Spatial-Temporal Outlier Sensing over Trajectory Data Streams

Xian Wu, Chao Huang and Lin Chen

Chinese Academy of Sciences
Beijing, China

Abstract

The increasing capability to track moving vehicles in city roads enable people to probe the dynamics of a city. In this paper, we address the problem of detecting outliers and anomalies sources with trajectory data. Unlike existing anomaly detection methods, both spatial and temporal information are considered to find the potential outliers. We identify anomalies according to not only the individual traffic regular patterns, but also the consistent traffic behaviors with adjacent road segments on road network. To analyze the major sources of anomalies, we then describe the anomalies diffusion process on the basis of information diffusion model. Furthermore, in order to make detection results not limited to only large scale events, the granularity of our detected traffic anomaly is on the level of road segments instead of spatial regions. Finally, experiments on a very large volume of real taxi trajectories in an urban road network show that the proposed approaches outperform the recent state-of-the-art algorithms.

Keywords: *Spatial-Temporal Outliers, Data Streams.*

1. Introduction

With the increasing popularity of GPS and Wi-Fi devices, the tracking mobile movements in major metropolitan cities has become a reality. Thus the very practical and important problem for detecting outlier over trajectories data has attracted much attention in literatures [1, 2, 3, 4, 5, 6]. Unusual patterns of moving objects trajectories generally reflect abnormal traffic patterns on road networks: popular sporting events draw crowds, holidays create disruptions, protests may result in road closures, etc. Therefore, the detection of outliers/anomalies from over trajectory data stream can help in sensing abnormal events and plan for their impact on the smooth flow of traffic.

In this paper, we propose a novel framework to detect outlier patterns and anomalies sources, which could be caused by traffic accidents and controls, large-scale business promotions and celebrations. Towards this end, firstly we mine the GPS trajectory data stream to detect significant traffic volume changes. A definition of an outlier in [7] motives us to detect outlier patterns not only utilizing the individual spatial-temporal information but also the agglomerate spatial-temporal information of adjacent road segments. Secondly, we piece together the mined outliers and infer sources which may have caused the anomalies to occur.

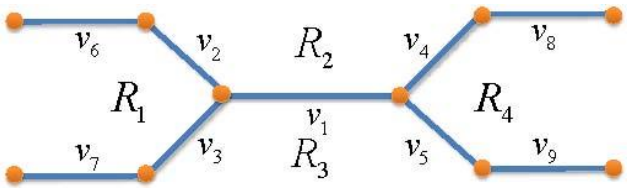

Fig 1. Concrete Example.

To give a concrete example as shown in Figure 1. There are nine road segments (i.e. V_1, V_2,…, V_9) in this road network. Here suppose that a traffic accident happens in V_1. In most cases, since the spatial correlations between among road segments, anomalies in one road segment may trigger other anomalies in its adjacent road segments. Specifically, the abnormal traffic patterns of V_1 may trigger anomalies of the road segments adjacent to V_1 (i.e. V_2, V_3, V_4 and V_5). Additionally, the anomalies of V_2 may have a certain impact on the traffic pattern of V_6 (which is farther away from V_1).

Our system uses a novel methodology to detect outlier and has the following advantages over the existing methods. First, it provides a comprehensive view of the anomalies, showing the anomalous road segment as well as the relationship between these road segments. For instance, the individual traffic volume-based method may not even be able to detect some extreme cases, where the traffic volume does not change significantly on each road segment. Second, the granularity of our detected traffic anomaly is on the level of road segments instead of spatial regions. The regions based methods may suffer from a boundary problem. Specifically, if two taxis travel via V_1, due to the imprecise recording of GPS devices, locations of one taxi may be recorded in Region R_2, but locations of another taxi may be recorded in Region R_3. This will lead to the inaccuracy of map-matching.

The contributions we make in this paper are as follows:

(1). We propose a deviation based method to detect traffic anomalies considering not only the individual traffic volume change, but also the affected spatial regions and relationship between individual road segments. By doing so, we can detect anomalous road segments that do not disrupt the traffic volume, but are not consistent with volume changes of adjacent road segments.
(2). We propose a diffusion based method to discover the major sources of outliers, by taking advantage of diffusion properties of anomalies in the road network. By doing so, we correlate disruptions in the traffic patterns with their sources, i.e. the outliers in this road segment causing the anomalies of corresponding region.
(3). We propose a diffusion based method to discover the major sources of outliers, by taking advantage of diffusion properties of anomalies in the road network. By doing so, we correlate disruptions in the traffic patterns with their sources, i.e. the outliers in this road segment causing the anomalies of corresponding region.

The rest of the paper is structured as follows. Section 2 elaborates on the preliminaries and the overview of system. Section 3 presents our offline mining approach. In Section 4, we detail our outlier and source detection method. Section 5 shows our experimental results and analysis. Section 6 describes the related work. Finally, conclusions are drawn in Section 7.

2. Preliminaries and System Overview

In this section, we introduce our notations, definitions and give an overview of the system.

2.1 Preliminaries

DEFINITION 1 (Trajectory) A trajectory Tr is a trace created by moving objects in geographical space. A Tr is represented by a set of time-ordered points, e.g., $Tr: p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow \ldots \rightarrow p_n$, where each point consists of a geospatial coordinate set and a timestamp, i.e. $p = (longitude, latitude, timestamp)$

In many situations, the relationships between road segments are directed, especially there are many one-way roads in the actual road network. The vehicles can moving from road segment v_i to v_j directly. But vehicles may be prohibited passing through v_i to v_j due to some traffic regulations. Motivated by this observation, we consider the directed relationships between road segments in road network.

DEFINITION 2 (Directed Adjacency.) Given a trajectory Tr, there exists a directed edge e (namely directed adjacent relationship) between road segment v_i and v_j, if there exists adjacent points, namely (p_1, t_1) and (p_2, t_2) (where $t_1 < t_2$) such that p_1 is in v_i and p_2 is in v_j, and v_i is not the same to v_j. Then we define that v_j is directed adjacent to v_i, and v_i is a directed spatial neighbor of v_j. Given an road segment v, we denote the set of all directed neighbors of v by $N(v)$.

Given a road segment v and a time interval $\Delta t = [t, t']$, an object is said to pass the road segment v in Δt, if there exists a timestamp $t \in [t, t']$ such that its estimated position at timestamp t is along the road segment v.

DEFINITION 3 (Traffic.) The traffic of road segment v in the time interval Δt, denoted by $\psi(v, \Delta t)$, is defined to be the total number of objects passing v in the time interval Δt.

Information Diffusion Model. The Information Diffusion model *Independent Cascade Model* is originally proposed by Lopez-Pintado [8], and is the common dynamic model in information diffusion [9, 10]. In the diffusing process, information always flows from a position with high energy to a position with low energy.

In this paper, we model diffusion of anomalies as processes of information energy diffusion. Actually, we observe that the process of anomalies influencing other road segments is very similar to the information diffusion phenomenon. In a road network, the abnormal road sources act as information sources, have a very high amount of energy. These road segments start to diffuse their anomalies to the early other road segments which "directed adjacent" to them, then the late majority.

The information flows throughout a geometric manifold with initial conditions can be described by the following second order differential equation:

$$\left\{ \begin{array}{l} \dfrac{\partial f(x,t)}{\partial t} - \Delta f(x,t) = 0 \\ f(x,0) = f_0(x) \end{array} \right\} \quad (1)$$

Where $f(x, t)$ is the information energy at location x at time t, beginning with an initial distribution $f_0(x)$ at time zero, and Δf is the *Laplace-Beltrami operator* on a function f.

The model has an important parameter called *diffusion speed* λ. This diffusion coefficient plays an important role in the anomalies diffusion process. If it has a high value, anomalies will diffuse very quickly. Otherwise, anomalies will diffuse slowly. However, the Independent Cascade Model model does not take edge weight into consideration. Actually, a road network is a directed weighted graph, and the weight should play an important role in anomalies diffusion. Intuitively, the more contacts between two road segments, the more likely the influence will happen. We extend the diffusion model to accommodate the edge weight. The influence rate raises as the weight between road segments increases.

DEFINITION 4 (Influence Degree.) The influence degree $\omega_{ij}(\Delta t)$ from road segment v_i to v_j in time interval Δt, is defined as: $\omega_{ij}(\Delta t) = Pout_{ij}(\Delta t) \times Pin_{ij}(\Delta t)$.

(1). $Pout_{ij}(\Delta t)$: the proportion of objects going through v_i to v_j among all objects moving out of v_i.

(2). $Pin_{ij}(\Delta t)$: the proportion of objects going through v_i to v_j among all objects moving into v_j.

Road Segment Graph. We extract the road traffic networks from the trajectory data and model it as a directed weighted graph: a road segment corresponds to a node; a directed edge from node v_i to v_j is established, if there exit directed adjacent relationship from v_i to v_j, with the corresponding influence degree as the weight of the edge. We denoted the graph as $G = (V, E, W)$, where V, E and W represent nodes, edges, and weights, respectively.

2.2 System Overview

The architecture of our system consists of three parts: offline mining, outlier detection, and outlier source analysis.

(1) Offline Mining. This step accumulates historical mobility data (e.g., GPS trajectories from vehicles) into a trajectory database and builds an index between road segments and the trajectories traversing them in order to enable online outlier detection.

(2) Online Outlier Detection. Outlier detection is an online inference step based on the recently received GPS trajectories of vehicles and the behavioral knowledge we obtained from offline mining. First, our system maps the received GPS trajectories of vehicles onto a road network using a map-matching algorithm presented in [11]. One copy of these processed trajectories is sent to the trajectory database for offline mining. Another copy is used for real-time traffic anomaly analysis. Similar to offline mining, we analyze the current vehicle flow for each road segment. By comparing the real-time information with our historical traffic knowledge, our system selects road segments, each of which real traffic with a certain deviation from its normal pattern.

(3) Outlier Source Analysis. The outlier analysis step aims to analyze the anomaly and find the major anomaly sources. We extract the information from the recent GPS trajectories of vehicles and the ordinary traffic behaviors. We then mine the major anomaly sources that their ``observed" outlier values (based on the deviation from expected traffic) with a certain deviation from their ``expected" outlier values (based on the diffusion of anomalies) in a given time interval.

3. Offline Mining

3.1 Building Road Segment Graph

In our study, we perform the map-matching operation to locate “recorded” trajectories (in the two-dimensional space) to the road segment graph. Performing this operation is beneficial since the boundary problem described in Section I does not appear after this operation is executed. In detail, we build a graph of road segments according to the following three steps.

(1). **Time Bin Partitioning**. In this paper, in order to define *outliers* more precisely, we divide all timestamps based on *30-minute time slots* and *day-of-the-week* as follows. Firstly, we divide a single 24-hour day 48 time slots, each of time slot lasts for 30 minutes. This is our default value in our experiments. It can be changed accordingly to the requirement of the time slot granularity in other applications. Each time slot is called a time bin. Let Γ be a set of all possible time bins. Given a time interval Δt, we define a mapping function $M(\cdot)$ which takes Δt as input and returns the time bin $\tau \in \Gamma$ that Δt belongs to.

Secondly, we partition all these time bins into a number of groups based on *days*. Since there are seven days of the week (i.e., Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday),

we create seven groups where each group denotes a day of the week, and each group contains a set of time bins which belong to this group. Each group is called a day-of-the-week group. In this partitioning, there are 48*7=336 possible time bins.

(2). **Traffic Estimating**. For each road segment $v \in V$ and each possible time bin $\tau \in \Gamma$, we can transfer each trajectory into traffic samples of the road segment v for this time bin τ, denoted by $\psi(v, \tau)$ (refer to Definition 2).

(3) **Influence Degree Estimating**. If there is a directed adjacent relationship between two roads segments, we connect the two roads segments with a directed edge (refer to Definition 3). Given a time bin τ, a directed edge $e_{ij} =< v_i, v_j >$ is associated with a weight ω_{ij}, namely the *influence degree* (refer to Definition 4). This weight is calculated as:

$$\omega_{ij} = Pout_{ij}(\tau) \cdot Pin_{ij}(\tau) = \frac{\psi(v_{ij}, \tau)}{\psi_{out}(v_i, \tau)} \cdot \frac{\psi(v_{ij}, \tau)}{\psi_{in}(v_i, \tau)} \quad (2)$$

(i). $\psi(v_{ij}, \tau)$: Total number of objects moving from v_i to v_j in the time bin τ.
(ii). $\psi_{out}(v_{ij}, \tau)$: Total number of objects moving out v_i in the time bin τ.
(iii). $\psi_{in}(v_{ij}, \tau)$: Total number of objects moving into v_j in the time bin τ.

During the process of offline mining, we build index structures between the trajectories and road segments. These index structures is built offline, but will be updated online as new trajectories are received.

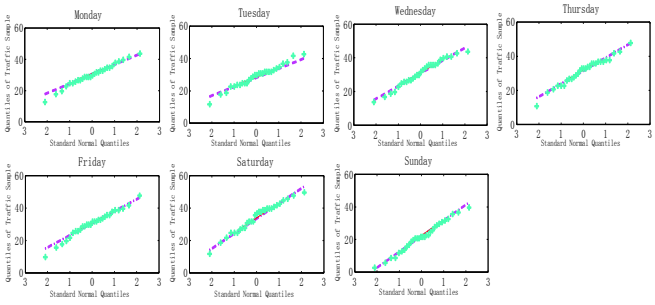

Fig. 2. Traffic samples of a road segment for a time bin versus a normal distribution.

3.2 Modeling Traffic Behavior

We analyze our real taxi trajectory dataset collected in Shenzhen from January 1, 2013 to June 30, 2013. Consider the day-of-the-week partitioning. Given an arbitrary road segment v. For each possible time bin τ representing a day of the week (e.g., Monday) and a 30-minute time slot (e.g., 11:00am to 11:30am). Figure 2 shows all possible traffic samples of v for τ versus a normal distribution. In the figure, the x-axis corresponds to the normal theoretical quantiles and the y-axis corresponds to the data quantiles. Each point (x, y) in the figure means that one of the quantiles from the normal distribution is x and the same quantile from the data (i.e., traffic samples) is y. Based on this figure, we observe that the points in this figure lie on a single line, which means that the distribution on traffic samples of v for this time bin τ, denoted by $D(\psi(v, \tau))$, is similar to a normal distribution.

In summary, we can model $D(\psi(v, \tau))$ as a normal distribution. Thus, based on all possible traffic values of v for τ, we can calculate the mean and the standard deviation, denoted by $\mu(v, \tau)$ and $\sigma(v, \tau)$. The "regular" traffic based on $D(\psi(v, \tau))$ is exactly equal to the mean of this distribution (i.e., $\mu(v, \tau)$). Motivated by this observation, we consider the "regular" traffic from the perspective of stable traffic distribution as the first basis for outlier detection.

Furthermore, after analyzing the distribution of traffic samples on adjacent road segments, we observe that in most cases, anomalies in one road segment can trigger other anomalies in adjacent roads segments. Consider the example in Figure 1, these adjacent roads segments (i.e. V_2, V_3, V_4 and V_5). Figure 3 shows that these road segments have *stable trend* of similar traffic from 00:00 to 16:00. About at 17:00, however, they all show an increase in traffic while V_3 remains the same. Road segment V_3 now become an outlier. Recall the outlier definition regarding "inconsistency". Motivated by this, we consider the "regular" traffic from the perspective of stable traffic trend as the second basis for outlier detection.

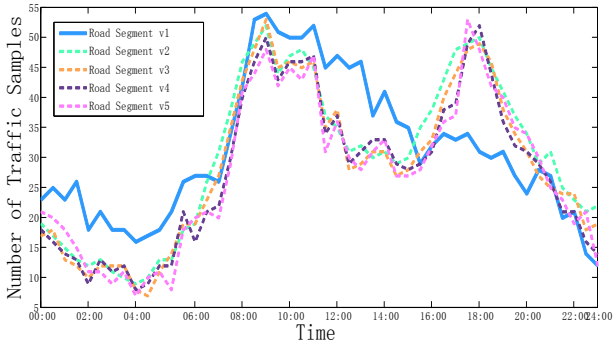

Fig 3. Historical traffic samples on serval road segments.

DEFINITION 5 (Historical Similarity Vector.) Every road segment v_i maintains a vector $\vec{\theta}_i$ to record historical similarities to its neighboring road segments. The length of this vector is $Num = |N(v_i)|$. Every j^{th} value, $\theta_{i,j} \in \Re$ records the historical similarity between road segment v_i and road segment v_j. A large $\theta_{i,j}$ indicates high historical similarity and vice-versa. Initially, all vector values are set to 0.

In every time bin, each value $\theta_{i,j}$ is adjusted up or down depending on the similarity between road segment v_i and v_j at that instant.

DEFINITION 6 (Instantaneous Similarity.) Let $d(v_i, v_j)$ be a distance function defined on the feature space (only consider traffic feature in this paper). Then, road segments v_i and v_j are similar with respect to an instant in time and the particular feature values at the current time bin (i.e., instantaneously similar) if $d(v_i, v_j) < \beta$, where β is a threshold to determine similarity. (According to the analysis of dataset, we set β to 10 in our experiment).

The temporal neighborhood vector values are updated differently based on whether instantaneous similarity holds true or false. Here, let $d(v_i, v_j)$ be the L-∞ distance (i.e., maximum absolute difference between feature values).

With instantaneous similarity defined, the next step is then updating the historical similarity vector temporally, where time is essentially a sequence of time bins. As mentioned briefly before, the amount of change in $\vec{\theta}_i$ in each time bin is the measure of "outlier-ness" for road segment v_i. Thus, it is critically important to have proper update rules. We use the the following intuition based on the existing similarity values.

(i). If two adjacent road segments are historically similar, then a new instantaneous similarity can be noted lightly.

(ii). If two adjacent road segments are historically similar, then a new instantaneous dissimilarity should be noted heavily.

(iii). If two adjacent road segments are not historically similar, then a new instantaneous similarity should be noted heavily.

(iv). If two adjacent road segments are not historically similar, then a new instantaneous dissimilarity can be noted lightly.

It turns out that exponential functions conveniently capture the above intuition. Let the period of update be a time bin and consider the similarity value $\theta_{i,j}^{\tau-1}$ between road segments v_i and v_j in the time bin $\tau - 1$. If these two road segments are instantaneously similar on the next time bin τ, the *gain* is defined as:

$$gain(v_i, v_j, \tau) = \exp(-\theta_{i,j}^{\tau-1}) \quad (3)$$

For progressively larger $\theta_{i,j}^{\tau-1}$ values, the gain naturally becomes smaller. The update function for $\theta_{i,j}^{\tau}$ is then $\theta_{i,j}^{\tau} = \theta_{i,j}^{\tau-1} + gain(v_i, v_j, \tau)$. The same equation applies to *loss* as well. If road segments v_i and v_j are instantaneously dissimilar in the time bin τ the loss is defined as:

$$loss(v_i, v_j, \tau) = \exp(\theta_{i,j}^{\tau-1}) \quad (4)$$

For progressively larger $\theta_{i,j}^{\tau-1}$ values, the loss becomes larger exponentially. The update function for $\theta_{i,j}^{\tau} = \theta_{i,j}^{\tau-1} - loss(v_i, v_j, \tau)$.

4. Outlier Detection

Problem Statement: Given a road segment network and a time interval Δt, firstly we aim at detecting all road segments in V with traffic pattern highly deviated from regular traffic pattern, denoted by $S(\Delta t)$. Secondly, we discover the road segments in $S(\Delta t)$ which are the major "sources" of the outliers. In the following sections, we will discuss the two major steps, anomalous road segment selection and anomalies sources detection.

4.1 Anomalous Road Segment Selection

In this section, we propose a deviation based detection method to find the anomalous road segments.

Firstly, we consider the deviation from the *stable distribution of traffic*. Since the distribution of traffic samples can be modeled as a normal distribution. Given a road segment v_i and a time bin τ, the first category of outlier value of road segment v_i in this time bin, denoted by $OV_{SD}(v_i, \tau)$, is defined as:

$$OV_{SD}(v_i, \tau) = 2 \cdot \frac{1}{1 + \exp\{\frac{|\psi(v_i, \tau) - \mu(v_i, \tau)|}{\sigma(v_i, \tau)}\}} - 1 \quad (5)$$

Since the expression above is a sigmoid function ranging from 0.5 to 1 for any non-negative real number (v_i, τ), $OV_{SD}(v_i, \tau)$ ranges from 0 to 1. If the difference between the real traffic (i.e., $\psi(v_i, \tau)$) and its regular traffic (ie., $\mu(v_i, \tau)$ is larger, then $OV_{SD}(v_i, \tau)$ will be larger. Otherwise, it will be smaller. For example, if the difference between $\psi(v_i, \tau)$ and $\mu(v_i, \tau)$ is equal to 3σ, then the $OV_{SD}(v_i, \tau)$ is equal to 0.952. Thus, Equation (5) captures the extent of the deviation of the traffic from the stable historical distribution. Nevertheless, it may potentially recall some outliers caused by the periodic fluctuation of traffic. In fact this may be a normal phenomenon, such as some major road segments during rush hour (i.e., the two time bins: 08:00 to 09:00 on every Monday).

Secondly, consider the deviation from the *stable trend of traffic*. As mentioned before, the historical similarity values are recorded in the historical similarity vector at each road segment. Intuitively, the most drastic or abnormal changes are ones that differ the most from historical values. Furthermore, the more stable the historical values are previously, the more the change should contribute to the outlier value. Fortunately, this intuition is easily captured in the *gain* and *loss* equations in the previous section. Because $\theta_{ij}^{\tau-1}$ is in the exponent of Equations (3) and (4), the large gain and loss will come from previously stable trends (either similar or dissimilar). Therefore, the outlier value of road segment v is equal to the sum of gain and loss. Given a road segment v_i and a time bin τ, the second category of outlier value of road segment v_i in this time bin, denoted by $OV_{ST}(v_i, \tau)$, is defined as:

$$OV_{ST}(v_i, \tau) = 2 \cdot \frac{1}{1 + \exp\{\sum_{j=1, j \neq i}^{Num} \| \theta_{ij}^{\tau} - \theta_{ij}^{\tau-1} \|\}} - 1 \quad (6)$$

Similarity, $OV_{ST}(v_i, \tau)$ also ranges from 0 to 1. For a given road segment, Equation (6) is calculated with respect to all its adjacent road segments in the road segments graph. However, it may potentially hide the outliers that affect a region of road segments. For example, suppose there is a very serious traffic accident that affects many road segments in a local area. Only take into account the deviation from stable trend of traffic may lead to a low outlier value and miss it altogether.

Fortunately, the false positives returned by first approach (deviation from stable distribution of traffic) will be assigned a low outlier value by the second approach (deviation from stable trend of traffic). Because periodic outliers always keep a stable trend of traffic. Additionally, the false negatives omitted by the second approach always deviates from their stable traffic distribution seriously. Motivated by this observation, we utilize the linear combination of these two categories of deviation. Thus, given a road segment $v \in V$ and a time bin $\tau \in \Gamma$, the outlier value of v in τ, denoted by $OV(v, \tau)$, is defined as:

$$OV(v, \tau) = OV_{SD}(v, \tau) + OV_{ST}(v, \tau) \quad (7)$$

Now, we introduce a user parameter called the *outlier threshold* ξ, which is a non-negative real number from 0 to 1, to determine whether a road segment v in a time bin τ is an outlier. Given a road segment $v \in V$ and a time bin $\tau \in \Gamma$, the road segment v in this time bin is said to be an *outlier* if $OV(v, \tau)$ is at least ξ.

4.2 Outlier Analysis

In this section, we propose a diffusion based detection method to find the major anomaly sources of the outliers found in the above section.

As mentioned before, we model a road network as a directed graph, and each road segment in the road network is defined as a node on this graph. The relationships between road segments are represented by edges that connect nodes. We propose our anomalies diffusion model on road segments graph

with prior knowledge of influence degree. In this directed graph, each node $v_j \in V$ is associated with a value of *information energy* (i.e., the degree of anomalies). We denote the amount of the energy for v_j at a timestamp t by $f(v_j, t)$ which is a non-negative real number. Each road segment v_j receives an amount of energy from its directed spatial neighbor v_i. During the time interval Δt whose start timestamp is equal to t. This amount is denoted by $\phi(v_j, v_i, t, \Delta t)$. It should be proportion to the duration of the time interval (i.e., $|\Delta t|$) and the difference in the degree of anomalies between v_i and v_j (i.e., $f(v_i, t) - f(v_j, t)$). Moreover, the anomalies flows from node v_i to v_j through the pipe $< v_i, v_{1j} > \in E$ that connects node i and j. $W = \{\omega_{ij} \mid$ there is an directed edge from v_i to $v_j\}$ is the set of all influence degrees between road segments. Based on this consideration, we assume that

$$\phi(v_j, v_i, t, \Delta t) = \lambda \cdot |\Delta t| \cdot (f(v_i, t) - f(v_j, t)) \cdot \omega_{ij} \quad (8)$$

Where λ is the diffusion speed with the value of non-negative. Thus, the total amount of anomalies that a road segments v_j receives between timestamp t and $t + \Delta t$, denoted by $\Phi(v_j, [t, t + \Delta t])$, is equal to the sum of the amount of anomalies that v_j receives from all of its directed spatial neighbors, which is defined as:

$$\Phi(v_j, [t + \Delta t]) = \lambda \cdot |\Delta t| \cdot \sum_{v_i \in N(v_j)} (f(v_i, t) - f(v_j, t)) \cdot \omega_{ij} \quad (9)$$

Furthermore, note that in our problem, since anomalies (e.g., road segments with car accidents) will be recovered by some external mechanisms (e.g., towing away cars in accidents). In our diffusion model, we model this external mechanism by an exponential decay process in which the amount of anomalies of a road segment decreases with time exponentially. Thus, the total amount of anomalies of v_j which can be kept at timestamp $t + \Delta t$ due to its original anomalies at timestamp t, denoted by $\Upsilon(v_j, [t, t + \Delta t])$, is defined as:

$$\Upsilon(v_j, [t, t + \Delta t]) = \exp(-\eta \mid \Delta t \mid) \cdot f(v_j, t) \quad (10)$$

Where η is the *decay factor* of the exponential decay process. Therefore, the total amount of anomalies of v_j at timestamp $t + \Delta t$ (i.e., $f(v_j, t + \Delta t)$) is defined as follows.

$$f(v_j, t + \Delta t) = \Phi(v_j, [t + \Delta t]) + \Upsilon(v_j, [t, t + \Delta t]) \quad (11)$$

That is,

$$f(v_j, t + \Delta t) = \lambda \cdot |\Delta t| \cdot \sum_{v_i \in N(v_j)} (f(v_i, t) - f(v_j, t)) \cdot \omega_{ij} + \exp(-\eta \mid \Delta t \mid) \cdot f(v_j, t) \quad (12)$$

We can rewrite Equation (12) as follows.

$$\frac{f(v_j, t + \Delta t) - \exp(-\eta \mid \Delta t \mid) \cdot f(v_j, t)}{\Delta t} = \lambda \cdot \sum_{v_i \in N(v_j)} (f(v_i, t) - f(v_j, t)) \cdot \omega_{ij} \quad (13)$$

Consider a road segment $v \in V$ and a time interval Δt. If the outlier value of v in Δt is large, then the amount of anomalies of v is large. Otherwise, the amount of anomalies of v is small. Thus, we assume that the "expected" outlier value of v in Δt can be regarded as the amount of anomalies of v at the starting timestamp of Δt in our model. In order to make this assumption holds, the initial amount of anomalies of each road segment should be set to the outlier value of v calculated based on the trajectory data by Equation (7).

The major idea of our diffusion based method for detecting the major anomaly sources is to find all road segments such that their "observed" outlier values (based on the deviation from regular traffic) deviates a lot from their "expected" outlier values (based on the diffusion of anomalies). The "observed" outlier value of a road segment in a given time interval can be calculated based on the trajectory data by Equation (7). In order to determine whether the deviation is large, we introduce a user parameter δ called the *major source threshold* (which is a non-negative real number). If the difference between the "observed" outlier value of a road segment v in a time interval δ and its "expected" outlier value is at least, we say that the anomalies of traffic of v in Δt is a major anomaly source. In this case, we re-start the anomalies diffusion model by re-initializing the initial value of the anomalies of each road segment v which is found to be a major anomaly source (representing the "expected" outlier value), to the current "observed" outlier value so that v can be

regarded as an anomaly source in the diffusion process.

Algorithm 1: Algorithm of Anomaly Sources Detection

Input: a major cause threshold δ and the outlier values $OV(v_i, \Delta t)$ for all road segments $v_i \in V$.
Output: A set S of major anomaly sources each in the form $(v, \Delta t)$ which is outputted in real time when found.

1 // Step 1 (Initialization)
2 $S \leftarrow \emptyset$
3 $t_{offset} \leftarrow 0$
4 $t \leftarrow 0$
5 $\Delta t \leftarrow [0, 30mins)$
6 **for** each $i \in [1, m]$ do
7 the i-th entry of **F**(0)$\leftarrow OV(v_i, \Delta t)$
8 **end for**
9 // Step 2 (Iterative Step)
10 **while true do**
11 $t \leftarrow t + 30$ mins
12 // Step 2(a) (Anomalies Diffusion)
13 compute F(t) according to Equation (18)
14 // Step 2(b) (Major Anomaly Source Finding)
15 **for** each $i \in [1, m]$ **do**
16 $F_i \leftarrow$the i-th entry of F(t)
17 **end for**
18 $\Delta t \leftarrow [t + t_{offset}, t + t_{offset} + 30$ mins)
19 **if** there exists an i such that $|OV(v_i, \Delta t) - F_i| \geq \delta$ **then**
20 **F**(0) $\leftarrow$ **F**(t)
21 **for all** i such that $|OV(v_i, \Delta t) - F_i| \geq \delta$ **do**
22 $S \leftarrow S \cup (v_i, \Delta t)$
23 **output** $(v_i, \Delta t)$
24 the i-th entry of **F**(0)$\leftarrow OV(v_i, \Delta t)$
25 **end for**
26 $t_{offset} = t_{offset} + t$
27 $t \leftarrow 0$
28 **end if**
29 **end while**

The process of discovering is shown in Algorithm 1. In this algorithm, we introduce three variables, namely S, t and t_{offset}. S is a variable used in this algorithm denoting the time difference between the starting timestamp (i.e., 0) and the timestamp of the current initial state. t is a variable used in this algorithm denoting the time difference between the current timestamp and the timestamp of the current initial state. We perform this algorithm according to the following two steps:

(i) *initialization step*. S is first initialized to $\varnothing$ (line 2). Since the starting timestamp is 0 and the timestamp of the current initial state is 0, t is set to 0 (line 4). Besides, we set a variable Δt to [0,30mins) (line 5). Then, we initialize **F**(0) by setting the i-th entry of **F**(0) to $OV(v_j, \Delta t)$ for each $i \in [1, m]$ where $\Delta t = [0,30 \min s]$ (lines 6-7).

(ii) *iterative step*. For each timestamp t which is a multiple of 30 minutes, we do the following two sub-steps. The first sub-step is called the anomalies diffusion step. In this sub-step, we compute **F**(t) (denoting the "expected" outlier values of all road segments) according to Equation (12) (line 13). The second sub-step called the major anomaly sources finding is to find all road segments which are major anomaly sources. Firstly, we find the "expected" outlier value of road segment v_i (i.e., the i-th entry of **F**(t) and set variable F_i to this value (line 16). Secondly, we set Δt to $[t + t_{offset}, t + t_{offset} + 30 \min s]$ (line 18). Thirdly, we check whether the difference between the "observed" outlier value of v_i and its "expected" outlier value is at least δ (i.e., $| OV(v_i, \Delta t) - F_i | \geq \delta$ (line 19). We have two cases. Case 1: there exists no i such that $| OV(v_i, \Delta t) - F_i | \geq \delta$. In this case, we find no major anomaly sources and thus do nothing. Case 2: there exists an i such that $| OV(v_i, \Delta t) - F_i | \geq \delta$. In this case, we find some major anomaly sources. Thus, we start a new heat diffusion process and reset the initial state by assigning the content of **F**(t) to **F**(0) (line 20). Besides, for such major anomaly sources $(v_i, \Delta t)$, we include it into S (line 22), output it (line 23) and update the i-th entry of **F**(0) with $OV(v_i, \Delta t)$ (line 24) since $(v_i, \Delta t)$ corresponds to a new heat source. At the end, since the initial state has been reset at this timestamp, we update t_{offset} by increasing it by t and reset t to 0 (line 26-27).

5. Experimental Evaluations and Analysis

5.1 Dataset and Parameters

Mobility Data: We use GPS trajectories as mobility data, As about 20% of traffic on road surfaces in Shenzhen is generated by taxicabs, the taxi trajectories represent a significant portion of the traffic flow on the road network. While we use taxi trajectory for validation, we believe our system and method are general enough to accept trajectory data generated by other sources, such as from public transit or location based check-in data, as long as they reflect mobility on the road network.

Traffic Anomaly Reports: We use the traffic anomaly reports published by transportation agencies [12, 13, 14] as the ground truth to evaluate the effectiveness of our approach, the statistics is shown in Table 3. Each event is associated with some road segments and the time interval for the event.

5.2 Evaluation Approach

In this evaluation, we carry out our method for outlier detection every 30 minutes and consider the taxi trajectories collected during this time interval as current data, and all the trajectories collected before as historical data to calculate the regular traffic behavior. The length of the time interval is a trade-off between the computational load and the timeliness of anomaly sensing.

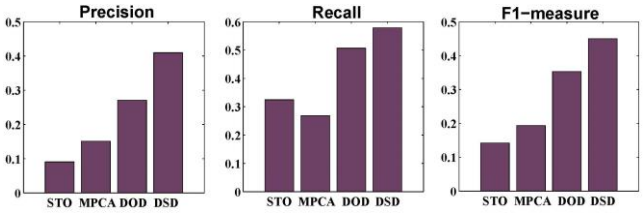

Fig. 4. Precision, recall and F1-measure of all algorithms evaluated based on the reported traffic incidents where the F1-measure of each algorithm is the greatest.

To evaluate the accuracy of our approach, we compare our algorithms, namely deviation-based outlier detection algorithm (DOD) and diffusion-based source detection algorithm (DSD), with a modified version of Principle Component Analysis (MPCA) in [6] and the STO algorithm in [5].

To evaluate the effectiveness of our approach, we consider the reported traffic incidents as ground truth, and evaluated all the algorithms in terms of three measurements, namely precision, recall and F1-measure.

5.3 Evaluation Result

(i) Comparison: Figure 4 shows the precision, the recall and the F1-measure of all algorithms evaluated based on reported traffic incidents. DSD has the greatest F1-measure value and DOD ((the one without being augmented with the "anomalies diffusion" property) has the second greatest F1-measure value. Therefore, we claim our approach significantly outperforms the pure traffic-volume approach. This is due to the fact that our approaches can detect the anomalies reflected not only from the traffic volume change of the road segment itself, but also from the traffic volume change of its adjacent road segments. In this figure, we note that the precision value of all algorithms are relatively low. This is due to the fact that the reported traffic incidents are not necessarily a complete set of ground truth.

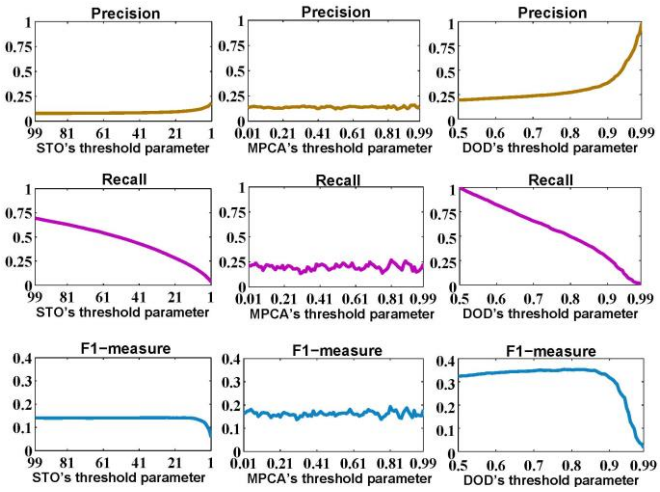

Fig. 5. Effect of input parameters of algorithms (i.e., STO, MPCA and DOD).

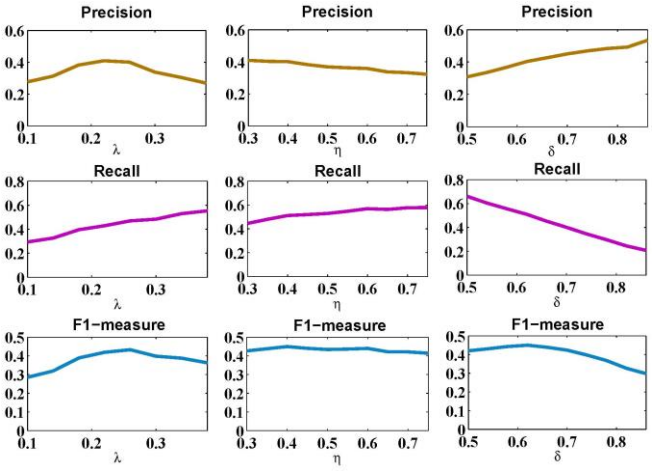

Fig. 6. Effect of input parameters of DSD.

(ii) Varying Parameter: To further show the superiority of our approach, we perform various evaluations to see how the measurements are affected when we vary the parameters of each algorithm. Note that STO requires one input parameter $h \in^{+} Z$, MPCA requires one input parameter $1 \in [0,1]$, and DOD requires one input parameter $\xi \in [0,1]$. Figure 5 shows the effect of input parameters of each algorithm. From these sub-figures, when the parameter of DOD set to a value smaller than 0.95, the F1-measure value of DOD greater than the F1-measure value of both STO and MPCA. This means that even if many possible parameters are tested on STO and MPCA, their F1-measure values are not better than DOD in most cases. In Figure 6, we also show the parameters effects of DSD by fixing two parameters (say, λ and η) while varying the other parameter (say δ). In most cases, the precision value, the recall value and the F1-measure

value of our proposed algorithms are larger than those of the two state-of-the-art algorithm, namely STO and MPCA. Moreover, compared with DOD, the number of false positives returned by DSD is smaller since it can identify the true positives more easily.

6. Conclusion

In this paper, we presented a novel framework to detect outlier patterns and sources using a large scale vehicular trajectory data. Furthermore, a wide set of experimental results show that our algorithms are superior compared with the state-of-the-art algorithms. The system for detecting such traffic anomalies can benefit both drivers and transportation authorities, e.g., by notifying drivers approaching an anomaly, as well as supporting traffic jam diagnosis. In the future, we plan to investigate the semantic meaning of traffic anomalies.

Reference

[1] J. Lee, J. Han, and K. Whang. "Trajectory clustering: a partition-and-group framework." In Proceedings of the ACM SIGMOD, pages 593-604, 2007.
[2] Y. Bu, L. Chen, A. Fu, and D. Liu, "Efficient anomaly monitoring over moving object trajectory streams." In Proceedings of the ACM SIGKDD, pages 159-168, 2009.
[3] Wu. E, Liu. W, and Chawla. S, "Spatio-temporal outlier detection in precipitation data." In Knowledge Discovery from Sensor Data, pages 115-133, Springer Berlin Heidelberg, 2010.
[4] F. Chen, C.-T. Lu, and A. P. Boedihardjo. "Gls-sod: a generalized local statistical approach for spatial outlier detection." In Proceedings of the ACM SIGKDD, pages 1069-1078, 2010.
[5] Liu. W, Zheng. Y, Chawla. S, Yuan. J, and Xing, X. "Discovering spatio-temporal causal interactions in traffic data streams." In Proceedings of the ACM SIGKDD, pages 1010-1018, 2011.
[6] Chawla. S, Zheng. Y, and Hu. J, "Inferring the root cause in road traffic anomalies." In Proceedings of the IEEE ICDM, pages 141-150, 2012.
[7] Barnett. V, and Lewis. T, "Outliers in statistical data." Vol. 3. New York: Wiley, 1994.
[8] D. Lopez-Pintado. "Diffusion in complex social networks." In Journal of Economic Literature, pages 573-590, 2004.
[9] W. Chen, Y. Wang, and S. Yang. "Efffcient influence maximization in social networks." In Proceedings of the ACM SIGKDD, pages 199-208, 2009.
[10] D. Kempel, J. Kleinberg, and E. Tardos. "Maximizing the spread of inffluence through a social network." In Proceedings of the ACM SIGKDD, pages 137-146, 2003.
[11] H. Wei, Y. Wang, and G. Forman. "Map Matching: Comparison of Approaches using Sparse and Noisy Data." In Proceedings of the ACM SIGSPTIAL, pages 434-437, 2013.
[12] "Shenzhen News online edition." In http://dtzbd.sznews.com/.
[13] "Shenzhen traffic online edition." In http://www.szjiaotong.com/.
[14] "Shenzhen Government online edition." In http://www.sz.gov.cn/cn/.