40% of the nodes in the network during the daytime, which is much less than the figure of 80% for Shanghai trace, while for Nanjing trace, the largest component takes up over 90% of the network during the daytime. These interesting facts indicate that the taxis in Shanghai and Nanjing are much easier to form a huge component due to the unique geological characteristics of these two cities, different to those of Beijing.

Since the dynamics of the largest component are very different during the daytime than during the nighttime, in Fig. 20 (a) and (b), we plot the ratios of the average largest component size $S_{\max}$ to the network size N during the daytime and the nighttime, respectively, based on Shanghai trace. Clearly, the largest component expands with the

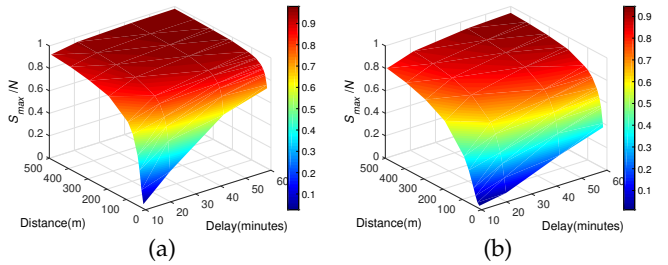


Fig. 20. The ratio of the average largest component size $S_{\max}$ to the network size N: (a) during the daytime, and (b) during the nighttime.

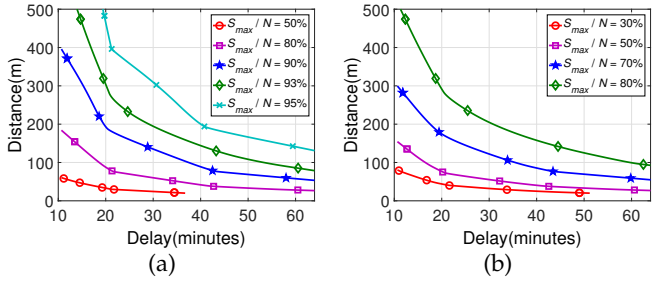


Fig. 21. Estimated network parameters of delay tolerance and transmission distance to achieve different percentages of $S_{\max}/N$: (a) during the daytime, and (b) during the nighttime.

increase of delay and transmission distance both during the daytime and nighttime. For the sake of clarity, we further plot the curves of the network parameters, d and δ, to achieve different percentages of $S_{\max}/N$ during the daytime and the nighttime in Fig. 21 (a) and (b), respectively. From Fig. 21 (a), for example, we can infer that the network will have a giant component with 80% of the nodes during the daytime, given $d = 100\,\text{m}$ and $\delta = 20\,\text{min}$ or $d = 50\,\text{m}$ and $\delta = 40\,\text{min}$. At night and under the similar network parameters, the largest component only takes 50% of the nodes.

Component dynamics versus network size N

We also study how the number of network nodes N impacts on the size of the largest component $S_{\max}$. As a larger network surely leads to larger components, we again analyze the ratio of $S_{\max}/N$. Fig. 22 (a), (b) and (c) depict the evolutions of the ratio of the largest component size to the network size, $S_{\max}/N$, for different N under the network conditions of a) $\delta = 10\,\text{min}$ and $d = 200\,\text{m}$, b) $\delta = 20\,\text{min}$ and $d = 100\,\text{m}$, and c) $\delta = 20\,\text{min}$ and $d = 200\,\text{m}$, respectively. Intriguingly, we observe that the percentage of the largest component in the networks increases as the size of the network N increases. For example, given the network condition of $\delta = 10\,\text{min}$ and $d = 200\,\text{m}$, during the daytime the largest component only takes around 30% of the network's nodes when $N = 500$, but the largest component reaches over 80% of the whole network when $N = 1500$. Under the network condition of $\delta = 20\,\text{min}$ and $d = 200\,\text{m}$, the ratios $S_{\max}/N$ are approximately 55% and 90%, respectively, for $N = 500$ and $N = 1500$, during the daytime. Also the ratio $S_{\max}/N$ is much larger and more stable during the daytime than during the nighttime.

Another important insight we can infer from Fig. 22 is that there is an intuitive limitation of $S_{\max}/N$. Specifically, for $\delta = 10\,\text{min}$ and $d = 200\,\text{m}$ as well as for $\delta = 20\,\text{min}$ and $d = 100\,\text{m}$, $S_{\max}/N$ saturates at around 80% as N

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

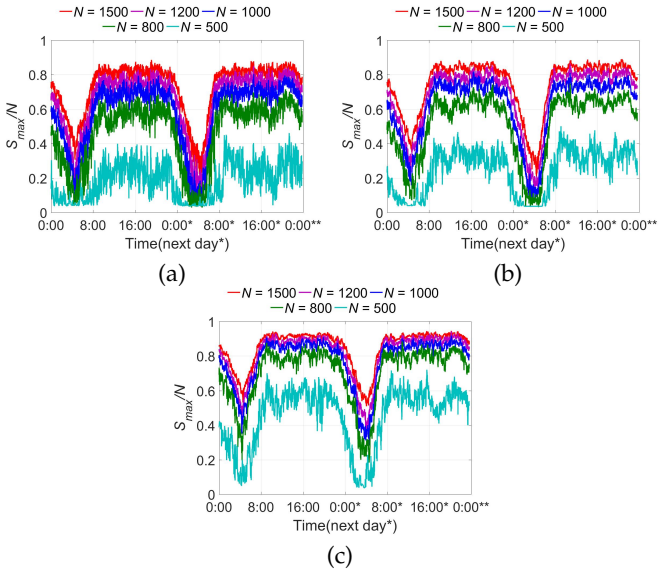


Fig. 22. Evolution of the ratio of the largest component size to the network size, $S_{\max}/N$, over 2 days for different N: (a) $\delta = 10$ min and $d = 200$ m, (b) $\delta = 20$ min and $d = 100$ m, and (c) $\delta = 20$ min and $d = 200$ m.

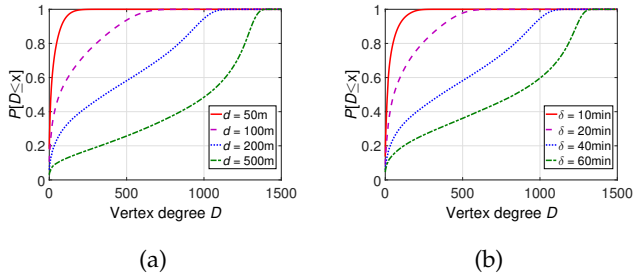


Fig. 23. CDFs of the vertex degree D aggregating all the samples over 2 days for: (a) different transmission distances d and $\delta = 40$ min, and (b) different delay tolerances δ and $d = 200$ m.

increases to near 1500, while for $\delta = 20$ min and $d = 200$ m, $S_{\max}/N$ saturates at around 90% as N increases to near 1500, which indicates that $N = 1500$ taxis are sufficiently dense in our traces. Further increasing N, while increasing the computational complexity dramatically owing to the complexity of the algorithm, will only make slight changes to our metrics.

In summary, the dynamical properties of the largest component are inextricably bounded up with the network parameters, specifically, the transmission distance d and the delay tolerance δ. In particular, the size of the largest component $S_{\max}$ is an increasing function of d and δ. Furthermore, the analysis on the influence of the network size N shows that $S_{\max}/N$ has an intuitive saturation value, indicating that $N = 1500$ taxis are a good choice to analyze the network connectivity, considering precision and timing budget.

4.4 Node-level Analysis

So far we focus on the properties of the whole network or the largest component. We now turn to the analysis on individual node level. Specifically, we aim to find the relationship between the vertex degree D of Definition 6 and the network parameters. In Fig. 23 (a), we plot the CDFs of D for different transmission distances d with $\delta = 40$ min.

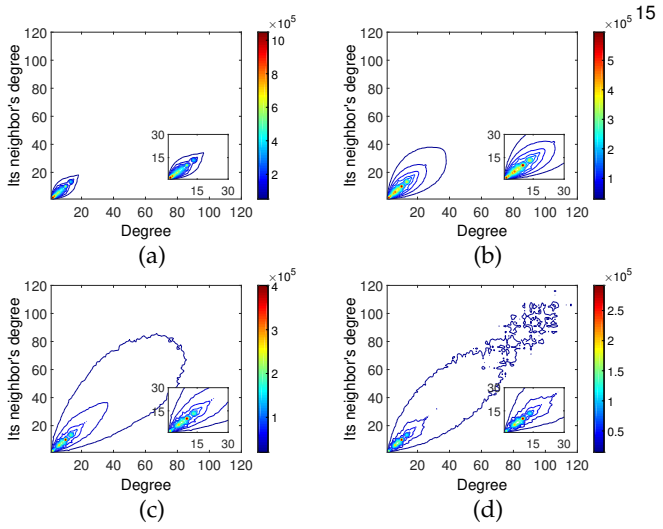


Fig. 24. Relationship expressed as isogram between node vertex degree D and the neighbor's vertex degree aggregating all the samples over 2 days under the network conditions of $d = 50$ m and different delay tolerances: (a) $\delta = 10$ min, (b) $\delta = 20$ min, (c) $\delta = 40$ min, and (d) $\delta = 60$ min.

The results show that the individual vertex degree increases rapidly with the increase of d. In particular, with $d = 100$ m, only 10% of the nodes have vertex degree over 500, and when the transmission distance is doubled to $d = 200$ m, the percentage rises to 40%, while when the transmission distance is increased to $d = 500$ m, the percentage increases to 80%. Similarly, Fig. 23 (b) shows the CDFs of the vertex degree D for different delay tolerances δ with $d = 200$ m. Observe that the individual vertex degree increases rapidly with the increase of δ. Specifically, there are litter nodes having vertex degree more than 500 with $\delta = 20$ min, while over half of the nodes can communicate with over 500 nodes in the network when the delay tolerance increases to $\delta = 40$ min.

Fig. 24 illustrates the relationship between the node vertex degree D and its neighbor's vertex degree under the network parameters of $d = 50$ m and different delay tolerances. In this figure, the numbers of pairs, i.e. (degree, neighbor's degree), are expressed as isogram and the details are portrayed in subplot. The high density area (with high appearance rate) is in red and typically lying on the diagonal. The results show the high associativity of the network nodes. To be more exact, the nodes with high vertex degree communicate with each other, while the low-degree ones communicate with other low-degree devices. As the delay tolerance increases from 10 minutes to 60 minutes, the isograms extend to high-degree area, which means that the whole network have more high-degree nodes whose neighbors are also high-degree nodes as the delay tolerance increases.

4.5 Summary and Implications

4.5.1 Results Summary

In this paper, we obtain the dynamical properties of taxi networks by analyzing ORG model, and our main findings are summarized as follows.

- We analyze the relationship between the component number C and the network parameters, specifically,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2878577, IEEE Transactions on Big Data

the transmission distance d and delay tolerance δ. Empirical results obtained from both Shanghai and Beijing traces indicate that C is an increasing function of d and δ, and the dynamical properties of C are very different during the daytime than at night. Moreover, it is found that the distribution of C can be accurately approximated by a uniform distribution, and we use the polynomial model to fit the two parameters of this uniform distribution. Our fitted uniform distribution is found to have high similarity with the empirical distribution of the real network.

- We analyze the geographical distribution of components. By categorizing the components into three classes of mini, medium and large components according to their sizes, we study the impact of the transmission distance d and delay tolerance δ on the distribution of components. The results indicate that when d and/or δ increase, first the numbers of medium components and large components will increase, and then the largest component begins to swallow up many smaller ones to become even bigger, which leads to the decrease of the total number of components.
- We then focus on the largest component, in which most multi-hop communications take place and ad-hoc network protocols are mainly designed to operate. The results obtained from all three traces show that the dynamical properties of the largest component is inextricably bound up with the network parameters, specifically, the transmission distance d, the delay tolerance δ and the number of the total nodes in the network N. In particular, the size of the largest component $S_{\max}$ is an increasing function of N, d and δ. Furthermore, we find that the ratio $S_{\max}/N$ has an intuitive saturation limit, indicating that $N = 1500$ taxis are a good choice to analyze the network connectivity, considering precision and timing budget.
- Finally, we focus on the individual node level, and analyze the distribution of vertex degree D to reveal the relationship between D and the network parameters d and δ. Isogram plots of the node vertex degree D and its neighbor's vertex degree further reveal the intriguingly high associativity of network nodes.

4.5.2 Implications

From the above analysis and summary, we observed taxi networking neighborhoods to be heterogeneous and assortative. Apart modeling the network connectivity and components for performance evaluations, these results and findings are highly relevant to the design of protocols for both vehicular-to-infrastructure and vehicular-to-vehicular communications. For example, our findings imply that one taxi is able to move quickly from an isolation component to being part of dense components with larger scales of connected taxis. Thus, in such highly time-varying dynamic environment, networking protocols and algorithms of power control, medium access control, data rate adaptation, etc. must be designed for rapid adaptability to the surrounding dynamic communication environment. Furthermore, we can use the observations with different system parameters of component number, largest component, node degree to design algorithms and protocols, and deploy them in different areas of the city according to the observed patterns of geographical distribution of components.

Another example is that from our results of low availability and limited reliability of large components when the tolerant delay is short, we obtain the evidence that it is difficult for routing or disseminating content within a purely ad-hoc taxi network. Then, intra-component connectivity, carry-and-forward transfers, and multi-hop routing protocols are needed. Our results under different tolerant delay and transmission time provide insights for the protocol and algorithm designs in these scenarios.

Overall, all the observations obtained in this paper reveal the spatio-temporal heterogeneity of vehicular connectivity revealed by the taxi networks. Therefore, these results stress that it is important to consider realistic and large-scale vehicular mobility datasets that comprise varied road traffic conditions in both the design and evaluation of vehicular networking protocols. They also suggest the dramatic importance of highly adaptive MAC and networking layer solutions to achieve effective vehicular-to-infrastructure and vehicular vehicle-to-vehicle communications by such opportunistic network connectivity.

5 Related Work

Recently, there have been continuous investigations to study vehicular mobility characteristics from various perspectives, and different mobility models have been proposed and studied. The study [19] considered three categories of mobility: macro-mobility, micro-mobility and bus network features. The work [20] also analyzed the mobility framework with macroscopic and microscopic metrics. The study [21] combined the stochastic model with traffic stream, car-following and flows-interaction to set up their simulation. Also there exist related works studying the mobility in different road conditions. For example, the work [22] focused on the mobility in intersections and two-dimensional road topology. Research [23] concentrated on the highway model and the work [24] was based on the lattice-shaped road network. However, our study is more generic since we consider the real urban cities – Shanghai, Beijing and Nanjing. The mobility model in our work is based on the real-world taxi traces, rather than a simulation trace. Most importantly we consider the networking problem with the two key network parameters, transmission distance and delay tolerance.

To study network topology, researchers have proposed various topology models based on different datasets and assumptions. For example, the studies [16], [25] investigated the instantaneous topology of a large-scale urban vehicular network. More specifically, Naboulsi and Fiore [16] studied the availability, connectivity and reliability of urban vehicular networks, based on a simulating vehicular dataset which may not reflect the true real vehicular behaviors in urban scenarios. The study [26] shares the same problem of [16] by only using a simulation trace, and the metric of node degree used in [26] is too simple to delineate the network topology. On the other hand, based on a real trace, the study [27] concentrated on the algorithms to improve the efficiency of

computing transitive closure of the networks. However, in their analysis, the authors of [27] only used 536 taxi nodes, which is far less than the number of nodes in our analysis. The Bologna dataset used in the study [28] is a good source of public dataset since the features of this dataset match well to those inferred from navigation services. However, there is no evidence that this dataset can be used to characterize the microscopic metrics, like dynamical connectivity in our study. Luo et al. [25] presented the characteristics of Shanghai trace, and used this real-world taxi dataset to discuss the connectivity and network performance, including link duration, average hops and connection rates. The connectivity in the study [25] is based on instantaneous metrics. The works [16], [22] also used an instantaneous model to characterize VANETs. Although the studies [20], [27] analyzed the temporal connectivity, it is actually the evolution of instantaneous topology. By contrast, our study aims to reveal the opportunistic topology in taxi networks, which is fundamentally different from the instantaneous topology. Moreover, our study employs three real-word large-scale taxi traces to calculate and to verify the proposed model and concepts and, therefore, our analysis reflects well the real taxi network behaviors in urban scenarios.

To compute the transitive closure of the time-varying graphs, Glacet et al. [13] also propose their way. However, they ignore the transmission distances in their graph, only assuming the link exists or not. In our model, we take the transmission distances into account, and research mainly on the effect of changing the transmission distance. Our algorithm can well support the computation of reachability graphs considering the transmission distance.

There exist many studies investigating the potential impacting factors on the connectivity, including topology, traffic signals, and vehicle traffic behaviors [19], [23], [29]. For example, Marfia et al. [29] focused on the stop-and-go behavior of traffic to study how it can cause network congestion and affect the connectivity. Artimy et al. [23] investigated the connectivity in VANETs and examined how the relative velocity as well as the number of lanes impact on the connectivity. Glacet et al. [13] research on the effect of store-carry-and-forward on the connectivity. Osman et al. [30] introduced the robustness model and analyze the effect of market penetration and traffic density on the robustness. Hou at el. [31] investigated the effect of component speed on the component size. Among various potential factors that influence the connectivity, the mobility is of greatest importance. Some studies [19], [32], [33], [34], [35], [36] did aim to explore the relationship between connectivity and mobility. However, all these works either study this relationship in general wireless networks [32], [33], [34], analyze the problem in a theoretical way [19], [35], [36], or analyzing some particular aspects of the network connectivity [13], [30], [31]. Thus, our study is the first work to reveal the fundamental properties and models of the opportunistic topology of taxi networks and to characterize the opportunistic connectivity in a large-scale urban mobility environment.

6 Conclusions And Future Work

We have characterized the opportunistic topology of taxi networks in real urban mobility environment in terms of connected components based on the ORG model. The opportunistic topology is largely different from the instantaneous topology without considering delay tolerance. Using real-world taxi traces in three big cities, i.e. Shanghai, Beijing and Nanjing, we have the in-depth analysis on how the opportunistic topology is affected under different network parameters, i.e. delay tolerance and transmission distance. The main metrics we use, i.e. the number, location and evolution of connected components and the size of the largest component, allow us to summarize the opportunistic topology in a high level on how nodes are temporally clustered and isolated.

Our study can be enhanced in several aspects. Firstly and obviously, we will revise the ORG model to better characterize the multi-hop process, for example, by introducing the concept of energy budget, etc. Secondly, from the dynamical connectivity property, we would like to obtain some other fundamental properties of the network, including throughput, that are capable of explaining critical performance of vehicular based communication networks. Finally, we will further consider the implementation of our theoretical models and analysis to the design, deployment and use of VANET's in real-world urban environments.

References

[1] G. Dimitrakopoulos and P. Demestichas, "Intelligent transportation systems," *IEEE Vehicular Technology Magazine*, vol. 5, no. 1, pp. 77–84, Mar. 2010.

[2] M. Khabazian, S. Aissa, and M. Mehmet-Ali, "Performance modeling of message dissemination in vehicular ad hoc networks with priority," *IEEE J. Selected Areas in Communications*, vol. 29, no. 1, pp. 61–71, Jan. 2011.

[3] J. Whitbeck, M. Dias de Amorim, V. Conan, and J.-L. Guillaume, "Temporal reachability graphs," in *Proc. MobiCom 2012* (Istanbul, Turkey), Aug. 22-26, 2012, pp. 377–388.

[4] A. Lindgren and P. Hui, "The quest for a killer app for opportunistic and delay tolerant networks," in *Proc. CHANTS 2009* (Beijing, China), Sep. 25, 2009, pp. 59–66.

[5] W. Zhao, M. Ammar, and E. Zegura, "Controlling the mobility of multiple data transport ferries in a delay-tolerant network," in *Proc. INFOCOM 2005* (Miami, FL), Mar. 13-17, 2005, vol. 2, pp. 1407–1418.

[6] C. Becker, M. Bauer, and J. Hähner, "Usenet-on-the-fly – supporting locality of information in spontaneous networking environments," in *Proc. CSCW 2002* (New Orleans, LA), Nov. 16-20, 2002, pp. 1–9.

[7] A. Casteigts, P. Flocchini, W. Quattrociocchi, and N. Santoro, "Time-varying graphs and dynamic networks," *Int. J. Parallel, Emergent and Distributed Systems*, vol. 27, no. 5, pp. 387–408, 2012.

[8] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Prentice Hall: Upper Saddle River, NJ, 1996.

[9] L. Cheng, B. E. Henty, D. D. Stancil, F. Bai, and P. Mudalige, "Mobile vehicle-to-vehicle narrow-band channel measurement and characterization of the 5.9 GHz dedicated short range communication (DSRC) frequency band," *IEEE J. Selected Areas in Communications*, vol. 25, no. 8, pp. 1501–1516, Oct. 2007.

[10] N. Akhtar, O. Ozkasap, and S. C. Ergen, "VANET topology characteristics under realistic mobility and channel models," in *Proc. WCNC 2013* (Shanghai, China), Apr. 7-10, 2013, pp. 1774–1779.

[11] M. Boban, T. T. V. Vinhoza, M. Ferreira, J. Barros, and O. K. Tonguz, "Impact of vehicles as obstacles in vehicular ad hoc networks," *IEEE J. Selected Areas Communications*, vol. 29, no. 1, pp. 15–28, Jan. 2011.

[12] C. Sommer, D. Eckhoff, R. German, and F. Dressler, "A computationally inexpensive empirical model of IEEE 802.11p radio shadowing in urban environments," in *Proc. WONS 2011* (Bardonecchia, Italy), Jan. 26-28, 2011, pp. 84–90.

[13] C. Glacet, M. Fiore, and M. Gramaglia, "Temporal connectivity of vehicular networks: the power of store-carry-and-forward," in *Proc. VNC 2015* (Kyoto, Japan), Dec. 16-18, 2015, pp. 52–59.

[14] H. Zhu, M. Li, L. Fu, G. Xue, Y. Zhu, and L. M. Ni, "Impact of traffic influxes: Revealing exponential intercontact time in urban VANETs," *IEEE Trans. Parallel and Distributed Systems*, vol. 22, no. 8, pp. 1258–1266, Aug. 2011.

[15] M. Li, H. Zhu, Y. Zhu, and L. M. Ni, "ANTS: Efficient vehicle locating based on ant search in ShanghaiGrid," *IEEE Trans. Vehicular Technology*, vol. 58, no. 8, pp. 4088–4097, Oct. 2009.

[16] D. Naboulsi and M. Fiore, "On the instantaneous topology of a large-scale urban vehicular network: the Cologne case," in *Proc. ACM MobiHoc 2013* (Bangalore, India), Jul. 29-Aug. 1, 2013, pp. 167–176.

[17] D. Hadaller, S. Keshav, T. Brecht, and S. Agarwal, "Vehicular opportunistic communication under the microscope," in *Proc. MobiSys 2007* (San Juan, Puerto Rico), Jun. 11-14, 2007, pp. 206–219.

[18] F. Bai, D. D. Stancil, and H. Krishnan, "Toward understanding characteristics of dedicated short range communications (DSRC) from a perspective of vehicular network engineers," in *Proc. MobiCom 2010* (Chicago, IL), Sept. 20-24, 2010, pp. 329–340.

[19] I. W. H. Ho and K. K. Leung, "Node connectivity in vehicular ad hoc networks with structured mobility," in *Proc. 32nd IEEE Conf. Local Computer Networks* (Dublin, Ireland), Oct. 15-18, 2007, pp. 635–642.

[20] F. D. Cunha, A. C. Vianna, R. A. F. Mini, and A. A. F. Loureiro, "Is it possible to find social properties in vehicular networks?" in *Proc. ISCC 2014*, (Madeira, Portugal), Jun. 23-26, 2014, pp. 1–6.

[21] M. Fiore, and J. Härri, "The networking shape of vehicular mobility," in *Proc. ACM MobiHoc 2008* (Hong Kong, China), May 26-30, 2008, pp. 261–272.

[22] W. Viriyasitavat, O. K. Tonguz, and F. Bai, "Network connectivity of VANETs in urban areas," in *Proc. IEEE SECON 2009* (Rome, Italy), Jun. 22-26, 2009, pp. 1–9.

[23] M. M. Artimy, W. Robertson, and W. J. Phillips, "Connectivity in inter-vehicle ad hoc networks," in *Proc. Canadian Conf. Electrical and Computer Eng. 2004* (Ontario, Canada), May 2-5, 2004, vol. 1, pp. 293–298.

[24] S. Shioda, J. Harada, Y. Watanabe, T. Goi, H. Okada, and K. Mase, "Fundamental characteristics of connectivity in vehicular ad hoc networks," in *Proc. IEEE PIMRC 2008* (Cannes, France), Sep. 15-18, 2008, pp. 1–6.

[25] P. E. Luo, H. Y. Huang, and M. L. Li, "Characteristics of trace data for a large scale ad hoc network - Shanghai Urban Vehicular Network," in *Proc. IET Conf. Wireless, Mobile and Sensor Networks 2007* (Shanghai, China), pp. 742-745, Dec. 12-14, 2007.

[26] H. Conceicao, M. Ferreira, and J. Barros, "On the urban connectivity of vehicular sensor networks," in *Proc. DCOSS 2008* (Santorini Island, Greece), Jun. 11-14. 2008, pp. 112–125.

[27] M. A. Hoque, X. Hong, and B. Dixon, "Efficient multi-hop connectivity analysis in urban vehicular networks," *Vehicular Communications*, vol. 1, no. 2, pp. 78–90, Apr. 2014.

[28] L. Bedogni, M. Gramaglia, A. Vesco, M. Fiore, J. Härri, and F. errero, "The Bologna ringway dataset: improving road network conversion in SUMO and validating urban mobility via navigation services," *IEEE Trans. Vehicular Technology*, vol. 64, no. 12, pp. 5464–5476, Dec. 2015.

[29] G. Marfia, G. Pau, E. De Sena, E. Giordano, and M. Gerla, "Evaluating vehicle network strategies for downtown Portland: opportunistic infrastructure and the importance of realistic mobility models," in *Proc. MobiSys 2007* (San Juan, Puerto Rico), Jun. 11-14, 2007, pp. 47–51.

[30] O. A. Osama and S. Ishak, "A network level connectivity robustness measure for connected vehicle environments," *Transportation Research Part C: Emerging Technologies*, vol. 53, pp. 48–58, Apr. 2015.

[31] X. Hou, Y. Li, D. Jin, D. O. Wu, and S. Chen, "Modeling the impact of mobility on the connectivity of vehicular networks in large-scale urban environments," *IEEE Trans. Vehicular Technology*, vol. 65, no. 4, pp. 2753–2758, Apr. 2016.

[32] T. K. Madsen, F. H. P. Fitzek, and R. Prasad, "Impact of different mobility models on connectivity probability of a wireless ad hoc network," in *Proc. 2004 Int. Workshop Wireless Ad-Hoc Networks* (Oulu, Finland), May 31-Jun. 3, 2004, pp. 120–124.

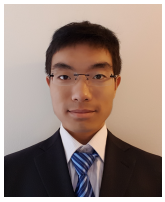
[33] F. Bai, N. Sadagopan, and A. Helmy, "IMPORTANT: a framework to systematically analyze the Impact of Mobility on Performance of Routing Protocols for Adhoc Networks," in *Proc. INFOCOM 2003* (San Francisco, CA), Mar. 30-Apr. 3, 2003, vol. 2, pp. 825–835.

[34] Q. Wang, X. Wang, and X. Lin, "Mobility increases the connectivity of K-hop clustered wireless networks," in *Proc. ACM MobiCom 2009* (Beijing, China), Sep. 20-25, 2009, pp. 121–132.

[35] X. Zhang, J. Kurose, B. N. Levine, D. Towsley, and H. Zhang, "Study of a bus-based disruption-tolerant network: mobility modeling and impact on routing," in *Proc. ACM MobiCom 2007* (Montreal, Canada), Sep. 9-14, 2007, pp. 195–206.

[36] H. Füβler, M. Torrent-Moreno, M. Transier, R. Küger, H. Hartenstein, and W. Effelsberg, "Studying vehicle movements on highways and their impact on ad-hoc connectivity," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 10, no. 4, pp. 26–27, Oct. 2006.

[37] P. Newson and J. Krumm, "Study of a bus-based disruption-tolerant network: mobility modeling and impact on routing," in *Proc. 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Seattle, Washington), Nov. 4-6, 2009, pp. 336–343 .



Ran Xu received the B.E. degree in electronic engineering from Tsinghua University. Now, he is a PhD student in the Department of Electrical and Computer Engineering at Purdue University. His research interests include vehicular networks, mobile computing, approximate computing and distributed storage system.

Yong Li (M'09-SM'16) received the B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007 and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. He is currently a Faculty Member of the Department of Electronic Engineering, Tsinghua University.

Dr. Li has served as General Chair, TPC Chair, TPC Member for several international workshops and conferences, and he is on the editorial board of three international journals. His papers have total citations more than 2300 (six papers exceed 100 citations, Google Scholar). Among them, eight are ESI Highly Cited Papers in Computer Science, and four receive conference Best Paper (run-up) Awards. He received IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers and Young Talent Program of China Association for Science and Technology.

Sheng Chen (M'1990-SM'1997-F'2008) obtained his BEng degree from the East China Petroleum Institute, Dongying, China, in January 1982, and his PhD degree from the City University, London, in September 1986, both in control engineering. In 2005, he was awarded the higher doctorate degree, Doctor of Sciences (DSc), from the University of Southampton, Southampton, UK. From 1986 to 1999, He held research and academic appointments at the Universities of Sheffield, Edinburgh and Portsmouth, all in UK. Since 1999, he has been with Electronics and Computer Science, the University of Southampton, UK, where he currently holds the post of Professor inIntelligent Systems and Signal Processing. Dr Chen is a Distinguished Adjunct Professor at King Abdulaziz University, Jeddah, Saudi Arabia. He is a Chartered Engineer (CEng) and a Fellow of IET (FIET). His recent research interests include adaptive signal processing, wireless communications, modeling and identification of nonlinear systems, neural network and machine learning, intelligent control system design, evolutionary computation methods and optimization. He has published over 470 research papers. Dr Chen is an ISI highly cited researcher in the engineering category (March 2004).