Inferring Traffic Cascading Patterns

Yuxuan Liang[1,2], Zhongyuan Jiang[1], Yu Zheng[2,1,3*]
[1]Xidian University, Xi'an, China
[2]Urban Computing Group, Microsoft Research, Beijing, China
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
{v-yuxuli,yuzheng}@microsoft.com,zyjiang@xidian.edu.cn

ABSTRACT

There is an underlying cascading behavior over road networks. Traffic cascading patterns are of great importance to easing traffic and improving urban planning. However, what we can observe is individual traffic conditions on different road segments at discrete time intervals, rather than explicit interactions or propagation (e.g., A→B) between road segments. Additionally, the traffic from multiple sources and the geospatial correlations between road segments make it more challenging to infer the patterns. In this paper, we first model the three-fold influences existing in traffic propagation and then propose a data-driven approach, which finds the cascading patterns through maximizing the likelihood of observed traffic data. As this is equivalent to a submodular function maximization problem, we solve it by using an approximate algorithm with provable near-optimal performance guarantees based on its submodularity. Extensive experiments on real-world datasets demonstrate the advantages of our approach in both effectiveness and efficiency.

CCS CONCEPTS

• **Information systems → Spatial databases and GIS**;

KEYWORDS

Spatio-temporal Data Mining, Urban Computing

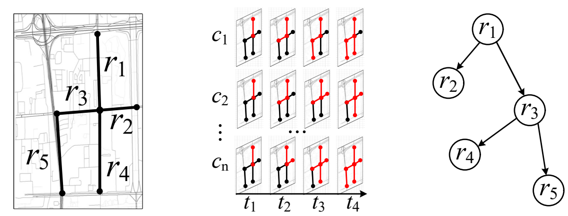
1 INTRODUCTION

Traffic is an important part of urban lives, affecting a city's sustainability. Traffic congestion in a city's road network usually spreads or relieves through the cascading patterns [11]. For example, based on the traffic conditions on road segments over a period of day (as shown in Figure 1(a)), we can find the cascading pattern (depicted in Figure 1(c)) that r_1 first gets congested leading to a possible congestion on r_2 and r_3; r_3 further results in the congestion on r_4 and r_5 possibly. Knowing such cascading patterns can help predict future traffic conditions and identify bottlenecks of road networks, thereby improving urban planning. However, mining traffic cascading patterns is challenging due to the following three reasons:

*The paper was done when the first author was an intern at Microsoft Research under the supervision of the third author. Yu Zheng is the correspondence author.

SIGSPATIAL'17, November 7–10, 2017, Los Angeles Area, CA, USA

ACM ISBN 978-1-4503-5490-5/17/11…$15.00
https://doi.org/10.1145/3139958.3139960



(a) Road Network (b) Observed Cascades (c) Cascading Pattern

Figure 1: A simple example of traffic cascading pattern.

1) **Implicit interaction**: Different from the explicit interactions (e.g. user A forwards a tweet from B) in social networks, the real propagation path of traffic conditions is implicit. What we can only observe is traffic conditions (e.g., travel speed) on individual road segments at discrete time intervals, i.e., we note *when* and *where* traffic congestion occurs but not how it reaches the specific road segments. Moreover, inferring traffic cascading pattern is not equivalent to mining frequent subgraphs over the graph consisting of real propagation paths. For example, the congestion on r_1 appears rarely in our observed data; But as long as r_1 gets congested, r_2 gets congested successively; Although the interaction from r_1 to r_2 occurs rarely, it has relatively high dependency.
2) **Multiple sources**: The traffic on a road segment comprises of two parts: the traffic flowing from other road segments and traffic originating from the neighboring road segments. Thus, there are two categories of factors affecting the traffic conditions on a road segment: a) Neighboring traffic: Traffic conditions on a road segment depend on that of its neighbors which may be physically connected like $r_1 \rightarrow r_2$ shown in Figure 2, or several-hops connected to the road segment, e.g., $r_2 \dashrightarrow r_4$. We define these two categories of influence from neighboring traffic as *direct influence* and *indirect influence*. b) Surrounding environment. Traffic originated from a road segment is influenced by its surrounding environment, such as POIs. For instance, the congestion on r_1 in Figure 2 derives from the *environmental influence* like the shopping malls. It is challenging to simultaneously capture both of them into the inference of traffic cascading patterns.
3) **Geospatial correlations**: In social networks, the information spreading is strongly correlated to the waiting time [7, 11], e.g., the time difference between two persons' tweet time. Besides the time difference of traffic congestion, there are geospatial correlations over traffic networks. We use a series of road segments in Figure 1 for an illustration. Suppose r_1 first gets congested, r_3 is easier to be impacted by r_1 than r_5 because of the spatial distances. Hence, how to integrate the temporal and geospatial correlations remains a challenge.

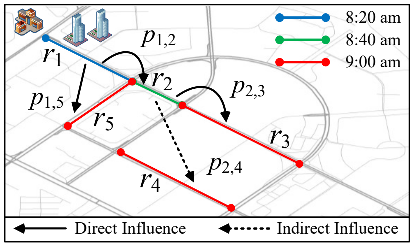


Figure 2: Illustration of two factors affecting traffic conditions, where the lines with different colors denote traffic congestion occurs at different time intervals. For example, we observe that r_1, r_2 and r_3 get congested at 8:20 am, 8:40 am, 9:00 am respectively.

There are no existing methods that can address these challenges simultaneously. To this end, we propose a generative probabilistic approach *CasInf* for the inference of traffic cascading patterns, which consists of three main components: an *Individual Transmission Likelihood (ITL)* model, an *EnvironMental inTensity (EMT)* model and a *Cascading Pattern Construction* algorithm. The contributions of our study lie in the following three aspects:

- *Modeling three-fold influences*: The *ITL* model captures the temporal and geospatial correlations to describe the direct and indirect influence between a pair of road segments. Incorporating the temporally- and spatially-related features, the *EMT* model utilizes a SVM-based method to infer the intensity of environmental influence.
- *Cascading pattern inference*: We propose a two-step generative probabilistic model that first formulates the influences from multiple sources into propagation trees over a graph and then applies an approximate algorithm with provable near-optimal guarantees to solve the NP-hard problem.
- *Real evaluation & case study*: We evaluate our method based on the real-world datasets over a period of nine months. Extensive experiments show the advantages of our method. We also present a case study to support our method.

2 OVERVIEW

Due to the equivalence between the spreading and relief process of traffic congestion, we take the spreading process as an example to illustrate how to infer the underlying cascading patterns.

2.1 Preliminary

Definition 1 (*Road network*): A road network R is composed of a set of road segments r, connecting each other in the format of a directed graph. Each road segment r is a directed edge having two terminal nodes, a length $r.len$, a level $r.lev$ denoting its capacity.
Definition 2 (*Traffic condition*): Following the state-of-the-art approach for congestion propagation detection [17], we assume the traffic condition on a road segment r at a specific time interval has a unique status: congested or smooth. If the travel speed on r at a given time slot exceeds 20km/h, we say r is congested [1].
Definition 3 (*Cascade*): A cascade $c = \{t_1, t_2, \ldots, t_n\}$ is a n-D (n-dimensional) vector extracted from one day's observed traffic conditions where t_i records when r_i gets congested and $r_i \in R$. In fact, most cascades hit not all road segments, so we set $t_i = \infty$ when r_i is not hit by a cascade.
Definition 4 (*Cascading Pattern*): A cascading pattern G is considered as an underlying *network* over which traffic congestion spreads in a given time span. In fact, the cascading pattern is time-evolving, thus we assume the cascading pattern in a specific time span (e.g., morning rush hours) is uniform. We denote the directed edges connecting two different road segments in the cascading pattern as *casual links*, which represent the casual correlations between them.

2.2 Problem Statement

Given a road network R, a time span ts including several time intervals from t_s to t_e, and the discrete traffic conditions at each time interval during a period of m days, we extract a set of cascades $C = \{c_1, \ldots, c_m\}$, where c_i denotes the cascade extracted from the traffic conditions in ts of the i-th day. Based on the set of cascades C, POIs located over R and the meteorological data, we aim to infer the cascading pattern G in ts that maximizes the likelihood of occurring cascades C, i.e., best explains the observed traffic conditions.

2.3 Framework

Figure 3 presents the framework of our approach, which consists of four major parts: 1) Data acquisition. We acquire four real-world datasets including POIs data, meteorological data, road networks data and taxi trajectories data from urban areas. 2) Multiple sources modeling. We propose the *ITL* model to infer the transmission likelihoods between road pairs. Meanwhile, we extract spatio-temporal features from real-world datasets and employ the *EMT* model to infer the intensity of environmental influence. (we use environmental intensity for short in the following writing) 3) Cascading pattern inference. We formulate the likelihoods and intensity from multiple sources into many propagation trees over a graph and use an approximate algorithm to efficiently solve the maximum likelihood problem. 4) Evaluation. We justify the effectiveness and efficiency of our approach in this part. We will detail the last three major parts in the following sections respectively.

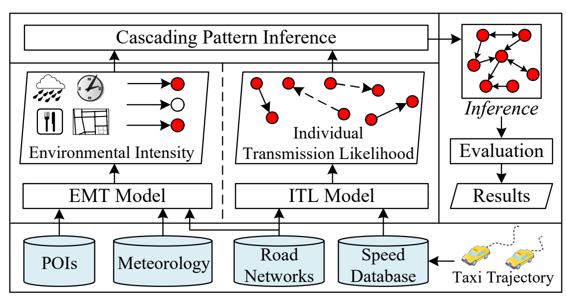


Figure 3: The framework of our approach.

3 MULTIPLE SOURCES MODELING

In this section, we first outline the general idea of the individual transmission likelihood, applying this model to infer the direct and indirect influence. Second, we analyze the spatio-temporal factors correlations and address the inference of environmental intensity.

3.1 Individual Transmission Likelihood

In general, traffic spreading between road segments can be expressed by the Independent Cascading model [7, 11] based on the congested time difference, which ignores the geospatial correlation between them. Here we develop an Individual Transmission Likelihood (*ITL*) model to infer both direct influence and indirect influence based on the spatio-temporal distances.

We consider the time difference between a pair of congested road segments as the waiting time that passed between the congestion appearance times t_i and t_j. Generally, more waiting time results in a less probability of spreading. e.g., as shown in Figure 2, though probability $p_{1,2}$ and $p_{1,5}$ are both induced by direct influence, it is more likely to see that r_1 infects r_2 than the latter pair. Moreover, the risk of direct influence is much higher than indirect influence owing to the spatial distance between roads, such as $p_{2,3} > p_{2,4}$.

Suppose r_j gets congested at time t_j and r_i gets congested at time t_i successively, i.e., $t_j < t_i$. Inspired by a well-studied monotonic exponential model in social networks [2, 11], here we combine the spatial and temporal distances into the exponential model to describe the *conditional likelihood* of transmission (i.e., individual transmission likelihood) from r_j to r_i as follows:

$$f(t_i|t_j; a_{j,i}, \lambda) \propto e^{-a_{j,i}(\Delta_{j,i}+\lambda * d_{i,j})}, \quad (1)$$

where $\Delta_{j,i} = t_i - t_j$ denotes the time difference (i.e., temporal distance) between the road congested time, $d_{i,j}$ denotes the road network distance from r_i to r_j (i.e., spatial distance), $a_{j,i}$ is the transmission rate assumed to be a constant value α for all pairs, and λ controls the importance of spatial distance as a trade-off. In particular, $\Delta_{j,i}$ and $d_{i,j}$ need to be normalized.

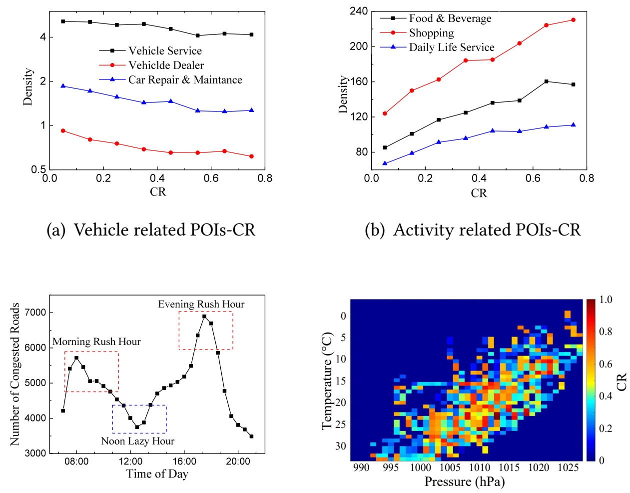
3.2 Environmental Intensity Inference

When inferring the environmental intensity, we face two challenges: 1) Uncertain correlations between traffic congestion and multiple environmental factors. 2) No available ground truths. It is a non-trivial task to obtain the congestion's occurring probability on a fixed road segment caused by the environmental factors.

3.2.1 Environmental Factors Correlation. To analyze the correlations between the traffic and the surrounding environment, we use the congestion ratio (CR) as the main indicator to show the long-term traffic conditions on each road, which describes the average frequency of a road segment in congested status during a day.

It has been well studied in the traffic-related researches [23, 24, 28] that the traffic condition has a strong correlation with the spatial factors, such as road networks (e.g., $r.len$, $r.lev$). Besides, the categories and density of POIs in a region indicate not only the land use and the function of the region but also the traffic patterns in the region [27]. As depicted in Figure 4(a), CR was decreased by the density of vehicle-related places since they are usually deployed in the remote areas instead of downtown. From Figure 4(b), it can be seen that denser human activity related POIs will lead to heavier traffic, which meets our common knowledge.

Besides spatial factors, temporal factors have indirect impacts on traffic flows, which provide complementary information of human behaviors. For example, the number of congested roads changing over time of day in Figure 4(c) reveals that the persons' commuting between their workplace and home causes the busy traffic. Likewise,



(a) Vehicle related POIs-CR (b) Activity related POIs-CR

(c) Traffic congestion-Time of day (d) Meteorology-CR

Figure 4: Correlations between traffic conditions and the environmental factors.

the meteorology is correlated to the traffic conditions indirectly. For example, when it goes rainy, people prefer driving their cars to where they want rather than travelling on foots. In Figure 4(d) where color, x- and y-axis denote CR, pressure and temperature respectively, there is an interesting observation that more instances with higher CR occur when temperature is increased by pressure, especially in the oval-shaped area.

In summary, we identify the following spatial features of a region: the POIs density in most related categories as the POIs features, total lengths and road density as the road networks features. We also extract temperature, pressure, and time of day as the temporal features. Single-view data may only tell us a part of the panoramic view of traffic. That is the reason why we need to incorporate multiple source features to infer the environmental intensity.

3.2.2 Supervised Learning. Due to the spatial dependency, we divide a city into disjoint grids (0.5 km * 0.5 km) denoting the different regions, assuming the environmental intensity (denoted as ε_g) is uniform for all road segments in a given grid g. To address the second challenge, we notice that the *percentage* of congested road segments in a small region g still reflects the magnitude of the environmental intensity in g. As detailed in Figure 5, suppose there is no diffusion between road segments, the percentage value can approximately denote the occurring likelihood of congestion on each road in the region. Hence, given a grid g and the spatio-temporal features, inferring the environmental intensity in g is approximately equivalent to predicting the percentage of congested road segments in g at that time interval.

We extract spatio-temporal features of different grids during a series of days and obtain the aforementioned percentage data from the historical traffic data. Incorporate the spatio-temporal features as the input of a supervised learning approach (we use SVM in this study), we try to predict the percentage value, so as to infer ε_g.

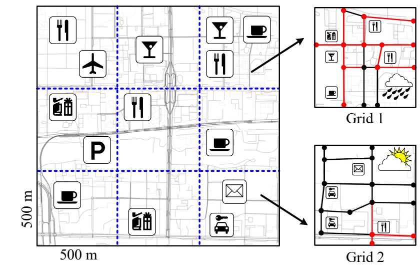


Figure 5: Traffic condition at 18:00, where red lines denote congested road segments. Traffic condition in g_1 is heavier than g_2, i.e., the environmental intensity in g_1 is stronger due to (1) POIs and time of day. People are more likely to go for dinner at that time. (2) Weather. In the rainy day, human's traveling mainly depends on vehicles. Here we can compare the environmental intensity in different regions by their congested roads percentage approximately.

4 CASCADING PATTERN CONSTRUCTION

In this section, we detail how to combine the three-fold influences into a cascading pattern through a generative probabilistic model. We use the word *graph* or *network* to denote the cascading pattern and *edge* to denote the casual link in the cascading pattern for short.

4.1 Model Formulation

The previous works consider the traffic congestion diffuses in the format of directed trees such as STCTree [17]. i.e., a road segment gets congested because of another single road segment. Following this propagation tree assumption, given a propagation tree T, we compute the likelihood of a single cascade c as follows:

$$f(c|T) = \prod_{(j,i)\in E_T} f(t_i|t_j;\alpha,\lambda), \tag{2}$$

where E_T is the edge set of propagation tree T. Generally, there are many possible propagation trees in a given cascading pattern G, i.e., an underlying graph G. For a given network G, considering all possible propagation trees T supported by G, we compute the likelihood of a cascade as follows:

$$f(c|G) = \sum_{T\in\mathcal{T}_c(G)} f(c|T)P(T|G), \tag{3}$$

where $\mathcal{T}_c(G)$ denotes the set of all the directed connected spanning trees over the subnetworks of G. Suppose we obtain a cascade c, there are many possible propagation trees that can create this cascade as illustrated in Figure 6. We assume r_j impacts r_i with a constant prior probability, hence $P(T|G)$ is equal for all trees T in cascade c and we simplify Equation 3 as follows:

$$f(c|G) \propto \sum_{T\in\mathcal{T}_c(G)} \prod_{(j,i)\in E_T} f(t_i|t_j;\alpha,\lambda), \tag{4}$$

For simplicity, we assume the cascades are conditional independent over a given network G, therefore the joint likelihood of a set of cascades C taking place in network G can be easily denoted as:

$$f(C|G) = \prod_{c\in C} f(c|G). \tag{5}$$

Given the observed traffic data including a set of cascades C obtained from a given time span of each day and pairwise transmission likelihoods based on spatio-temporal distances, we formulate this problem as a *Network inference problem* [4]. Our target is to find a network $\hat{G}$ such that

$$\hat{G} = \arg\max_{|G|\le k} f(C|G), \tag{6}$$

where $\hat{G}$ best explains the observed cascades, and the maximization is over all probable graphs G of at most k edges since real-world networks are always sparse such as road networks.

4.2 Alternative Target

Since it is intractable to directly optimize the target function in Equation 6, it can be changed into an alternative target [20]. The matrix tree theorem [3] states that the number of nonidentical spanning trees of a graph G is equal to any cofactor of its Laplacian matrix. Through this theorem, Equation 4 can be simplified into

$$f(c|G) \propto \prod_{t_i\in c} \sum_{t_j\in c:t_j\le t_i} f(t_i|t_j;\alpha,\lambda). \tag{7}$$

Remember the environmental influence, which can induce the *first* congested road segment and create disjoint cascades, occurs everywhere in a graph even an empty graph. To address this issue, the general network inference problem considers the external factors as an additional node that can infect *any* nodes with a constant probability ε [4, 20], but in reality it is time- and location-varying. Different from conventional studies, we infer the environmental intensity in different grids detailed in Section 3.2. Then we compute the improvement of log-likelihood for cascade c under graph G over an empty graph $\bar{K}$ as follows:

$$F(c|G) = \sum_{t_i\in c} log \sum_{t_i\in c:t_j\le t_i} w_c(j,i), \tag{8}$$

where $w_c(j,i) = \theta\varepsilon_g^{-1} f(t_i|t_j;a_{j,i},\lambda)$, which can be considered as the weight of edge (r_j, r_i) in c. The variable ε_g denotes the environmental intensity of grid g where r_i locates in, and θ is a trade-off parameter between the three-fold influences in the log-likelihood function. Finally, the optimization target in Equation 6 is equivalent to maximizing the following objective function:

$$\hat{G} = \arg\max_{|G|\le k} F(C|G) \tag{9}$$

where $F(C|G) = \sum_{c\in C} F(c|G)$ is non-negative monotonic and the maximization is over all probable graphs G of at most k edges. Next, we will present an approximate algorithm based on submodular function optimization to maximize the objective function $F(C|G)$.

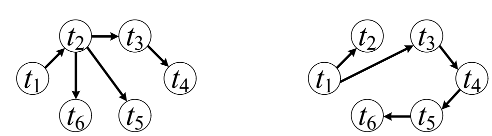


Figure 6: Two simple examples of possible propagation trees of a cascade $c = \{t_1, t_2, \ldots, t_n\}$, where $t_i > t_{i-1}$.

4.3 Optimization

It is proved that finding the optimal solution to a network inference problem is NP-hard [4, 20]. Equivalent to the MAX-k-COVER problem, it satisfies *submodularity*, a natural diminishing returns property. The *proof* is stated as follows:

Given a cascade c, we first prove the submodularity of $F(c|G)$, i.e, prove $F(c|G \cup e) - F(c|G) \geq F(c|G' \cup e) - F(c|G')$, where graph $G \subseteq G'$ and edge $e = (r_1, r_2)$ is not contained in G'. Let $w_c(r_1, r_2)$ be the weight of edge (r_1, r_2) in G. We notice that $w_c(r_1, r_2)$ is never more than its weight $w'_c(r_1, r_2)$ in G'. i.e., $w'_c(r_1, r_2) \geq w_c(r_1, r_2) \geq 0$. Particularly if (r_1, r_2) is contained in both G and G', then $w_c(r_1, r_2) = w'_c(r_1, r_2)$. Let $T_{A,e} = \sum_{i \in A \setminus \{r_1\}} w_c(i, r_2)$ and it satisfies that $T_{G',e} \geq T_{G,e}$. Thus, we have

$$
\begin{aligned}
F(c|G \cup e) - F(c|G) &= log \frac{T_{G,e} + w_c(r_1, r_2)}{T_{G,e}} \\
&\geq log \frac{T_{G',e} + w_c(r_1, r_2)}{T_{G',e}} \\
&= F(c|G' \cup e) - F(c|G').
\end{aligned} \quad (10)
$$

Since the conditional likelihood $F(c|G)$ satisfies submodularity, the objective function $F(C|G)$, which is a nonnegative linear combination of $F(c|G)$, is also submodular. This function is provable to be optimized through an approximate algorithm like MultiTree [20] that obtains at least a constant fraction of $(1 - 1/e)$ of the optimal value achievable using k edges [15].

In reality, traffic congestion occurs frequently on core road segments. It means that there are massive potential casual links as candidates, which lead to a low speed of optimization. Common sense tells us that if r_i is far away from r_j, the potential edge (r_i, r_j) exists with just a tiny probability. Hence, as described in Algorithm 1, we propose an approximate algorithm as an extension of MultiTree with *spatial constraints* in line 2. G starts with an empty graph $\bar{K}$ (line 1), and it adds edges from the potential edge set P that maximize the marginal gain $\delta_{j,i}$ sequentially (line 15-16). That is to say, in each iteration i, we select the best edge e_i which satisfies

$$
e_i = \underset{e \in G_i \setminus G_{i-1}}{\arg\max} \; F(C|G_{i-1} \cup e) - F(C|G_{i-1}). \quad (11)
$$

We further study the time complexity of Algorithm 1. Utilizing the submodularity property of the objective function, each iteration stage is able to be finished in time $O(|C| * N^2)$ instead of super-exponential time, where N denotes the size of the largest cascade. Moreover, Algorithm 1 also allows two speeds-up: *localized updates* and *lazy evaluation* [4].

Algorithm 1 Approximate algorithm for *CasInf*

Input:
k : the number of edges in cascading patterns we infer.
C : the set of cascades obtained in a time span.
$\mathcal{D}$: a constant denoting the spatial constraint.
Output: G: the inferred cascading pattern.

1: $G \leftarrow \bar{K}$;
2: $P \leftarrow$ all pairs (j, i): $\exists c \in C$ with $t_j < t_i$ and $d_{j,i} < \mathcal{D}$
3: **while** $|G| \leq k$ **do**
4: **for** all $(j, i) \in P \setminus G$ **do**
5: $\delta_{j,i} = 0$;
6: **for** all $c : t_j < t_i$ **do**
7: $\varepsilon_g \leftarrow$ *environmental intensity inference*
8: $w_c(m, n) \leftarrow$ *weight of* (m, n) *in* $G \cup (j, i)$;
9: **for** all $m : t_m < t_i$ and $m \neq j$ **do**
10: $\delta_{c,j,i} = \delta_{c,j,i} + w_c(m, i)$;
11: **end for**
12: $\delta_{j,i} = \delta_{j,i} + log(\frac{\delta_{c,j,i} + w_c(j,i)}{\delta_{c,j,i}})$;
13: **end for**
14: **end for**
15: $(j^*, i^*) \leftarrow argmax_{(j,i) \notin G} \delta_{j,i}$;
16: $G \leftarrow G \cup (j^*, i^*)$;
17: **end while**

5 EVALUATION

In this section, we describe the real-world datasets and evaluate the effectiveness and efficiency of our approach.

5.1 Datasets

We use the following real-world datasets as detailed in Table 1:
Taxi Trajectories. We employ a GPS trajectory dataset generated by 32,670 taxicabs in Beijing from two time periods. GPS-equipped taxis can be regarded as mobile sensors probing the travel speed on road segments. Due to data sparsity, we fill in the missing values through *Context-Aware Tensor Decomposition* [23], i.e., we estimate the historical travel speed on each road at per 20 minutes.
Road networks. We employ a road network database from Bing Maps. Since people are more concerned about traffic conditions on the major roads instead of small lanes, we conduct our experiment on the road segments with level ranging from 0 to 2 (i.e., with high capacity). Additionally, road segments with CR more than 0.5 (i.e., there is often heavy traffic jam) are removed because the congestion on these road segments is more likely to be originated from the surrounding environment instead of its neighbors.
Meteorological data. We collect the fine-grained meteorological data, consisting of weather, temperature, humidity, pressure, wind speed from a public website every hour.
POIs. The POIs data is also obtained from Bing Maps, where each POI has a name, a category, and geo-coordinates.

Table 1: Details of the datasets

Datasets		Values
Taxi Trajectories	Number of Taxies	32,670
	Time Spans	03/01/2015-06/30/2015 11/01/2015-03/31/2016
Road Networks	Number of Segments	249,080
	Number of Nodes	186,266
	Total Length	25,638 km
POIs	Number of POIs	651,016
	Number of Categories	20
Meteorological Data	Number of Stations	16
	Time Spans	01/01/2015-12/31/2016

5.2 Baselines and Variants

We compare our approach with several baselines in the following two areas: 1) Diffusion network inference. 2) Congestion propagation detection. We first introduce the following three approaches for diffusion network inference:

- *NetInf* [4]: Since Equation 6 is intractable to optimize, this algorithm proposes an alternative target as follows: for each cascade, only the most likely propagation tree is considered, instead of considering all possible trees.
- *stNetInf*: stNetInf is a spatio-temporal implement of NetInf by integrating the *ITL* model and the *EMT* model.
- *MultiTree* [20]: It efficiently finds a solution to the network inference problem with provable sub-optimality guarantees by exploiting submodularity property.

Then we compare our methods with two approaches based on frequent subgraph mining in congestion propagation area as follows:

- *Frequency-Based Method (FBM)*: Inspired by the frequent pattern mining, we simply select k casual links with top-k occurring frequency during a number of consecutive days.
- *STC-DBN* [17]: The state-of-the-art approach for detecting congestion propagation constructs STCTrees based on the spatio-temporal information of identified congestions and apply an Apriori-based algorithm to detect the frequent substructures of the forests. It uses a Dynamic Bayesian Network (DBN) to simulate the congestion propagation process over road networks. In general, STC-DBN only considers the diffusion between connected segments. Here we apply a variant of it, which considers indirect influence.

To evaluate each component of our method, we also compare it with different variants of *CasInf*:

- *CasInf-gd*: We use *geometrical distance* as the spatial distance instead of road network distance in the *ITL* model.
- *CasInf-td*: We can easily set $\lambda = 0$ to implement this variant, where only the *temporal distance* is used. i.e., we ignore the spatial relationship between road segments.
- *CasInf-ne*: There is *no EMT model* in this variant. We can derive it by setting ε_g a constant ε for all urban grids.
- *CasInf-ni*: This variant does not consider the *indirect influence*. i.e., traffic congestion only spreads between adjacent road segments instead of neighboring road segments.

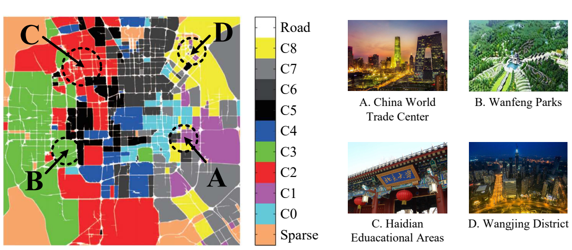


Figure 7: Functional regions discovered by DRoF [27] where C1, C3, C2, C8 denote developing commercial area, nature and parks, education and science area and emerging residential area respectively. We select the 4 specific areas A to D according to the aforementioned different function of regions.

5.3 Ground Truths & Parameter Settings

To justify the effectiveness of our method, we need to evaluate our approach based on some ground truths. Suppose a road segment gets congested at a specific time interval, we know which roads will get congested successively according to the inferred cascading patterns. For instance, there is a casual link (r_1, r_2) in the cascading pattern. It means that if r_1 gets congested at t_1, r_2 will get congested with relatively high risk over a period of time ranging from t_1 to $t_1 + \gamma$, where γ is the size of the time window. If a time span's observation satisfies the casual link, we say we have a *hit*. On the contrary, if such sequential occurrence of congestion does not exist in a time span's observed data, it reveals that our inference makes a non-hit (i.e., an error). Given the size of time window, we calculate the number of hits based on the observations of time spans from a series of M days, then use the metric *occurring probability* to validate the effectiveness of each casual link as follows:

$$prob_{j,i} = \frac{\#hit_{j,i}(\gamma)}{m_j}, \tag{12}$$

where $\#hit_{j,i}(\gamma)$ denotes the number of hits detected during M days specified by the window size γ, m_j means how many times r_j gets congested during these days, and $prob_{j,i}$ denotes the occurring probability of the casual link (r_j, r_i) in a cascading pattern $\hat{G}$, i.e., if we know a road segment gets congested, the accuracy of predicting next congestion occurs on its neighboring road segments in a given time window. To validate the correctness of the cascading pattern, we define the *score* (i.e., correctness) of a cascading pattern $\hat{G}$ with k casual links as follows:

$$score@k(\hat{G}) = \frac{\sum_{(j,i)\in E_{\hat{G}}} prob_{j,i}}{k}, \tag{13}$$

where $E_{\hat{G}}$ is the edge set of $\hat{G}$ and $k = |\hat{G}|$.

Particularly, to study the effectiveness and efficiency of our method changing over both the scale of road networks and the time spans, we evaluate our approach in the following four different sized areas of different functions shown in Figure 7 during *Morning rush hour* (8:00-10:00), *Noon lazy hour* (11:00-13:00) and *Evening rush hour* (17:00-19:00) respectively: A) *The China World Trade Center (CBD)*: A typical Central Business District with 769 road segments. B) *Wanfeng Parks (NPA)*: There are 659 road segments with relatively smooth traffic in this Nature and Parks Area. C) *Haidian District (ESA)*: An Education and Science Area with 1,574 road segments, including many colleges and IT companies. D) *Wangjing District*: We choose a small area composed of the residential places and business districts with 279 road segments to perform a case study.

Since our approach is scalable and parametric, we need to find a proper value for the number of inferred edges (i.e., k). As detailed in Equation 9, our target is to infer a relatively sparse cascading pattern over the road networks. We set k 500 to guarantee that the sparseness of the graph ranges from 0.1% to 1% in each area. i.e., we mainly use the metric *score*@500 for an experimental setting and further study on the impacts of different value of k in next section. In addition, we test different parameters for all baselines, finding the best setting for each.

5.4 Effectiveness Studies

5.4.1 Model Comparison. The upper half of Table 2 demonstrates the effectiveness of different methods in term of the metric *score*@500, where we set the parameter γ an hour. The result in Table 2 is the average value of the score in the three time spans detailed in the former section. The score seems small as compared to one road's congestion but relatively high with a 2-segment propagation. For example, both r_1 and r_2 have 30% chances getting congested in a day. The occurring probability of $r_1 \rightarrow r_2$ during the day is always less than 30% obviously. Moreover, we evaluate our method based on the road segments with relatively small congestion ratio, which increase the difficulty of the inference task.

Table 2 shows that our method outperforms the baselines on the overall score as well as the results in each tested area. More precisely, we divide the experimental analysis into two aspects as follows:

Network inference perspective: Our approach outperforms the temporal models including NetInf and MultiTree by 0.247 and 0.218 respectively. Besides, integrating the components including the *ITL* model and the *EMT* model, spatio-temporal approaches like stNetInf bring significant improvements (79.1%) beyond NetInf, which reveals that the insight of our method contributes not only to our model but also the general methods for network inference in spatio-temporal space. Only considering the most likely propagation trees, stNetInf performs slightly worse than *CasInf*.

Frequent subgraph perspective: *CasInf* shows great superiority against FBM because the latter one focuses on the occurrence times of frequent interactions but ignores the spatio-temporal dependency between two congested road segments. Specifically, *CasInf* achieves a better performance than STC-DBN by 0.172 (around 79.2% improvement). The reason is that DBN only captures the spatio-temporal dependency between two consecutive time frames, and it focuses on the frequent substructures of STCTrees.

We also get several interesting observations as follows. Firstly, our approach achieves the highest performance among various methods in both the large area (ESA) and the small areas (CBD and NPA), which verifies that *CasInf* is a flexible method that can be effectively applied to road networks with different scales. Secondly, we notice that *CasInf* presents higher performance in the areas with often busy and complex traffic like ESA since the congestion occurs more densely and has more spatio-temporal correlations than other smooth areas.

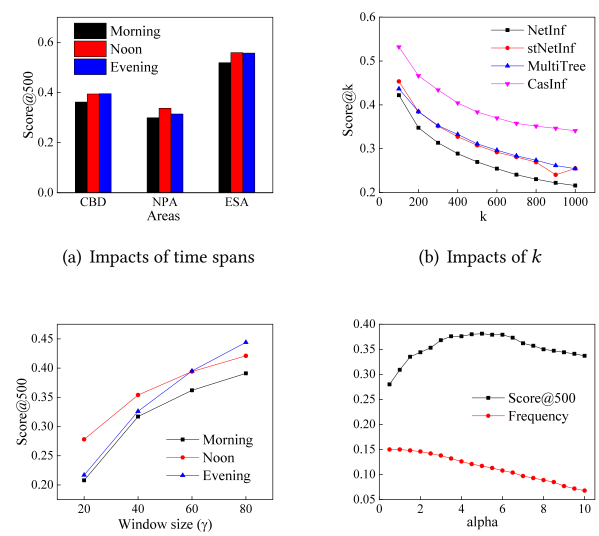
Table 2: Predictive performance comparison among various approaches over next 1 hour in the four areas.

Methods	CBD	NPA	ESA	Overall
NetInf	0.270	0.119	0.116	0.168
stNetInf	0.308	0.201	0.394	0.301
MultiTree	0.311	0.140	0.141	0.197
FBM	0.287	0.193	0.171	0.217
STC-DBN	0.307	0.198	0.225	0.243
CasInf-gd	0.359	0.258	0.488	0.368
CasInf-td	0.336	0.199	0.203	0.246
CasInf-ne	0.363	0.298	0.515	0.392
CasInf-ni	0.197	0.129	0.215	0.180
CasInf	**0.384**	**0.317**	**0.545**	**0.415**

5.4.2 Variant Comparison. The experimental results in the bottom half of Table 2 illustrate the strength of each component of our method on the accuracy metric. From this table, we can draw the following conclusion: (1) It is better to use road network distance rather than geometrical distance according to the improvement of performance between *CasInf* and *CasInf-gd*. It reveals that road network distance truly describes the spatial correlation between different road segments. (2) Lack of spatial distance, *CasInf-td* achieves a significantly worse performance than *CasInf* by 0.169, which illustrates the importance of the *ITL* model. That is because the diffusion of traffic congestion depends on the spatio-temporal properties, not only temporal properties. (3) Utilizing the *EMT* model, our approach achieves a slight increase (5%) in the score since the environmental factors are time- and location-varying in reality instead of constant in the conventional methods. (4) Our combination method shows a significantly better performance beyond *CasInf-ni* in term of accuracy, which reveals that the indirect influence plays an important role in the diffusion process.

5.4.3 Impacts of Parameters. We further study on the impacts of the model parameters including four parts as follows.

Different Time Spans. We discuss on the impacts of different time spans on *score*@500. Figure 8(a) shows the results obtained by *CasInf* in each area during the three time spans. There is a similar trend in the three areas changing over time of days, where *CasInf* shows the best performance at the noon time and the worst performance in the morning time. And it is easy to find that the results in the large area (ESA) with busy traffic during all the time spans are higher than that in other areas, which is also mentioned in Section 5.4.1.



(a) Impacts of time spans (b) Impacts of k

(c) Impacts of γ (d) Impacts of α

Figure 8: Impacts of the model parameters.

Different k values. We present the score-k-curve of the four methods including NetInf, stNetInf, Multitree and our approach in Figure 8(b). As the number of added edges k increases, the metric $score@500$ obtained by all the methods drop. i.e., the more edges we add into the cascading pattern, the lower score we will get. The reason is that our approach first selects the edges with relatively high (not always the highest) occurring probability during the running time. These edges always take an important role in the traffic congestion propagation process. In addition, it can be easily seen that our approach shows superiority over other methods from the beginning to the end of the curve.

Different Window Sizes. Figure 8(c) plots the performance of *CasInf* changing over γ. As the window size increases, the predictive performance increases particularly in the evening, since a higher γ allows a longer-term spreading process between a pair of roads, which causes an increase to the number of hits. Figure 8(c) also indicates that our method achieves better performance during the noon hours when the window size γ is lower than 1 hour. In section 5.4.1, we mainly choose $\gamma = 1$ hour to conduct the experiments.

Different Transmission Rates. Figure 8(d) shows the impact of transmission rate α on our approach based on the experiment during the morning hours of CBD. A large α may lose the diversity of individual transmission likelihoods between pairwise roads, but a very small α makes our approach focus on the frequency instead of our performance metric. That is why $score@500$ first increases and then decreases after a threshold like $\alpha = 5.5$. We also plot the average occurring frequency of each casual links during a series of days in Figure 8(d). The frequency seems small as compared to the congestion on a road segment but is relatively high with a pair of propagation. It can be seen that transmission rate α is a trade-off between the values of score and frequency.

5.5 Efficiency Studies

Using a single core of a server (with a 3.4GHz CPU and 16 GB memory) and the configurations introduced in Section 5.3, we further evaluate the efficiency of our approach against other methods in this part. The experimental results are shown in Figure 9, where we use the average running time per edge added as the metric.

We first compare our approach with the temporal models including MultiTree and NetInf. While the computational complexity of classical models in each iteration like MultiTree and NetInf is upper bounded by $O(|C| * N^2)$, which is the same as *CasInf*, our method runs much faster than classical ones. That is because there are less potential links constrained by the spatial distance $\mathcal{D}$ in algorithm 1. To our surprise, NetInf only considers the most likely propagation trees but runs slightly slower than MultiTree in our experiments. Besides the temporal models, we observe that the spatio-temporal models achieve significant improvement in efficiency study compared to temporal models, which indicates that our components can be practically applied to speed up classical models. Particularly, STCTree-based methods are not discussed in this part since it is not a scalable approach aiming at search k edges. This category of methods uses quadratic running time to construct trees and spend time on learning the whole network. The more potential links, the

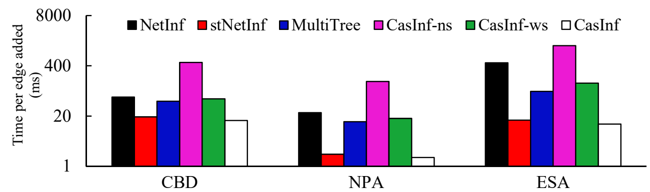


Figure 9: Efficiency of various approaches in the three areas.

more computational costs. This is why the running time per edge added in ESA is much slower than other small areas.

We further study on the running time of two additional variants including *CasInf-ns* (*CasInf without speeds-up* including localized updates and lazy evaluation) and *CasInf-ws* (*CasInf without spatial constraints*) to show the effectiveness of all these speeds-up components. Figure 9 verifies the advantage of our method integrating the the speeds-up components. Our method

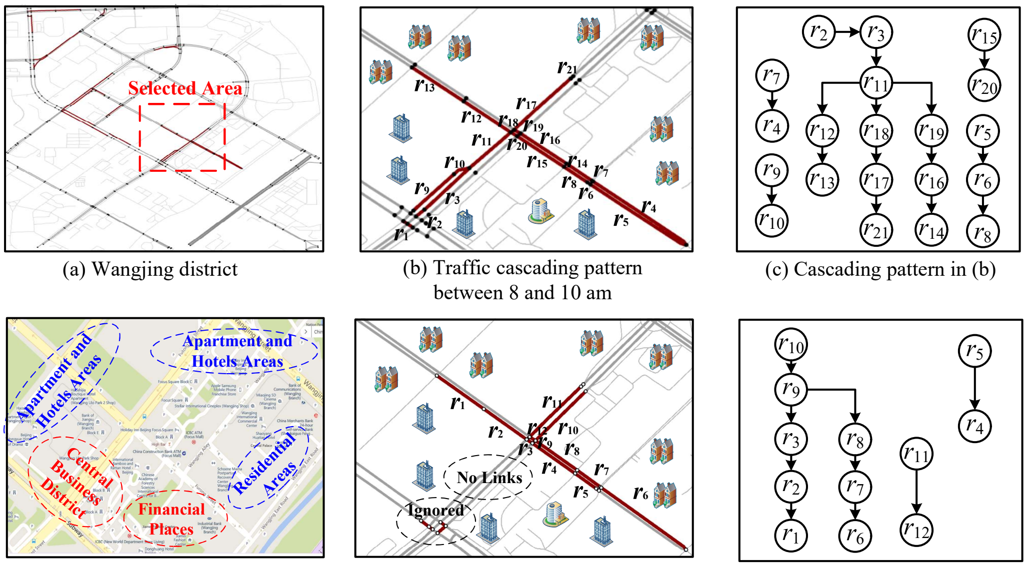
6 CASE STUDY

Now we perform a case study in Wangjing District. In Figure 10(a), there are many red lines denoting the road segments contained in the pattern during the morning rush hours. Additionally, we notice there exist several disjoint subnetworks in this figure. To explain the inferred pattern, we specifically select a partitioned area shown in Figure 10(d). According to the subfigure, it can be clearly seen that we are able to partition this area into *central business district* and *residential areas* due to the different functional POIs such as New World Department Store (a shopping mall) and Central Palace (an apartment). We will explain why and how the traffic congestion spreads in this area based on the facts as follows.

Morning rush hours. We present the cascading pattern composed of 21 road segments between 8 and 10 am in Figure 10(b-c). The latter figure clearly shows us that how the traffic congestion diffuses over this pattern. As it is known by us all that traffic conditions are closely related to human behaviors, people are commuting from the residential places to their workplaces in the morning, which results in a strong traffic flow into the surrounding of workplaces such as r_2 and r_9. With the time going on, we will see the traffic congestion occurs on r_3, r_{11} sequentially, following the pattern we inferred ($r_2 \rightarrow r_3 \rightarrow r_{11} \rightarrow r_{12} \rightarrow r_{13}$). Same goes for $r_5 \rightarrow r_6 \rightarrow r_8$, $r_{15} \rightarrow r_{20}$ and so on.

Evening rush hours. On the contrary, when reaching evening rush hours, people are more likely to go back home. Supporting this fact, our inferred pattern contains no congested roads (red lines) around the workplaces just like r_9, r_5 in Figure 10(b). i.e., the casual links between those road segments become less significant than it used to be in the morning. As depicted in Figure 10(f), our result is realistic and explainable since the traffic congestion spreads from residential places to workplaces.

Hence, the cascading pattern inferred by our approach is significant to understand the process of traffic congestion propagation according to the case study. Additionally, we can find the traffic bottlenecks (e.g., r_{11} in Figure 10(b-c)) through our inferred pattern, thereby improving urban planning.



(a) Wangjing district (b) Traffic cascading pattern between 8 and 10 am (c) Cascading pattern in (b)

(d) Selected area (e) Traffic cascading pattern between 7 and 9 pm (f) Cascading pattern in (e)

Figure 10: A case study in Wangjing District. Figure (a) shows the traffic cascading patterns comprised of many road segments (specified by red lines) during the morning rush hour. We select a small area of Wangjing shown in Figure (d) with different functional POIs. Figure (b) and (e) illustrate a subnetwork of the pattern in different time spans. Figure (c) and (f) are the visualization of the patterns. In the case study, the indirect influence is ignored because of the simplicity to visualization.

7 RELATED WORK

7.1 Urban Computing

Traffic congestion in the urban areas has been one of the main focuses for many years. M. Bando first presents a dynamical model of traffic congestion based on the equation of motion of each vehicle [1]. In reality, most instances of congestion mainly result from traffic bottlenecks, therefore many researchers study on such bottlenecks to alleviate traffic congestion [9, 16].

Besides finding the bottleneck, understanding the diffusion process of traffic congestion is of great importance to easing traffic condition and improving urban planning. Until now, there is a lack of research focusing on the traffic congestion propagation. W. Liu first studies causal interactions between different regions through causality trees [12]. These trees based on temporal and spatial information is able to be used for detecting the propagation of congestion as well [17]. However, there still remains several limitations in the previous studies: 1) *Data sparsity*. In general, previous studies are deployed in the datasets including small regions with sparse loop detectors under road segments during a short period of time. 2) *Multiple influential factors*. The formulations of previous works do not consider the influence of surrounding environment such as POIs and meteorology. 3) *Spatio-temporal correlations*. The previous work constructs propagation trees based on the traffic conditions at two consecutive time intervals, which ignores the long-term spreading process of traffic congestion. Additionally, the spatial correlations in diffusion are not well addressed.

Nowadays, the increasing availability of GPS-embedded taxicabs provides us with an unprecedented wealth to understand the properties of traffic flow, e.g., the travel speed. Y. Wang proposes a context-aware tensor decomposition approach to achieve a high accuracy in speed estimation [23]. Incorporating spatio-temporal features extracted from the real-world datasets, many data-driven approaches demonstrate their advantages in both flexible and extendibility in many ubiquitous applications [13, 14, 29]. Based on these studies, we propose a data-driven approach to solve aforementioned problems and infer the traffic cascading patterns efficiently.

7.2 Inferring Diffusion Networks

In the context of social networks, modeling how information spreads as cascades is of importance to analyzing misinformation and stopping the spread of virus [6, 18, 21]. How to infer the connectivity of a network based on the diffusion traces has been extensively studied these years [10, 22]. Assuming the nodes are only activated by another single node, NetInf aims to maximize the likelihood of observed data. MultiTree algorithm, which considers all possible propagation trees, directly solve this problem with provable near-optimal guarantees based on submodularity property. Extending NetInf for generality, NetRate [19] infers the pairwise transmission rates by solving a convex maximum likelihood problem. Above methods assume that the hypothetical network is static while InfoPath [5] infer a time-varying network changing over time.However, to the best of our knowledge, based on network inference, our approach is the first work incorporating spatio-temporal factors to infer the spreading of traffic congestion over road networks, which can also be applied into general spatio-temporal applications.

7.3 Frequent Subgraph Mining

Frequent Subgraph Mining (FSM) is a fundamental building block of graph mining. The objective of FSM is to extract all the frequent subgraphs that appear more number of times than a given threshold in a graph dataset. Graph datasets can be divided into two categories: datasets consisting of many stand-alone (small) graphs, called transactional setting, and datasets with a single large graph. The representative FSM methods with a transactional setting include the FSG [8], gSpan [25] and CloseGraph [26]. To discover congestion propagation pattern in spatio-temporal data, Nguyen first constructs causality trees based on the spatio-temporal datasets and then apply Apriori Subtree(a variant of FSG) to detect the frequent subtrees. But in reality, the underlying patterns may not only contain the frequent substructures of observed correlation graphs. The causal links with a low frequency but high dependency also play an important role in the cascading patterns.

8 CONCLUSION AND FUTURE WORKS

In this paper, we propose a data-driven approach *CasInf* to infer the traffic cascading patterns from multiple spatio-temporal datasets, which is comprised of the *ITL* model, the *EMT* model, and the cascading pattern construction algorithm. The first two models are applied for the inference of three-fold influences existing in traffic propagation process. Combining these influences into multiple propagation trees over a graph, the last model maximizes the likelihood of the observed traffic data through an approximate algorithm with provable near-optimal guarantees. We evaluate our approach based on four real-world datasets including taxi trajectories data, road networks data, POIs data, and meteorological data in Beijing. The experimental results show that our method outperforms the baselines in terms of predictive score. Our method gets an overall score around 0.415, which outperforms the temporal models including NetInf and MultiTree. Besides, our method shows superiority against the state-of-the-art methods STC-DBN, which is based on frequent subgraph mining. We present a running time study as well and the result experimentally verifies the efficiency of our approach. The code has been released for research use.

In the future, we plan to infer the transmission rate a for each casual link based on convex optimization instead of regarding it as a constant. Additionally, we would like to extend our study into general spatio-temporal problems, such as inferring cascading patterns of human flows from an area to another area.

ACKNOWLEDGEMENTS

The work was supported by the National Natural Science Foundation of China (Grant No. 61672399, No. U1609217 and No. 61502375).

REFERENCES

[1] Masako Bando, Katsuya Hasebe, Akihiro Nakayama, Akihiro Shibata, and Yuki Sugiyama. 1995. Dynamical model of traffic congestion and numerical simulation. *Physical review E* 51, 2 (1995), 1035.
[2] Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 7039 (2005), 207–211.
[3] Francis Buekenhout and Monique Parker. 1998. The number of nets of the regular convex polytopes in dimension? 4. *Discrete mathematics* 186, 1-3 (1998), 69–94.
[4] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. 2010. Inferring Networks of Diffusion and Influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*. ACM, New York, NY, USA, 1019–1028. https://doi.org/10.1145/1835804.1835933
[5] Manuel Gomez-Rodriguez, Jure Leskovec, and Bernhard Schölkopf. 2013. Modeling Information Propagation with Survival Theory.. In *ICML (3)*. 666–674.
[6] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. 2013. Information diffusion in online social networks: A survey. *ACM SIGMOD Record* 42, 2 (2013), 17–28.
[7] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 137–146.
[8] Michihiro Kuramochi and George Karypis. 2001. Frequent subgraph discovery. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE, 313–320.
[9] Wei-Hsun Lee, Shian-Shyong Tseng, Jin-Lih Shieh, and Hsiao-Han Chen. 2011. Discovering traffic bottlenecks in an urban network by spatiotemporal data mining on location-based services. *IEEE Transactions on Intelligent Transportation Systems* 12, 4 (2011), 1047–1056.
[10] Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. 2010. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research* 11, Feb (2010), 985–1042.
[11] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. 2007. Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining*. SIAM, 551–556.
[12] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. 2011. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1010–1018.
[13] Ye Liu, Yuxuan Liang, Shuming Liu, David S Rosenblum, and Yu Zheng. 2016. Predicting urban water quality with ubiquitous data. *arXiv preprint arXiv:1610.09462* (2016).
[14] Ye Liu, Yu Zheng, Yuxuan Liang, Shuming Liu, and David S Rosenblum. 2016. Urban water quality prediction based on multi-task multi-view learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*.
[15] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functionsąłI. *Mathematical Programming* 14, 1 (1978), 265–294.
[16] Gordon Frank Newell. 1998. A moving bottleneck. *Transportation Research Part B: Methodological* 32, 8 (1998), 531–537.
[17] Hoang Nguyen, Wei Liu, and Fang Chen. 2016. Discovering congestion propagation patterns in spatio-temporal traffic data. *IEEE Transactions on Big Data* (2016).
[18] Qiang Qu, Siyuan Liu, Feida Zhu, and Christian S Jensen. 2016. Efficient Online Summarization of Large-Scale Dynamic Networks. *IEEE Transactions on Knowledge and Data Engineering* 28, 12 (2016), 3231–3245.
[19] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. 2011. Uncovering the temporal dynamics of diffusion networks. *arXiv preprint arXiv:1105.0697* (2011).
[20] Manuel Gomez Rodriguez and Bernhard Schölkopf. 2012. Submodular inference of diffusion networks from multiple trees. *arXiv preprint arXiv:1205.1671* (2012).
[21] Everett M Rogers. 2010. *Diffusion of innovations*. Simon and Schuster.
[22] Tristan Mark Snowsill, Nick Fyson, Tijl De Bie, and Nello Cristianini. 2011. Refining causality: who copied from whom?. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 466–474.
[23] Yilun Wang, Yu Zheng, and Yexiang Xue. 2014. Travel time estimation of a path using sparse trajectories. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 25–34.
[24] Fei Wu, Hongjian Wang, and Zhenhui Li. 2016. Interpreting traffic dynamics using ubiquitous urban data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 69.
[25] Xifeng Yan and Jiawei Han. 2002. gspan: Graph-based substructure pattern mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, 721–724.
[26] Xifeng Yan and Jiawei Han. 2003. CloseGraph: mining closed frequent graph patterns. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 286–295.
[27] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 186–194.
[28] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3 (2014), 38.
[29] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. 2013. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1436–1444.