Discovery of Critical Nodes in Road Networks Through Mining From Vehicle Trajectories

Ming Xu, Jianping Wu, Mengqi Liu, Yunpeng Xiao, Haohan Wang, and Dongmei Hu

***Abstract*— Road networks are extremely vulnerable to cascading failure caused by traffic accidents or anomalous events. Therefore, accurate identification of critical nodes, whose failure may cause a dramatic reduction in the road network transmission efficiency, is of great significance to traffic management and control schemes. However, none of the existing approaches can locate city-wide critical nodes in real road networks. In this paper, we propose a novel data-driven framework to rank node importance through mining from comprehensive vehicle trajectory data, instead of analysis solely on the topology of the road network. In this framework, we introduce a trip network modeled by a tripartite graph to characterize the dynamics of the road network. Furthermore, we present two algorithms, integrating the origin-destination entropy with flow (ODEF) algorithm and the crossroad-rank (CRRank) algorithm, to better exploit the information included in the tripartite graph and to effectively assess the node importance. ODEF absorbs the idea of the information entropy to evaluate the centrality of a node and to calculate its importance rating by integrating its centrality with the traffic flow. CRRank is a ranking algorithm based on eigenvector centrality that captures the mutual reinforcing relationships among the OD-pair, path, and intersection. In addition to the factors considered in ODEF, CRRank considers the irreplaceability of a node and the spatial relationships between neighboring nodes. We conduct a synthetic experiment and a real case study based on a real-world dataset of taxi trajectories. Experiments verify the utility of the proposed algorithms.**

***Index Terms*— Tripartite graph, ranking algorithm, road network, vehicle trajectories, OD entropy.**

I. Introduction

ROAD networks are the critical infrastructure of the urban transportation system and can be regarded as one of the largest and most complex systems of modern society. With the rapidly accelerating demand for mobility, road networks have suffered from increasing pressure. An anomalous event, such as a traffic accident, may readily cause severe traffic congestions for some road segments (or intersections). In recent years, there have been many studies revealing a phenomenon of cascading failures in road networks [1]–[3]. This indicated that when a few critical nodes fail, they would cause other nodes to fail through connections between the nodes, which would dramatically reduce the transmission efficiency of the network and even result in the global collapse. Therefore, accurately identifying the critical nodes and prioritizing protective strategies is an effective approach for congestion control and to improve the transportation efficiency.

In previous studies, the road network is generally abstracted as a connected graph, in which nodes represent intersections and links represent road segments that connect two neighboring intersections. In this way, a family of methodologies based on network science and graph theory can be used to analyze the road network infrastructure. However, effectively and efficiently discovering the critical nodes in the road network is still a significant challenge due to the following reasons.

1) Road networks have a homogeneous degree distribution. This means that the majority of nodes in the road network have similar degrees, which leads to the ineffectiveness of a class of methods based on the local metrics, such as degree centrality.

2) Road networks are very large scale and include tens of thousands of nodes and edges. For such a scale network, the simulation or the global metrics based methods, such as the betweeness centrality and closeness centrality, cannot be applied due to high cost, in spite of their acceptable effectiveness on smaller networks.

3) The existing methods, regardless of the local metrics or the global metrics, evaluate the node importance based only on the topology of the network. However, significant useful information, such as the quality of the road segments, delay at the signalized intersections and the surroundings of the roads, is very difficult to collect, and is therefore generally neglected when modeling the road network as a node-edge network. This information is related to the path selection that directly affects the ranking of the node importance.

4) With the development of cities and societies, mobility and travel demands, such as traffic flow between different regions, travel distance, mode of transport, etc., have changed greatly. The changeable flow distribution and unchangeable infrastruc-

Manuscript received March 10, 2017; revised July 27, 2017 and January 26, 2018; accepted March 14, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61363001 and in part by the Henan Provincial Science and Technology Project of China under Grant 162102210214. This paper was presented at the 3rd SigKDD Workshop on Urban Computing (UrbComp 2014) [25]. The Associate Editor for this paper was Z. Ding. *(Corresponding author: Jianping Wu.)*

M. Xu, J. Wu, and D. Hu are with the Department of Civil Engineering, Tsinghua University, Beijing 100084, China (e-mail: mxu@tsinghua.edu.cn; jianpingwu@tsinghua.edu.cn; hudm13@tsinghua.edu.cn).

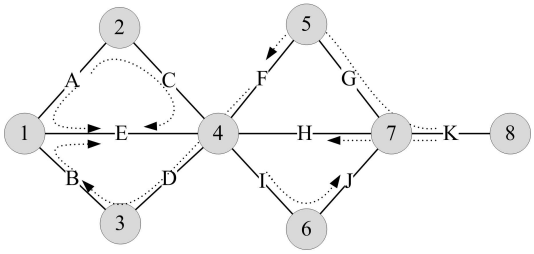
M. Liu is with the Department of Telecommunication Engineering with Management, Beijing University of Posts and Telecommunications, Beijing 100084, China (e-mail: victorialiu0628@gmail.com).

Y. Xiao is with the School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: xiaoyp@cqupt.edu.cn).

H. Wang is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: haohanw@andrew.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2018.2817282



(a)

	p_1	p_2	p_3	p_4	p_5	p_6	p_7
$A \rightarrow E$	10	10	0	0	0	0	0
$B \rightarrow E$	0	0	10	0	0	0	0
$F \rightarrow B$	0	0	0	5	0	0	0
$K \rightarrow F$	0	0	0	0	10	0	0
$K \rightarrow H$	0	0	0	0	0	10	0
$I \rightarrow J$	0	0	0	0	0	0	25

(b)

	1	2	3	4	5	6	7	8
p_1	1	0	0	0	0	0	0	0
p_2	0	1	0	1	0	0	0	0
p_3	1	0	0	0	0	0	0	0
p_4	0	0	1	1	0	0	0	0
p_5	0	0	0	0	1	0	1	0
p_6	0	0	0	0	0	0	1	0
p_7	0	0	0	0	0	1	0	0

(c)

Fig. 1. An example network consisted of 8 nodes and 11 edges. (a) the network topology. (b) OD-Path matrix. (c) Path-Intersection matrix.

ture of the road network may cause extreme unbalance of the flow distribution, which may transform the role of hubs defined in road network planning.

A simplified example is presented in Fig. 1. Fig. 1(a) gives the topology of a small network including 8 nodes (1-8) and 11 edges (A-K). The objective is to rank the node importance. In cases where only the network topology is considered, it can be intuitively judged that node 4 may be the most important due to its location in the network. In fact, it has the highest betweenness centrality. Now let us consider the flow distribution denoted by the dashed directed lines in (a); the traffic flow on each path of each origin-destination (OD) is given by an OD-path matrix shown in (b); whether or not a path traversing a node is indicated by a path-intersection matrix shown in (c). Specifically, node 6 carries the largest flow from one OD-pair; both node 1 and node 7 have the second biggest mixed flow from two distinct OD-pairs. Taking this into consideration, can we confidently clarify which node has the greatest impact on the network if it fails? Moreover, what about a larger scale network? This paper will answer these questions.

Although there are currently no exact answers to these issues, it is clear that the existing methods based on topology analysis or simulations cannot perform well. The information neglected by the existing methods, such as the flow distribution and route selection, is likely to be critical, perhaps more critical than the topology information, since such information contains not only the dynamics of the road network, but also rich experiential knowledge of the drivers. This experiential knowledge indirectly captures unquantifiable structural features of the road network, which are difficult to collect directly.

In recent years, massive trajectory data are widely available, which motivates us to tackle this challenge from a data-driven perspective. In this paper, we introduce a novel trip network to characterize the dynamics of the road network by extracting information from the comprehensive vehicle trajectory data, rather than simply represent the road network as a node-edge network. Specifically, we model the trip network using a tripartite graph, which reflects the traffic flow between different regions and route preferences of drivers. Based on the trip network, we propose two algorithms, integrating the OD Entropy with Flows (ODEF) algorithm and the Crossroad-Rank (CRRank) algorithm. ODEF presents a new global centrality measure inspired by information entropy. Moreover, it strikes a proper balance between the traffic volume and the centrality of a node to assess its significance, which rectifies the shortcomings of the methods based only on the topology analysis. Like the PageRank [11] and HITS algorithm [10], CRRank is also a variant of the eigenvector centrality measure. It implicates a mutually reinforcing concept: a heavy-load trip is likely to choose popular paths and a popular path is likely to be chosen by a heavy-load trip. Similarly, a popular path is likely to traverse many important intersections and an important intersection is likely to be traversed by many popular paths. In addition to having the advantages of ODEF, CRRank considers the irreplaceability of the path and spatial relationships between neighboring nodes to rank the node importance. Specifically, CRRank slightly enhances the significance of the nodes on the irreplaceable paths, such as a bridge connecting two regions. Since the congestion of the downstream node has the potential to cause the upstream node to be congested, the downstream nodes of big flows may obtain a relatively high ranking. We validate the effectiveness of the two proposed algorithms using a synthetic experiment and a real trajectory dataset of taxicabs, respectively. The results indicate that both the ODEF and CCRank can identify city-wide important nodes in the global connectivity perspective with a low computing cost. In particular, CRRank shows an improved performance over ODEF.

The remainder of the paper is organized as follows. In Section II, related studies on the discovery of important nodes are presented. In Section III, we briefly discuss the original intention behind our model. Then, the proposed framework and related definitions are introduced. In Section IV, related concepts of the tripartite graph are given, and the utilization of the tripartite graph in the trip network modeling is described. Furthermore, two algorithms are presented. In Section V, the experiments on the validity of the two proposed algorithms are presented. Section VI concludes the paper with a summary.

II. Related Work

In this section, we briefly review two research fields that are relevant to this paper.

A. Evaluation of Node Importance

In complex networks, due to the cascading failure phenomena resulting from the failure of a small number of important nodes, identifying important nodes is of vital significance to prevent collapse and to maintain the functionality of the network. The methods in this research area mainly exploit the

network topology and can be divided into four categories: node importance ranking based on 1) the local centrality measure, 2) global centrality measure, 3) eigenvector centrality, and 4) node removing. The referred evaluation metrics of the first category are the degree centrality and the semi-local centrality [4]. The idea of the degree centrality method is that the more direct neighbors a node has, the more important it is. The semi-local centrality considers the number of both direct and indirect neighbors of a node in a limited area. Recently, more works have focused on improving such methods. The study [5] considered the mutual dependence among nodes. In [6], a method was proposed based on both the degree value and the importance of edges. The advantages of this category of methods are simplicity and intuition, but the results are not as good as other approaches. The works in the second category sought to evaluate the importance of a node from a global network perspective. The representative evaluation metric is the betweenness centrality of a node, which is the number of the shortest paths that pass through this node for every pair of nodes. In addition to the betweenness centrality, the global centrality measure involves a range of metrics, such as the closeness centrality, Kats centrality, flow betweenness centrality [7], communicability betweenness centrality [8], routing betweenness centrality [9], etc. Compared with the local centrality, the global centrality can obtain better results. However, it is difficult to apply the global centrality over a large scale network due to their inherent high complexity. The most relevant methods to our work are the third category of methods, which consider not only the number of neighboring nodes but also their relative importance. This category of quintessential methods includes HITs [10], PageRank [11], LeaderRank [12], and SALSA [13]. However, due to the homogeneous degree distribution of the road network, these methods cannot be directly applied. The last category of methods evaluates the node importance by following the two steps: 1) remove the node from the network, 2) measure the decline rate of the network efficiency [14]. This category also faces the high computing cost, and cannot be extended to large networks.

A common problem in all the aforementioned methods is that they only consider the network topology, neglecting some critical structural features that are difficult to collect and quantify. To the best of our knowledge, this paper is the first research that discovers important nodes in road networks using a data-driven framework and rectifies the aforementioned shortcomings.

B. Data-Driven Road Network Analysis

The flourishing of sensing technologies and large scale computing infrastructures have produced a variety of big and heterogeneous data for urban spaces, e.g., human mobility, vehicle trajectories, air quality, traffic patterns, and geographical data [15]. These data inspire a variety of novel data mining methods for road network analysis. A LDA-based inference model that combined road network data, points of interests, and massive taxi trajectories to infer the functional regions in a city was proposed in [16]. The study [17] discovered the bottleneck of road networks using the corresponding taxi trajectories. In [18], a PCA-based algorithm was presented to identify the link anomalies, and then an optimization technique was used to infer the traffic flow that produced the traffic anomalies. The study [19] used frequent sub-graph mining to discover anomalous links at each time interval, and then outlier trees were formed to reveal the potential flaws in the design of the existing traffic networks. The detection of traffic anomalies in road networks using crowd sensing with human mobility and social media data was given in [20]. The study [21] proposed a spatio-temporal graph mining method to discover black holes (traffic sink) and volcanoes (traffic source) using human mobility data in a city, which can be used to quickly detect the traffic anomalies, such as concerts and football matches. Similarly, this paper also addresses the problem of transportation planning using massive vehicle trajectories.

III. Framework

In a road network, a node is critical if its failure could lead to a dramatic reduction in the transmission efficiency of the entire network as well as the failure of numerous other nodes. Intuitively, when a node fails, its load will be redistributed to its neighbor nodes, and then the neighboring nodes may be overloaded and fail in turn. Therefore, the failure of a higher capacity node is likely to have a greater impact on its neighboring nodes; the failure of a higher centrality node would also impact a wider range of nodes. Consequently, it can be inferred that the importance of a node can be determined by its capacity and centrality, which has already been verified in the previous studies [22], [23]. The capacity indicator of a node can be measured approximately by calculating its traffic volume, whereas the centrality indicator has multiple different definitions and measures as illustrated in the related work section. As a path-level centrality measure, the betweenness centrality could find only structural vulnerabilities instead of true vulnerabilities, since it employs pure topological information following a hypothesis that traffic flow is assigned according to the shortest paths. This is unrealistic and fails to capture the complex structural characteristic of road networks. In many cases, the shortest path is not the drivers' preference. The factors that affect a driver's route selection are extremely complicated and uncertain, and these factors may include the number of turns, the road environment, traffic flow, road quality, etc. For example, consider an intersection near a famous tourist attraction, the Forbidden City in Beijing, which is located in an urban center area. We find that it is traversed by the shortest paths for many region pairs, and thus, it has a high betweenness centrality. However, in reality, due to the traffic chaos caused by many tourists visiting this site, drivers usually prefer detour paths rather than the shortest path passing through this intersection. As a result, this node cannot act as a hub. Although route preferences and flow characteristics are extremely crucial, it has not previously been possible for these factors to be fully considered in the flow assignment model. Fortunately, these unquantifiable factors are reflected in the vehicle trajectories. Therefore, we introduce a data driven framework based on the comprehensive vehicle trajectories

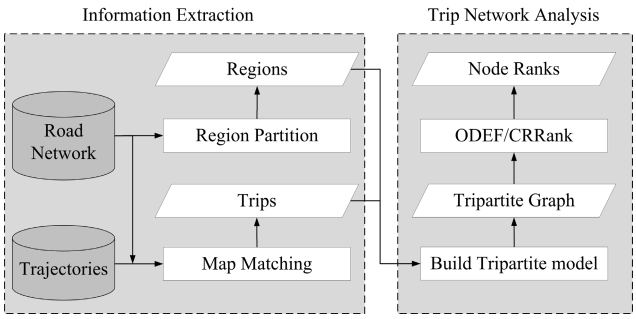


Fig. 2. The framework of the proposed approach to rank node importance.

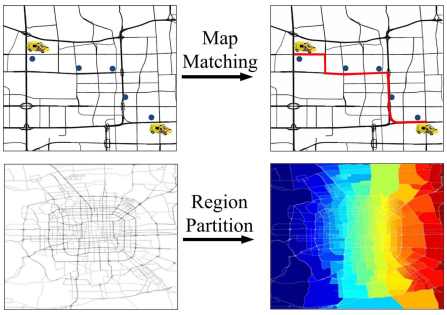


Fig. 3. Illustrations of the map matching and region partitioning.

and geographic information of the road network to identify the significant nodes. Before describing our approach, we first introduce some concepts used throughout this paper.

Definition 1 (Road Network): A road network is a directed weighted graph $G(V, E)$, where V is the set of nodes representing intersections or end points of road segments; E is the set of weighted edges representing the road segments that link two intersections. Each node v_i is associated with an id $v_i.id$; each edge e_i is associated with an id $e_i.id$ and a level indicator $e_i.level$. In detail, if e_i is an expressway, freeway or arterial road, $e_i.level$ equals 1; if e_i is a sub-arterial road, $e_i.level$ equals 2; if e_i is a bypass, $e_i.level$ equals 3.

Definition 2 (Region): A city map can be partitioned into non-overlapping regions by the high-level road segments ($e.level = 1$) in the road network. The set R of these regions is represented by $\{r_1, r_2, \cdots, r_{|R|}\}$, where $|R|$ is the total number of regions; each region r is associated with an id $r.id$. Some low-level road segments can also be included within a region. The partition method can preserve the semantic meaning of a region, e.g., schools, parks, business areas, residential areas, etc.

Definition 3 (OD-Pair): An OD-pair s is represented by a triple $< o, d, \rho_s >$, where $o \in R$ represents the origin region; $d \in R$ represents the destination region; $\rho_s : \{p_1, p_2, \cdots, p_{|\rho_s|}\}$ is the set of candidate paths for s. s is valid if it has a flow.

Definition 4 (OD-Flow): An OD-flow f is a sequence of vehicles from a particular origin to a particular destination. f is represented by a tuple $< s, m >$, where s is an OD-pair; m represents the size of f (i.e. the number of the vehicles).

Definition 5 (Path): A path p is an intersection sequence including all the intersections traversed by a trip for an OD-pair, $p : \{v_1, v_2, \cdots, v_{|p|}\}$, where v_i is the i^{th} intersection in p; $|p|$ is the number of intersections in p.

Our data-driven framework can be divided into the following two parts: information extraction and trip network analysis, as depicted in Fig. 2.

A. Information Extraction

In this phase, we extract the raw trip information of each vehicle from the GPS trajectories that refer to their geospatial coordinate readings (longitude and latitude) with their sampling timestamps. However, due to the noise in the GPS samples, the readings often deviate from the actual road segments. To solve this problem, a map matching algorithm [15] is used to map each GPS reading onto the corresponding road segment, as presented in the upper part of Fig.3. Each record in a raw trip includes three items: the trip id, road segment id and the sampling timestamp. In addition, the area segmentation algorithm [16] is used to partition the urban map into regions, as depicted in the bottom of Fig. 3.

B. Trip Network Analysis

In this phase, as depicted in the right hand side of Fig. 2, we use a tripartite graph to represent the trip network, which is described in detail in the next section. To construct the tripartite graph, we need to collect the OD-pairs and paths. All the valid OD-pairs can be obtained according to which region the origin road and the destination road of each raw trip are falling in. All of the paths can be obtained via transforming the road segment sequence of each raw trip into the intersection sequence based on the topology of the road network. Then, ODEF and CRRank algorithms are implemented to rank the node importance.

IV. Methodology

A. Tripartite Graph of Trip Network

Differing from the road network reflecting the geometric structure, the trip network characterizes the dynamics of the road network, such as the traffic flows from different OD-pairs and path selections. In this paper, the trip network is represented as a tripartite graph. A tripartite graph is a graph whose nodes can be partitioned into three disjoint sets so that no two nodes within the same set are directly connected. The relationships encoded by the edges between two sets of nodes can be summarized into a pair of adjacency matrixes. Specifically, the tripartite graph of the trip network is given in Fig. 4, and can be represented formally by $G'(S \bigcup P \bigcup V, W \bigcup U)$, where S, P and V represent the sets of the nodes corresponding to valid OD-pairs, paths, and intersections, respectively; $|S|$, $|P|$ and $|V|$ are the number of nodes in these three categories, respectively; W denotes the adjacency matrix corresponding to the OD-pair-path edges; the entry $w_{s,p}$ of W is the size of flow on path p from OD-pair s; U denotes the adjacency matrix corresponding to the path-intersection edge. If p contains the intersection v, $u_{p,v}$ equals 1; otherwise, $u_{p,v}$ equals 0. Next, we describe the two algorithms for the node importance ranking based on the tripartite graph.

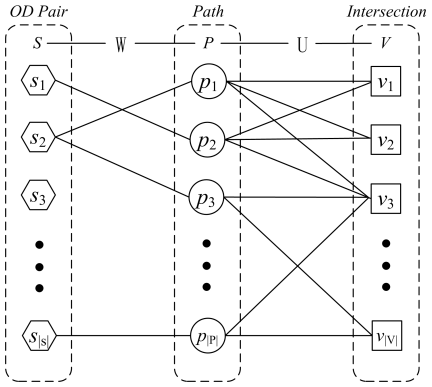


Fig. 4. The tripartite graph that represents the trip network.

B. The ODEF Algorithm

To accurately evaluate the centrality of the nodes and locate the traffic hub, we introduce the concept of OD entropy. And then we integrate the OD entropy with the traffic flow to calculate importance ratings of each node. In information theory, entropy is a measure of the unpredictability or the amount of information from a message source. As a powerful tool, entropy-based methods have also expanded to various fields, such as mathematical chemistry, cybernetics, computational physics and pattern recognition. In particular, graph entropy is widely applied to quantify structural information of graphs in graph theory [27], [28]. Motivated by these concepts, the OD entropy of a node in the road network is given by

$$H(v_k) = -\sum_{s=1}^{|S|} p_s(v_k) \log p_s(v_k) \tag{1}$$

where $p_s(v_k)$ represents the probability that the flow on node v_k is from OD-pair s. Intuitively, the higher the OD entropy of a node, the poorer the predictability of a specific OD-pair passing through it, and the wider areas this node connects. Hence, the OD entropy is very appropriate to measure the centrality of a node, reflecting the route selection of the drivers. Next, the importance of node v_k is calculated by integrating the OD entropy with the traffic flow and given by

$$I(v_k) = \sum_{s=1}^{|S|} f_k p_s(v_k) \left[1 + \log p_s^{-1}(v_k)\right]^{\frac{1}{5}} \tag{2}$$

where f_k is the sum of all the flows through the intersection v_k. Both f_k and $p_s(v_k)$ can be calculated according to the tripartite graph and are given by

$$f_k = \sum_{s=1}^{|S|} \sum_{i=1}^{|P|} w_{s,i} u_{i,k} \tag{3}$$

$$p_s(v_k) = \frac{\sum_{i=1}^{|p|} w_{s,i} u_{i,k}}{\sum_{j=1}^{|S|} \sum_{i=1}^{|P|} w_{j,i} u_{i,k}} \tag{4}$$

C. The CRRank Algorithm

Mutually reinforcing relationships among linked objects in the graphs are ubiquitous and considered as the basis of many ranking algorithms, e.g., in the HITS algorithm, each webpage has both a hub score and an authority score. The intuition is that a good authority is pointed to by many good hubs and a good hub points to many good authorities. The hub and authority of each webpage will converge to reasonable values through an iterative calculation, which is similar to the idea of CRRank.

To implement the score propagation over the tripartite graph, each node is assigned a meaningful score that is updated during each iteration. We use the vector L to denote the OD-load, which represents the scores of each OD-pair. In fact, the road network is a finite transportation resource. Once a trip of any OD-pair is generated, the load on the road network is increased, while the load caused by different OD-pairs is different. Similarly, vector H represents the popularity of each path, and vector C represents the importance of each intersection. To integrate more priori knowledge about the traffic flow distribution and topology of the road network into CRRank, we define three profile vectors $L^{(0)}$, $H^{(0)}$ and $C^{(0)}$, which are the initial score vectors of the OD-load, path popularity and intersection importance. These vectors are given by

$$L^{(0)} = \left[\frac{\sum_j w_{1,j}}{\sum_i \sum_j w_{i,j}} \quad \frac{\sum_j w_{2,j}}{\sum_i \sum_j w_{i,j}} \quad \cdots \quad \frac{\sum_j w_{|S|,j}}{\sum_i \sum_j w_{i,j}} \right]^T \tag{5}$$

$$H^{(0)} = \left[\frac{\sum_i w_{i,1}}{\sum_j \sum_i w_{i,j}} \quad \frac{\sum_i w_{i,2}}{\sum_j \sum_i w_{i,j}} \quad \cdots \quad \frac{\sum_i w_{i,|P|}}{\sum_j \sum_i w_{i,j}} \right]^T \tag{6}$$

$$C^{(0)} = \left[\frac{\sum_{d_1} \gamma_1^{d_1}}{\sum_{d_v} \sum_{d_v} \gamma_v^{d_v}} \quad \frac{\sum_{d_2} \gamma_2^{d_2}}{\sum_{d_v} \sum_{d_v} \gamma_v^{d_v}} \quad \cdots \quad \frac{\sum_{d_{|V|}} \gamma_{|V|}^{d_{|V|}}}{\sum_{d_v} \sum_{d_v} \gamma_v^{d_v}} \right]^T \tag{7}$$

where $\gamma_v^{d_v}$ is level score of the d^{th} edge of the intersection v and is defined as

$$\gamma_v^{d_v} = \begin{cases} 1+\lambda & \text{if } e.level = 1 \\ 1 & \text{if } e.level = 2 \\ 1-\lambda & \text{if } e.level = 3 \end{cases} \tag{8}$$

Here, the value of λ is set to 0.1. These definitions are accordance with our intuition. That is, the OD pair with a larger flow perhaps becomes a heavier load on the road network; therefore, it has a higher initial load score. In a similar way, the path with a larger traffic flow should have a higher initial popularity score, and the intersection linking higher level roads should have a higher initial importance score. The final scores for all node types can be synchronously calculated with the weight matrixes via an iterative mode. The weighted matrixes between S and P, denoted by $X^{S\to P}$ and $X^{P\to S}$ respectively, can be obtained by transforming from the adjacency matrix using column normalization and are given by

$$X_{s,p}^{S\to P} = \frac{w_{s,p}}{\sum_{i=1}^{|P|} w_{s,p}} \tag{9}$$

$$X_{p,s}^{P\to S} = \left(\left[X^{S\to P} \right]\right)_{p,s}^T \tag{10}$$

The weighted matrixes $Y^{P\to V}$ and $Y^{V\to P}$ between P and V are calculated by

$$Y^{P\to V}_{p,v} = (u_{p,v} + \delta_{p,v})c^{(0)}_v \tag{11}$$

$$Y^{V\to P}_{v,p} = u_{v,p}h^{(0)}_p \tag{12}$$

where $c^{(0)}_v$ is the v^{th} entry of the vector $C^{(0)}$; $h^{(0)}_p$ is the p^{th} entry of the vector $H^{(0)}$; $\delta_{p,v}$ is used to model the impact on the upstream intersection from the downstream intersection. In principle, if the downstream intersection is congested with a large flow, the congestion will spread rapidly from the downstream to the upstream. Therefore, the direction of the flow is very useful and also considered in CRRank. To calculate this, a bi-directional index structure is built. In the forward index, each distinct path is indexed by all the intersections that it contains, and we define the function $J(p, i)$ that returns id of the i^{th} intersection in the path p; In the reverse index, each intersection is indexed by every path that traverses it, and the function $R(p, v)$ is used to return the index of the intersection v in path p. If v is not included in p, the number "0" is returned. Commonly, the impact between two neighboring intersections disappears with the increase in the distance between them, so it is reasonable to define the function δ as being dependent on the distance. However, for simplicity, here we set $\delta_{p,v} = \delta_{p,v'} + 0.01$, where v' is the last upstream intersection of v in path p, i.e., $v' = J(p, R(p, v) - 1)$. Our experiments show that such a simplification still obtains an advantageous effect. The score calculation on the tripartite graph consists of a forward phase and a reverse phase. In the forward phase, score propagation begins from the OD-pair nodes, and the OD-load scores are transformed into the popularity scores of their neighboring path nodes through the weighted matrix $X^{S\to P}$. Then the importance score of the intersections can be acquired using the weighted matrix $Y^{P\to V}$ via a similar procedure. H and C can be updated successively through

$$H = \alpha\left[X^{S\to P}\right]^T L + (1-\alpha)H^{(0)} \tag{13}$$

$$C = \alpha\left[Y^{P\to V}\right]^T H + (1-\alpha)C^{(0)} \tag{14}$$

where α is the damping factor, and is set to 0.85 like PageRank [29]. In the reverse phase, the scores spread in the opposite direction. H and L can be updated successively through

$$H = \alpha\left[Y^{V\to P}\right]^T C + (1-\alpha)H^{(0)} \tag{15}$$

$$L = \alpha\left[X^{P\to S}\right]^T H + (1-\alpha)L^{(0)} \tag{16}$$

The forward phase and the reverse phase are applied in an alternating fashion until convergence. To prevent the scores from an unbounded increase, H, C and L need to be normalized by $l_i = l_i/||L||$, $h_i = h_i/||H||$ and $c_i = c_i/||C||$ after each iteration. The meaning behind the score propagation algorithm is the mutual reinforcement to boost the co-linked nodes on the tripartite graph. In detail, in one way, a heavy-load OD-pair will choose a few or only one popular path, and a popular path will traverse many important intersections. In another way, an important intersection will be traversed by many popular paths, and a popular path will be chosen by a heavy-load OD-pair. CRRank has some unique advantages: a variety of key factors that affect the node ranking, such as whether a path can be substituted by others and the impact on the upstream node from the downstream are incorporated into the score propagation.

D. Time Complexity

In our framework, the computing cost of the information extraction phase depends primarily on the map matching, region partitioning and other external data preprocessing technologies. Thus, here we only focus on the time complexity of calculating the importance score. Suppose that the number of non-zero elements of $X^{S\to P}$ and $Y^{P\to V}$ are $|X|$ and $|Y|$. ODEF directly calculates the node importance according to the tripartite graph with its time complexity $|Y|$, since one path can only be assigned to one OD-pair. The score propagation of CRRank using multiplication of sparse matrix takes $O(n(|X| + |Y|))$ time with n iterations. In practice, the convergence of CRRank needs only a limited number of iterations. Thus, the efficiency is mainly determined by the numbers of OD pairs, paths and intersections. In brief, both ODEF and CRRank can efficiently rank the node importance in the road network from a city-wide perspective.

E. Example

In this subsection, we continue to consider the small network example in Fig. 1. To intuitively understand the two proposed algorithms, we compare them with two baselines: the flow-based algorithm (FA) and the betweenness-based algorithm (BA), both of which evaluate the importance of a node according only to its traffic flow and betweenness, respectively. All four algorithms are performed, and their node importance rankings are listed in Table 1. As presented, FA ranks node 6 as No. 1 due to the greatest flow, and BA ranks node 4 as No. 1 due to the highest betweenness. In the result of ODEF, node 1 and node 7 are tied for the first place, since both of them have the biggest mixed flows from two OD pairs, i.e., the flow and OD entropy of node 1 and node 7 are equal, respectively, and therefore ODEF cannot distinguish them. In comparison, CRRank also ranks node 7 as No. 1; however, node 1 is merely ranked as No. 3. Recall that when the scores propagate in the forward phase of CRRank, if an OD-pair has more than one path, the score of this OD-pair will be distributed to each path according to the corresponding weight matrix. This means that the more alternatives a path has, the lower its popularity, which lowers the importance of the intersections on this path. Note that the OD-pair $A \to E$ that traverses node 1 has two paths, while node 7 is on the only path of $K \to H$ and $K \to F$. Therefore, compared with ODEF, CRRank slightly raises the importance of node 7. Likewise, node 6 has only one biggest flow $I \to J$, while node 4 has two flows $A \to E$, $F \to B$. Particularly, $F \to B$ has only one path. Thus, CRRank ranks node 4 as No. 2, which is more important than node 6. In addition, node 2 and node 5 have only one equal flow, so they have the same ranks in

TABLE I
THE RANKING RESULT OF BA, FA, ODEF AND CRRANK ON THE EXAMPLE NETWORK

	FA		BA		ODEF		CRRank	
Rank	ID	Value	ID	Value	ID	Value	ID	Value
1	6	25	4	0.619	1,7	25.1984	7	0.4428
2	1,7	20	7	0.309	1,7	25.1984	4	0.4295
3	1,7	20	1	0.024	6	25	1	0.3987
4	4	15	2,3,5,8,6	0	4	16.254	6	0.3728
5	2,5	10	2,3,5,8,6	0	2,5	10	5	0.3449
6	2,5	10	2,3,5,8,6	0	2,5	10	2	0.3040
7	3	10	2,3,5,8,6	0	3	5	3	0.2876
8	8	0	2,3,5,8,6	0	8	0	8	0.1648

FA, BA and ODEF, but they have different ranks in CRRank. This difference is caused by node 5 being downstream from node 7 (No. 1). Imagine that if node 5 fails, it will likely cause the failure of node 7. To summarize this example, ODEF comprehensively considers flow and centrality to calculate the node importance. In addition to the factors considered in ODEF, CRRank incorporates the irreplaceability of the path and the impact on the upstream node from the downstream node.

V. EXPERIMENTS

In this section, we evaluate the performance and generality of the proposed algorithms. To achieve this, we carried out both a synthetic experiment and a real case study.

A. Synthetic Experiments

We generate a variety of large-scale networks using the NetworkX package of python.

1) **Random Network** is constructed using the Erdős–Rényi model, in which each pair of nodes has a uniform and independent probability p of linking to each other. We set $p = 0.4$, the average degree $\langle k \rangle = 4$ and the network size $N = 1024$.

2) **Small World Network** is constructed using the Watts-Strogatz model, which starts as a lattice ring, with each node connected to its neighbors, then randomly rewires each link of this network with a probability that forbids repeated wiring. We set $N = 1024$, $p = 0.01$, and $\langle k \rangle = 4$.

3) **Scale Free Network** is constructed with the Barabási-Albert model, which started with $m_0 = 4$ nodes. In each step, a new node is created with m links that connects to m nodes already in the network. We set $\langle k \rangle = 4$ and $N = 1024$.

4) **Regular Lattices** is the network, in which each node has exactly the same number of links. In our simulation, each node is connected to its four nearest neighbors. We set $N = 1024$.

For a given network, suppose that at each time step, a certain amount of traffic is generated between some random pairs of links. Each node is characterized by the designed capacity and the maximum length of its queue. The designed capacity e_i of node v_i is defined as the maximum traffic flow it can deliver in one unit of time under ideal conditions; the maximum length of the queue of v_i is denoted as q_i, $e_i < q_i$. When the vehicle number in the queue of a particular node exceeds its designed capacity, this node is considered to fail. For simplification, we ignore the effects of traffic signals. Each node in the network can be a router for trips between any pair of links. However, when the queue of a node is full, its upstream node cannot deliver a vehicle to it until there is a vacancy in its queue. Suppose that the duration of traversing a node for a vehicle is a function of the number of vehicles in its queue. Intuitively, this function should be monotonically increasing. To simulate such congestive effects, we use the well-known U.S. Bureau of Public Roads (BPR) function [22], [26], given by

$$d_i = \underline{d}[1 + \eta \left(\frac{q_i}{e_i}\right)^{\tau}] \tag{17}$$

where d_i is the cost of a vehicle passing through node v_i. $\underline{d}$ is the cost of a vehicle passing through a node at free flow; q_i is the number of vehicles in its queue. Here, $\eta = 0.15$ and $\tau = 4$ are constants. Each vehicle always takes the minimal travel cost path. We trace each vehicle and record its trajectories in the database.

We compare ODEF and CRRank with FA and BA, and define the following metric to evaluate the effectiveness of each algorithm.

Network efficiency is used to measure the transmission efficiency of the entire network. In the previous definition of the network efficiency, the path length is considered as the travel cost [6], [22]. Differing from these studies, we put more emphasis on the congestion phenomena, and thus the path cost is dynamically changed. We extend the previous definition of network efficiency as follows:

$$E(G) = \frac{1}{|R|(|R|-1)} \sum_{i \neq j \in R} \frac{d_{ij}}{\hat{d}_{ij}} \tag{18}$$

where d_{ij} and $\hat{d}_{ij}$ are the travel cost under free flow and the actual average travel cost, respectively, from edge i to edge j; $|R|$ is the total number of edges. Each edge can serve as an origin or a destination.

FA, BA, ODEF and CRRank are implemented individually on the four simulated networks, and distinct results of the node rankings are obtained. To compare these algorithms, at each turn, the designed capacity of one node is reduced to 10% according to the node rankings in descending order, and then the impact of the network on $E(G)$ is measured. After n turns, we can obtain the Decline Rates of $E(G)$ (DRE) corresponding to the distinct ranking nodes. The above steps are repeated five times to obtain the average results. Fig. 5

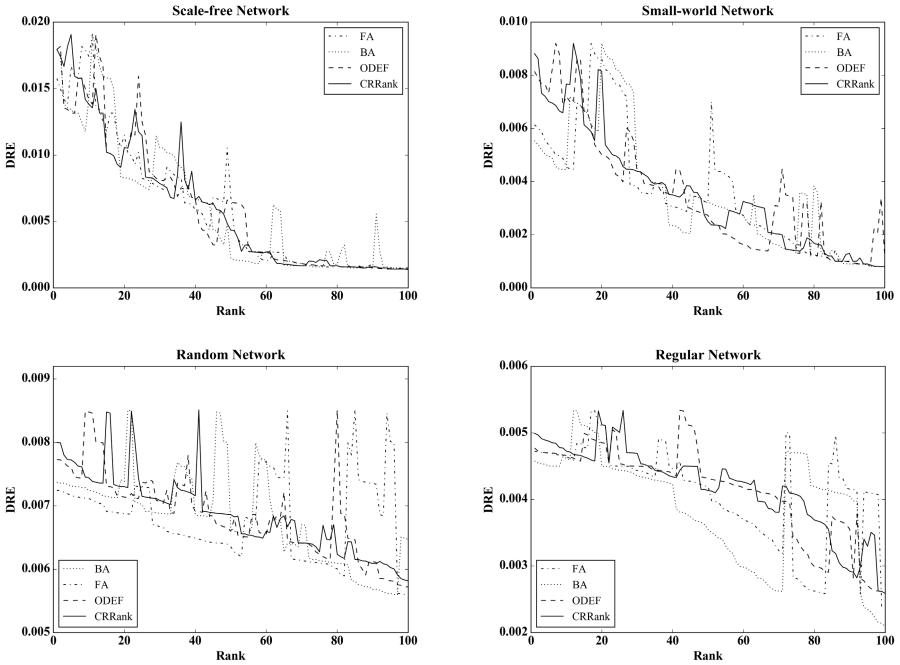


Fig. 5. The relation between DRE and the node ranking.

TABLE II
THE DIFF VALUE AND AVERAGE DRE OF RESULTS OF BA, FA, ODEF AND CRRANK

	Scale-free		Small-world		Random		Regular	
Algorithm	Diff	Avg.DRE	Diff	Avg.DRE	Diff	Avg.DRE	Diff	Avg.DRE
FA	0.0736	0.00616	0.1692	0.00356	0.2892	0.00690	0.3504	0.00370
BA	0.0828	0.00616	0.1448	0.00355	0.4992	0.00692	0.3880	0. 00351
ODEF	0.0718	0.00616	0.1392	0.00356	0.1960	0.00706	0.2216	0. 00403
CRRank	0.0706	0.00616	0.1012	0.00358	0.1244	0.00758	0.1520	0. 00419

shows the experiment results of the top-100 ranking nodes. In general, the variation in $E(G)$ can directly demonstrate the accuracy of the node ranking. That is, a higher DRE is expected following the failure of a more important node, and thus a high quality of ranking result should show a descending trend of the DRE when the node ranking decreases. To evaluate the ranking results, we define the difference of a ranking list Z from its expected order as

$$Diff\,(Z) = \frac{\sum_{i=1}^{n} |R\,(z_i, Z) - R\,(z_i, desc(Z))|}{\lfloor n^2/2 \rfloor} \qquad (19)$$

where z_i is the i^{th} element in Z; n is the length of Z; $desc(Z)$ is to sort Z in descending order; $R(z_i, Z)$ is the position number of z_i in Z. If Z is arranged exactly in descending order, which is expected, the value of $Diff(Z)$ is 0, whereas if Z is arranged in ascending order, the value of $Diff(Z)$ is 1, $0 \leq Diff(Z) \leq 1$. Table 2 presents the *Diff* values and average DRE of the ranking results of each algorithm for each simulated network. In the scale free network and the small world network, there is a clear correlation between DRE and the node rankings for all four algorithms, and the effectiveness of these algorithms are nearly the same (see Fig. 5 and Table 2). The reason is that the nodes with large traffic flows are more likely to have relatively high betweenness centralities and vice versa in these two networks. Therefore, regardless of the algorithms are based on traffic volume, betweenness centrality or both of them, they will give the similar ranking results. In the random network, both of FA and BA exhibit large fluctuations of DRE and have high *Diff* values, while ODEF and CRRank have a relatively clear decline trend. In particular, CRRank shows the best results. In the regular network, CRRank still outperforms the other algorithms, and ODEF performs better compared with FA, while BA shows the worst performance. Furthermore, we evaluate the impact on the simulated network of top-20 nodes for each algorithm. Specifically, as a function of time, the cumulative number $F_c(t)$ of failed nodes is investigated in Fig. 6. To clearly distinguish the different algorithms, the common nodes discovered by all four algorithms are excluded. For all the networks, $F_c(t)$ exhibits an initial rapid growth, and then tends to be stable with the increase of time. Across the measured time, CRRank has the most rapid growth and obtains the highest steady-state value, demonstrating its advantages over all other algorithms. Compared with BA and FA, ODEF has a relatively good and stable performance.

B. Case Study

Road network: the road network of Beijing contains 13,722 intersections and 25,178 road segments.

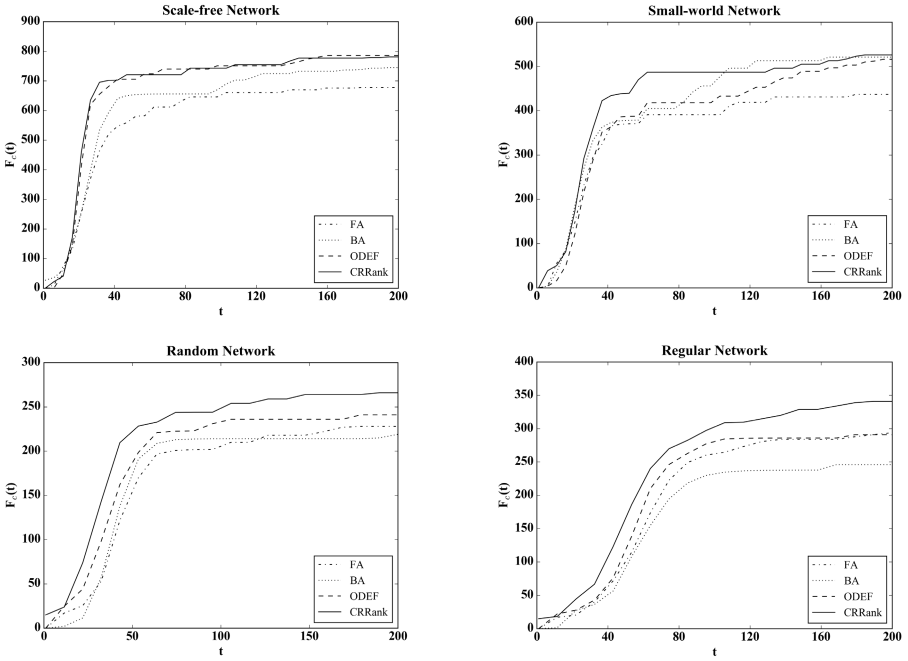


Fig. 6. The relation between the cumulative number of failed nodes and the increased time.

The map is segmented into 478 regions using the high-level roads.

Vehicle trajectories: we use a trajectory dataset generated by approximately 30,000 taxicabs in Beijing over a period of one month (October, 2016). The sampling interval of the dataset is between 30 seconds and 60 seconds. Considering the morning and evening traffic peaks of the road network, we choose the trajectories during 7:30∼10:00 and during 17:00∼19:30. As a result, we extract approximately 161,000 valid trips (carrying passengers). Due to noise in the trajectories and mismatches of the map matching, we remove the invalid OD-pairs with less than 5 trips. As a result, we obtain 4,756 OD-pairs and 13,270 paths. The tripartite graph is built based on this information.

Fig. 7 highlights the visualized results of the ranking intersections of BA, FA, ODEF and CRRank. The intersections marked with points in different size are the top-60 most important nodes. We observe that Beijing road network formed four ring-shape loops from the center outwards, which are called the 2^{nd}, 3^{rd}, 4^{th} and 5^{th} ring road, respectively. These ring roads are the arteries for the urban traffic. The No. 1 and No. 2 most important nodes discovered by all four algorithms are the same. However, many important intersections discovered by BA are located mainly within the 2^{nd} ring road. The other three algorithms tend to identify the important nodes on the ring roads in a wider range. We further discover that although some intersections within the 2^{nd} ring have high betweenness centralities and seem to be transport hubs, they do not have significant traffic flow. In fact, these intersections are very close to the shopping malls, railway stations or famous scenic spots, which are generally origins or destinations of trips. For example, intersection A near one of the largest shopping centers "Xidan Market" has a high ranking at No. 14 in the results of BA, whereas its ranking are 45^{th}, 58^{th}, and 82^{nd} in FA, ODEF and CRRank, respectively. For the daily trips, many drivers prefer the ring roads that are detours with less congestion to reach their destination quickly. To guarantee the free flow on the ring roads, many overpasses have been built instead of traditional signalized intersections. The reduction in the number of intersections on the ring roads leads to their high popularity. Consequently, the importance of the remaining intersections on the ring roads is further boosted.

Comparing with FA, ODEF and CRRank both slightly lower the rankings of the intersections on the 4^{th} ring road. The reason is that the number of OD-flows on the 4^{th} ring road is relatively small, and thus the centralities of the intersections are not as high as those on the 2^{nd} or 3^{rd} ring road. Moreover, the OD-flows on the 4^{th} ring road are commonly distant and large, and their origins or destinations are likely near the 4^{th} ring, which is consistent with our daily experience. With the continuous expansion of the city, many residential areas have migrated outside the 4^{th} or 5^{th} ring, which results in the emergence of many distant OD-flows.

A difference between ODEF and CRRank is that CRRank can raise the rankings of the intersections on the irreplaceable path or the downstream intersections of large flows. As presented in Fig. 7, the rankings of intersection B on the only path to the airport are 56^{th}, 51^{st}, 47^{th} and 19^{th}, given by BA, FA, ODEF and CRRank, respectively. Although the centrality of intersection B and its traffic are not relatively very high, its failure will greatly reduce the network efficiency and seriously affect the experience of travelers. Therefore, such a node should be given adequate attention, as is done in CRRank.

Another function of CRRank is that the OD-load and path popularity can be evaluated. Four OD-pairs and their main candidate paths are shown in Fig. 8. The OD-pair s_1,

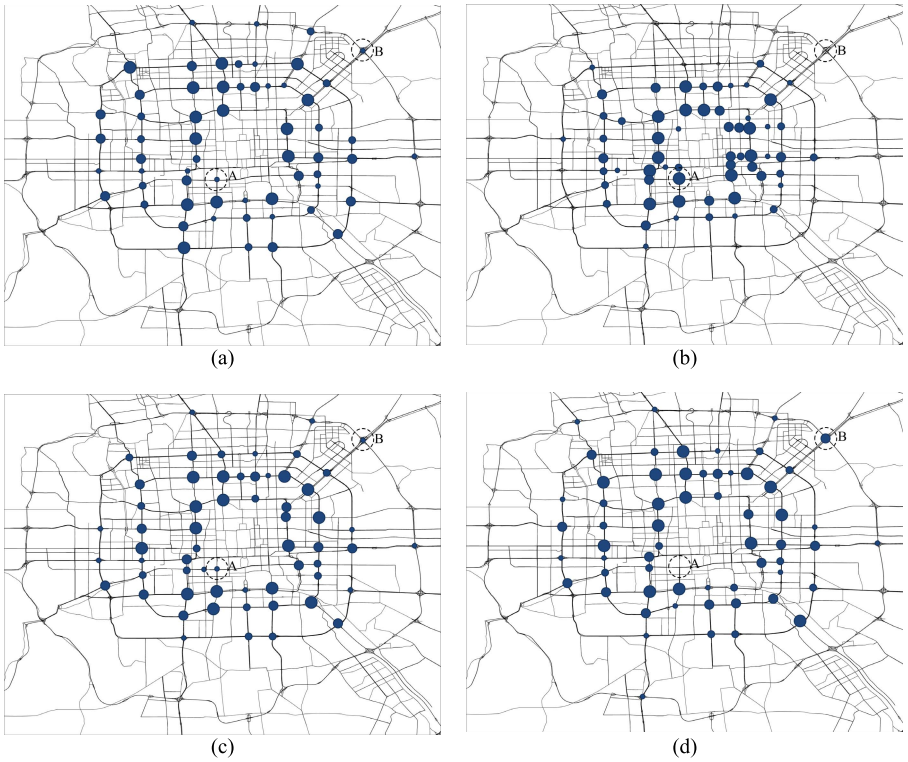


Fig. 7. The top-60 important intersections in real road network evaluated by different methods. (a) FA. (b) BA. (c) ODEF. (d) CRRank.

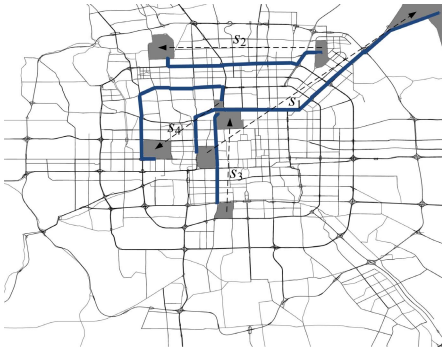


Fig. 8. An example of evaluation on OD-load derived from different OD-pairs.

from "Jinrong" Street (a famous CBD) to the airport, has 4,821 trips. It has the largest flow for all the OD-pairs and traverses many important intersections. Accordingly, CRRank determines that s_1 places the highest load on the road network, and its path has the highest popularity. The OD-pair s_2, from the "Wangjing" sub-district (Beijing's Koreatown) to "Wudaokou" (a commercial and educational center), has 728 trips. In CRRank, its initial ranking is higher than the OD-pair s_3, which has 630 trips from Beijing's South Railway Station to the "Houhai" scenic location. However, the final ranking of s_2 is lower than that of s_3. This is because the path of s_3 traverses a larger number of more important intersections located on the 3rd ring road. Although the traffic flow ofs_3 is smaller, a confluence of s_3 and other OD-flows at these important intersections may have a greater impact on the road network. Given the last example, despite the path of the OD-pair s_4 (from the Chinese Academy of Engineering to the Military Museum) traverses many important intersections, it has still a weak impact on the road network, since it has only 11 trips. CRRank merely boosts the ranking of s_4 from 4234^{th} to 3998^{th}.

VI. Conclusion and Future Work

We propose a data-driven framework to rank the node importance in a city-wide road network via data mining from massive vehicle trajectories. In our framework, we model the trip network characterizing the dynamics of the road network using a tripartite graph. Based on the trip network, we present two ranking algorithms, ODEF and CRRank. The synthetic experiment demonstrates that our methods outperform typical baseline approaches in the simulated networks with different topologies. We also evaluate our methods using a case study with a trajectory dataset generated by over 30,000 taxis. Compared with the baseline approaches, the results of our methods are more reasonable. In addition, some other interesting findings, such as the heavy-load OD-pairs and popular paths, can be observed using CRRank.

In the future, we would like to apply our methods to more road networks of additional cities. The importance rankings of intersections would help traffic bureaus avoid widespread congestion and improve the efficiency of road networks.

References

[1] D. Li, J. Yinan, K. Rui, and S. Havlin, "Spatial correlation analysis of cascading failures: Congestions and blackouts," *Sci. Rep.*, vol. 4, Jun. 2014, Art. no. 5381.

[2] Y. Qian, B. Wang Y. Xue, J. Zeng, and N. Wang, "A simulation of the cascading failure of a complex network model by considering the characteristics of road traffic conditions," *Nonlinear Dyn.*, vol. 80, nos. 1–2, pp. 413–420, 2015.

[3] S. Lämmer, B. Gehlsen, and D. Helbing, "Scaling laws in the spatial structure of urban road networks," *Phys. A, Stat. Mech. Appl.*, vol. 363, no. 1, pp. 89–95, 2006.

[4] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Physica A, Stat. Mech. Appl.*, vol. 391, pp. 1777–1787, Feb. 2012.

[5] P. Hu, W. Fan, and S. Mei, "Identifying node importance in complex networks," *Physica A, Stat. Mech. Appl.*, vol. 429, pp. 169–176, Jul. 2015.

[6] J. Liu, Q. Xiong, W. Shi, X. Shi, and K. Wang, "Evaluating the importance of nodes in complex networks," *Physica A, Stat. Mech. Appl.*, vol. 452, pp. 209–219, Jun. 2016.

[7] G. Yan, T. Zhou, B. Hu, Z.-Q. Fu, and B.-H. Wang, "Efficient routing on complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 73, no. 4, p. 046108, 2006.

[8] E. Estrada and N. Hatan, "Communicability in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 77, no. 3, p. 036111, 2008.

[9] S. Dolev, Y. Elovici, and R. Puzis, "Routing betweenness centrality," *J. ACM*, vol. 57, no. 4, p. 25, 2010.

[10] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the Web," Stanford InfoLab, Palo Alto, CA, USA, Tech. Rep. SIDL-WP-1999-0120, Nov. 1999.

[12] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PLoS ONE*, vol. 6, no. 6, p. e21202, 2011.

[13] R. Lempel and S. Moran, "The stochastic approach for link-structure analysis (SALSA) and the TKC effect," *Comput. Netw.*, vol. 33, nos. 1–6, pp. 387–401, 2000.

[14] S. Iyer, T. Killingback, B. Sundaram, and Z. Wang, "Attack robustness and centrality of complex networks," *PLoS ONE*, vol. 8, no. 4, p. e59613, 2013.

[15] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, p. 38, 2014.

[16] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. KDD*, Beijing, China, 2012, pp. 186–194.

[17] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proc. UrbComp*, Beijing, China, 2011, pp. 89–98.

[18] S. Chawla, Y. Zheng, and J. Hu, "Inferring the root cause in road traffic anomalies," in *Proc. ICDM*, Brussels, Belgium, 2012, pp. 141–150.

[19] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing, "Discovering spatio-temporal causal interactions in traffic data streams," in *Proc. KDD*, San Diego, CA, USA, 2011, pp. 1010–1018.

[20] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *Proc. GIS*, Orlando, FL, USA, 2013, pp. 344–353.

[21] L. Hong, Y. Zheng, D. Yung, J. Shang, and L. Zou, "Detecting urban black holes based on human mobility data," in *Proc. GIS*, Seattle, WA, USA, 2015, p. 35.

[22] J. J. Wu, Z. Y. Gao, and H. J. Sun, "Effects of the cascading failures on scale-free traffic networks," *Physica A, Stat. Mech. Appl.*, vol. 378, no. 2, pp. 505–511, 2007.

[23] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Map-matching for low-sampling-rate GPS trajectories," in *Proc. GIS*, Seattle, WA, USA, 2009, pp. 352–361.

[24] N. J. Yuan, Y. Zheng, and X. Xie, "Segmentation of urban areas using road networks," Microsoft Corp., Redmond, WA, USA, Tech. Rep. MSR-TR-2012-65, 2012.

[25] M. Xu *et al.*, "Discovery of important crossroads in road network using massive taxi trajectories," presented at the 3rd Int. Workshop Urban Comput., New York, NY, USA, Aug. 2014.

[26] J. J. Wu, Z. Y. Gao, H. J. Sun, and H. J. Huang, "Congestion in different topologies of traffic networks," *Europhys. Lett.*, vol. 74, no. 3, p. 560, 2006.

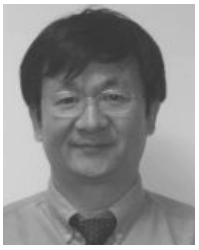
[27] J. Shetty and J. Adibi, "Discovering important nodes through graph entropy the case of Enron email database," in *Proc. LinkKDD*, Chicago, IL, USA, 2005, pp. 74–81.

[28] M. Dehmer, "Information processing in complex networks: Graph entropy and information functionals," *Appl. Math. Comput.*, vol. 201, nos. 1–2, pp. 82–94, 2008.

[29] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, 1998.

Ming Xu received the Ph.D. degree in computer science from Beijing University of Posts and Telecommunications, Beijing, China, in 2015.

He currently holds a postdoctoral position with the Department of Civil Engineering, Tsinghua University. His research interests include urban computing, spatio-temporal data mining, deep learning, reinforcement learning, and application in intelligent transport systems.



Jianping Wu received the B.S. and M.S. degrees in civil engineering from Zhejiang University, in 1982 and 1984, respectively, and the Ph.D. degree in transportation engineering from University of Southampton, U.K., in 1994.

He is currently a Professor with the Department of Civil Engineering, Tsinghua University, Beijing, China, and the Director of Tsinghua University–Cambridge University and MIT Low Carbon Alliance Center for Future Transport Research. He is also a Visiting Professor with University of Southampton, U.K. He was a recipient of the prestigious Chong Kong Scholar Professorship awarded by the Ministry of Education of China.

His research interests include drivers' behavior and microscopic simulation, sustainable transport systems and low carbon transport, and smart city and ITS.

Mengqi Liu is currently pursuing the B.S. degree in telecommunication engineering with Beijing University of Posts and Telecommunications, Beijing, China.

From 2016 to 2017, she was a Research Assistant with the Future Transport Research Center, Tsinghua University, Beijing, China. Her research interests include urban computing, data mining, deep learning, and machine learning.

Yunpeng Xiao received the B.Sc. degree in information system from Chongqing University of Posts and Telecommunications, China, and the master's and Ph.D. degrees in computer science and engineering from Beijing University of Posts and Telecommunications, China. He is currently an Assistant Professor with the Software Engineering College, Chongqing University of Posts and Telecommunications, China. His current research interests are in the areas of social network analysis, big data analysis, and mobile Internet.

Haohan Wang received the bachelor's degree from Beijing University of Posts and Telecommunications and the master's degree from Carnegie Mellon University, in 2012 and 2014, respectively, where he is currently pursuing the Ph.D. degree with the School of Computer Science, Language Technologies Institute, all in computer science. His research interest mainly covers computational biology, and statistical machine learning methods with an emphasis on deep learning methods.

Dongmei Hu received the M.Sc. degree in environmental science and engineering from Taiyuan University of Technology in 2013. She is currently pursuing the Ph.D. degree in civil engineering with Tsinghua University. She is currently a Visiting Graduate Researcher in environment health sciences with University of California at Los Angeles, Los Angeles.

Her research interest is primarily in the field of traffic and air pollution. Specifically, her current research focuses on data mining in urban air quality and quantitative exposure/risk assessments on fine particles and ultrafine particles from various indoor and outdoor sources.