# A New Framework for Traffic Anomaly Detection *

Jinsong Lan, Cheng Long, Raymond Chi-Wing Wong, Youyang Chen, Yanjie Fu
Danhuai Guo, Shuguang Liu, Yong Ge, Yuanchun Zhou, Jianhui Li

**Abstract**

Trajectory data is becoming more and more popular nowadays and extensive studies have been conducted on trajectory data. One important research direction about trajectory data is the anomaly detection which is to find all anomalies based on trajectory patterns in a road network. In this paper, we introduce a *road segment-based* anomaly detection problem, which is to detect the abnormal road segments each of which has its "real" traffic deviating from its "expected" traffic and to infer the *major causes* of anomalies on the road network. First, a *deviation-based* method is proposed to quantify the anomaly of reach road segment. Second, based on the observation that one anomaly from a road segment can trigger other anomalies from the road segments nearby, a *diffusion-based* method based on a *heat diffusion model* is proposed to infer the major causes of anomalies on the whole road network. To validate our methods, we conduct intensive experiments on a large real-world GPS dataset of about 23,000 taxis in Shenzhen, China to demonstrate the performance of our algorithms.

## 1 Introduction

Nowadays, with the advanced development of communication and information infrastructure, the movement of objects can be captured in many ways and a large number of location traces of moving objects have been accumulated. Particularly, GPS tracking devices have been installed on taxis in many cities over the world. In the literature, there are a lot of existing studies [1, 2, 3, 4, 5, 6, 7] which analyzed the location traces of moving objects. For instance, [1, 2] proposed different approaches to find the "periodic" movement patterns. [3, 4] are to find frequent movement patterns. [5, 6] proposed methods to find abnormal movement patterns. In fact, finding abnormal movement patterns, formally called *anomaly detection*, has become a hot topic [5, 6, 8, 9, 10, 11] due to the various applications of anomaly detection such as traffic jam detection and traffic pattern monitoring.

Let us first introduce a concrete example to illustrate the problem of anomaly detection, which we will focus on in this paper. Consider the road network as shown in Figure 1(a). In this figure, there are 7 road segments, namely $e_1, e_2, ..., e_7$. In this figure, each road segment is associated with two numbers in the form of $(\mu, \sigma)$ where $\mu$ denotes the *expected* number of taxis passing this road segment at 11:00am every Friday, and $\sigma$ denotes the standard deviation of the number of taxis passing this road segment at the same time. These two numbers can be obtained from the past taxi movement data. Here, $\mu$ corresponds to the *expected traffic* of this road segment. Figure 1(b) shows the same road network but each road segment is associated with a single number, called the *real traffic*, denoting the *real* number of taxis passing this road segment at 11:00am on June 1, 2012 (Fri). Let us consider the road segment $e_1$. The expected number of taxis passing $e_1$ at 11:00am every Friday is 100 (Figure 1(a)) and the real number of taxis passing $e_1$ at 11:00am on June 1, 2012 (Fri) is 150 (Figure 1(b)). Since the real traffic of $e_1$ at that time (i.e., 150) is "statistically" much greater than the expected traffic of $e_1$ (i.e., 100) (in this case, 150 is greater than $100+3\sigma$), the real traffic of $e_1$ at that time is considered as an *anomaly*. Similarly, we can conclude that all five road segments "near" to $e_1$, namely $e_2, e_3, e_4, e_5$ and $e_6$, have their real traffic considered as abnormal.

In most cases, since there are some correlations among the traffic on different road segments, an anomaly on one road segment *triggers* a lot of anomalies in other road segments "near" this road segment. Although outputting all anomalies can help people understand all anomalies in the road network, in some cases returning the *major cause* of the anomaly is more interesting to people because people can better understand the reason why there are a lot of anomalies. For instance, suppose we can find that there is a car accident on the road segment $e_1$ in the figure 1(b), which triggers the abnormal traffic of all road segments "near" to $e_1$, the abnormal traffic of the road segment $e_1$ causes not only the

*J. Lan, Computer Network Information Center, Chinese Academy of Sciences, China, lanjinsong@cnic.cn (University of Chinese Academy of Sciences, China). C. Long, R. C.-W. Wong, Hong Kong University of Science and Technology, Hong Kong, {clong,raywong}@cse.ust.hk. Y. Chen, Beijing Institute of Technology, China, chen.youyang@qq.com. Y. Fu, Rutgers University, USA, yanjie.fu@rutgers.edu. D. Guo, Computer Network Information Center, Chinese Academy of Sciences, China, guodanhuai@cnic.cn. S. Liu, Chongqing Institutes of Green and Intelligent Technology, Chinese Academy of Sciences, China, liushuguang@cigit.ac.cn. Y. Ge, University of North Carolina at Charlotte, USA, yong.ge@uncc.edu. Y. Zhou, Computer Network Information Center, Chinese Academy of Sciences, China, zyc@cnic.cn. J. Li (contact author), Computer Network Information Center, Chinese Academy of Sciences, China, lijh@cnic.cn


(a) 11:00am every Friday

(b) 11:00am June 1, 2012 (Friday)

Figure 1: An example showing anomalies

abnormal traffic of each of its adjacent road segments (i.e., $e_2, e_3, e_4$ and $e_5$) but also the abnormal traffic of a road segment farther away from $e_1$ (i.e., $e_6$).

To this end, in this paper, we study the anomaly detection of traffic with the following two goals. The first goal is to find the road segments with abnormal traffic. The second goal is to find the major cause(s) of the anomalies found in the previous goal.

In the literature, there are many works about traffic anomaly detection with location traces. Particularly [5, 6] proposed different methods to detect the anomaly with taxi GPS data. However those methods could not meet our two goals. To explain the reason, let us first describe their algorithms. Both algorithms share the following two-step framework. The first step is to partition the whole space into a number of *regions* via all major road segments in the road network. In Figure 1(a), there are 4 regions, namely $A, B, C$ and $D$. For example, Region $B$ is enclosed by road segments $e_1, e_2$ and $e_4$. The second step is to find each abnormal traffic *path* from a region to another region. In Figure 1(a), some possible abnormal traffic paths are "Region $A$ to Region $B$" and "Region $A$ to Region $C$". Note that [5] and [6] are different in the second step under the two-step framework.

Now, we describe how those algorithms do not meet our goals. Firstly, they cannot find the abnormal traffic of *road segments*. Instead, they could only find the abnormal traffic *paths* from regions to regions. Secondly, they cannot find the major causes of the anomalies on *road segments*, while they could only find the major causes of some abnormal *paths*. Thirdly, more importantly, the underlying two-step framework suffers from a fundamental *boundary* problem. Specifically, even though two different taxis move on the same major road segment in the road network, due to the imprecise recording of the GPS device equipped in the taxi, the location of the taxi can be recorded in one region and the location of the other taxi can be recorded in another region. To illustrate this, let us consider the road segment $e_1$ in Figure 1(a). If two taxis move on $e_1$, the location of one taxi may be recorded in Region $B$ but the location of another taxi may be recorded in Region $C$. The major reason for this problem is that the taxi is moving along the *boundary* among regions.

In this paper, we propose a novel method for each of the two goals in the anomaly detection problem, which does not have the above drawbacks. For the first goal, we propose a *deviation-based* method. This method finds all road segments on which the real traffic statistically *deviates* from its expected traffic a lot. For the second goal, we propose a *diffusion* model which finds the major causes of anomalies in the road network. Specifically, we observe that the abnormal traffic of a road segment affects the abnormal traffic of road segments "near" to this road segment *progressively*. This phenomenon is similar to the heat diffusion process on an object. In this process, the heat energy spreads from a single source to other places in the object *progressively*. Motivated by this observation, we propose a *heat diffusion model* for the second goal where we regard the major causes of abnormal traffic in our problem as the sources of heat energy in the heat diffusion model. Both of our methods do not suffer from the boundary problem in the state-of-the-art algorithms since our methods are based on road segments instead of paths and regions.

The following shows our contributions. First, we study the road-segment based anomaly detection problem which is more reasonable than the existing *path-based* one in applications where the abnormal traffic of road segments (e.g., car accidents on road segments) is important. We have two goals. The first goal is to find all anomalies in the road network and the second goal is to find the major causes of anomalies. Second, we propose two methods, namely a *deviation-based* method and a *diffusion-based* method, for the two goals. Third, we conduct experiments on a large real dataset (600 GB) containing 23,000 taxis in Shenzhen from January 1, 2012 to August 28, 2012 to demonstrate the performance of our methods.

This paper is organized as follows. Section 2 gives our problem definition. Section 3 presents our proposed methods. Section 4 shows our experimental results. Section 5 describes the related work. Section 6 concludes our paper.

## 2 Problem Definition

We are given a directed graph $G = (V, E)$, representing a road network, where $V$ denotes a set of vertices in the network and $E$ denotes a set of edges in the network. Each vertex in $V$ is associated with its spatial location, namely the latitude and the longitude. Each edge in $E$ is also called a *road segment*.

An objects (i.e., vehicle), equipped with the GPS device, is moving in the road network. The GPS device records the movement of this object at some regular timestamps. Specifically, at each timestamp $t$, it records the spatial

location of this object, namely its latitude and its longitude. Thus, each object is associated with a sequence of entries in the form of $(p, t)$ where $p$ is the spatial location of an object at time $t$. This sequence is called the *trajectory* of this object. Suppose that we are given a set of $T$ trajectories from $n$ objects.

For the trajectory of a moving object, since the GPS device records its spatial location in a regular time period, it does not record its spatial location in some timestamps. However, following existing studies, we can estimate the spatial location of a moving object at some timestamps not recorded by the GPS device. Suppose the GPS device records two entries, namely $(p_1, t_1)$ and $(p_2, t_2)$, where $t_1 < t_2$ and we want to know its estimated position at timestamp $t$ where $t_1 < t < t_2$, there are a lot of existing ways which could estimate the position at timestamp $t$. One way is to use the linear interpolation method by assuming a constant speed of this object traveling from $p_1$ to $p_2$ in the road network. Since the position estimation method is not our focus in this paper, we simply adopt the linear interpolation method. But other position estimation methods can also be used.

Given two timestamps $t$ and $t'$ where $t < t'$, we represent a *time interval* in the form of $[t, t')$, denoting all timestamps at least $t$ and smaller than $t'$. Given a road segment $e \in E$ and a time interval $\triangle t = [t, t')$, an object is said to *pass* the road segment $e$ in $\triangle t$ if there exists a timestamp $t \in [t, t')$ such that its estimated position at timestamp $t$ is along the road segment $e$. Given a road segment $e \in E$ and a time interval $\triangle t = [t, t')$, the *traffic* of the road segment $e$ in the time interval $\triangle t$, denoted by $f(e, \triangle t)$, is defined to be the total number of objects passing $e$ in the time interval $\triangle t$.

In this paper, in order to define *anomalies* more precisely, we divide all timestamps based on both *30-minute time slots* [1] and *days*. Firstly, we divide all timestamps based on *30-minute time slots* as follows. We divide a single 24-hour day into 48 time slots each of which lasts for 30 minutes. For example, the first time slot is from 00:00 to 00:30, and the second time slot is from 00:30 to 01:00. Each time slot is called a *time bin*. Secondly, we partition all these time bins into a number of groups based on *days*. There are two ways of partitioning, namely the *day-of-the-week partitioning* and the *weekday-weekend partitioning*. For the day-of-the-week partitioning, since there are seven days per week , 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 $48 \times 7 = 336$ possible time bins. For the weekday-weekend partitioning, we create two groups where one group is for the weekdays and the other group is for the weekends, and each group contains a set of time bins which belong to this group. Each group is called a *weekday-weekend group*. In this partitioning, there are $48 \times 2 = 96$ possible time bins. Note that there are other different ways of partitioning. For the sake of space, we focus on the above two ways of partitioning.

In the following, we consider one way of partitioning (e.g., the day-of-the-week partitioning). Let $\mathcal{B}$ be a set of all possible time bins based on this partitioning. Given a time interval $\triangle t$, we define a mapping function $M(\cdot)$ which takes $\triangle t$ as input and returns the time bin $b \in \mathcal{B}$ that $\triangle t$ belongs to. For example, suppose that $\triangle t$ is the time interval from 11:00am to 11:30am on June 1, 2012 (Fri), and $b$ is the possible time bin in $\mathcal{B}$ which is the time slot from 11:00am to 11:30am on Friday. Then, $b = M(\triangle t)$. We say that $\triangle t$ is a *time sample* for $b$.

Consider a time interval $\triangle t$. Given a time bin $b \in \mathcal{B}$ and a road segment $e \in E$, if $\triangle t$ is a time sample for $b$, then we say that the traffic of $e$ in $\triangle t$ is a *traffic sample* of $e$ for $b$.

Based on the past dataset, for each road segment $e \in E$ and each possible time bin $b \in \mathcal{B}$, we can find the *distribution* on the traffic samples of this road segment $e$ for this time bin $b$, denoted by $D(e, b)$. We will describe how to find $D(e, b)$ later in this paper.

In this paper, we study the following two goals for the anomaly detection problem. The first goal is detecting the *anomaly road segment*. That is, given a time interval $\triangle t$, we find a set of all road segments in $E$, denoted by $S(\triangle t)$, such that the traffic of each road segment $e$ in the time interval $\triangle t$ in this set deviates a lot from its "expected" traffic based on $D(e, M(\triangle t))$. Later, we will describe what we mean by "expected". Each road segment in $S(\triangle t)$ is called an *anomaly*.

The second goal is inferring the *major anomaly causes*. That is, given a time interval $\triangle t$, we find all road segments in $S(\triangle t)$ which are the major "causes" of the anomalies in $S(\triangle t)$. Later, we will describe what we mean by "causes".

## 3 Methodology

In this section, we first present how we process the trajectory data which cannot be found in the state-of-the-art methods [5, 6] (Section 3.1). Then, we present our deviation-based method for the first goal (Section 3.2) and our diffusion method for the second goal (Section 3.3).

**3.1 Processing Trajectory Data** In this section, we introduce an operation called *map matching*, a well-established technique, to locate a "recorded" trajectory (in the two-dimensional space) to the road network. Performing this operation is beneficial since the boundary problem described in

[1] We use time slots with the duration equal to 30 minutes in this paper and the reason is that time slots with shorter durations (e.g., 15-minutes) would suffer from the data sparsity issue while those with longer slots (e.g., 1-hour) would cause serious imprecise problems. However, this setting can be changed accordingly based on the requirement of the granularity of the duration of a time slot in other applications.

Section 1 does not appear after this operation is executed.

Specifically, we first perform a *map matching* operation which maps each trajectory in the dataset to a sequence of road segments in $E$. The map matching algorithm we used is [12]. Note a trajectory is originally represented by a sequence of 2-tuples in the form of $(p, t)$ where $p$ is a *spatial location* and $t$ is the timestamp when an object is at $p$. After we perform this operation on a trajectory, we obtain a sequence of 2-tuples in the form of $(e, t)$ where $e$ is an *edge* in $E$ and $t$ is the estimated timestamp when an object is at $e$. In the following, for clarity, when we describe a trajectory, we mean the sequence obtained after the map matching operation.

**3.2 Deviation-based Method** In this section, we present a *deviation-based method* for the first goal (i.e., the anomaly road segment finding goal). Although this method is simple, it is useful in practice in some cases compared with state-of-the-art methods (which will be shown in our experiments).

Given a time interval $\triangle t$, we want to find a set of all road segments in $E$, denoted by $S(\triangle t)$, such that the traffic of each road segment $e$ in the time interval $\triangle t$ in this set deviates a lot from its "expected" traffic based on $D(e, M(\triangle t))$. There are the following three issues. The first issue is to define $D(e, b)$ precisely where $b$ is a possible time bin. The second issue is to give a precise description of the "expected" traffic based on $D(e, b)$. The third issue is to describe the meaning of "deviates a lot".

Consider the first issue. We analyze our real taxi trajectory dataset collected in Shenzhen from January 1, 2012 to August 28, 2012 according to two different partitioning ways. The detailed description of this dataset can be found in Section 4. Consider the weekday-weekend partitioning and an arbitrary road segment $e$. For each possible time bin $b$ representing weekday/weekend and a 30-minute time slot (e.g., 11:00am to 11:30am), we want to draw two figures to analyze the distribution on the traffic samples of $e$ for this time bin $b$ (i.e., $D(e, b)$). The first figure is the histogram on all possible traffic samples of $e$ for $b$. In this figure, the x-axis denotes all possible traffic values of $e$ for $b$, and the y-axis denotes the number of traffic samples which have the corresponding possible traffic value of $e$ for $b$. Figure 2 shows this histogram on our real dataset. The second figure is the Q-Q plot (or the quantile-quantile plot) of all possible traffic samples of $e$ for $b$ versus a normal distribution. In the figure, the x-axis corresponds to the normal theoretical quantiles and the y-axis corresponds to the data quantile. 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 $y$. Figure 3 shows the Q-Q plot on our real dataset.

We have the following two observations based on the figures. Firstly, from the histogram, we observe that the traffic samples in our real dataset look like a normal distribution. Secondly, from the Q-Q plot, we observe that the points in the plot lie on a single line, which means that the distribution on the traffic samples of $e$ for this time bin $b$ (i.e., $D(e, b)$) is similar to a normal distribution. Based on these observations, we model $D(e, b)$ as a normal distribution.

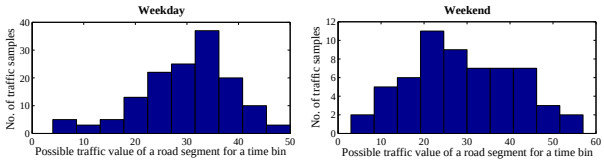


Figure 2: Histogram on all possible traffic samples of a road segment for a time bin (Weekday-weekend Partitioning)

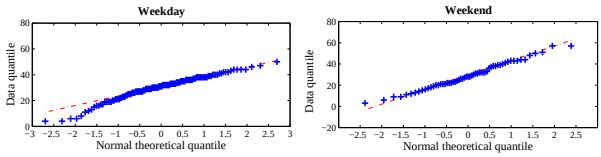


Figure 3: Q-Q plot of all possible traffic samples of a road segment for a time bin versus a Normal Distribution (Weekday-weekend Partitioning)

Similarly, we have the histogram figure and the Q-Q plot figure when the day-of-the-week partition is adopted, based on which, we can also model $D(e, b)$ as a normal distribution. Due to space limit, these figures are not shown here.

In summary, regardless of which partition is used, $D(e, b)$ can be modeled as a normal distribution. Thus, based on all possible traffic values of $e$ for $b$, we can calculate the mean and the standard deviation, denoted by $\mu(e, b)$ and $\sigma(e, b)$, for these values, two parameters in the normal distribution.

The second issue can be easily addressed after the first issue is addressed. Since $D(e, b)$ can be modeled as a normal distribution, the "expected" traffic based on $D(e, b)$ is exactly equal to the mean of this distribution (i.e., $\mu(e, b)$).

The third issue can be addressed by introducing a concept of "anomaly value". Given a road segment $e \in E$ and a time interval $\triangle t$, the *anomaly value* of $e$ in $\triangle t$, denoted by $A(e, \triangle t)$, is defined as follows.

$$A(e, \triangle t) = 2 \cdot \frac{1}{1 + exp\{-\frac{|f(e,\triangle t) - \mu(e,b)|}{\sigma(e,b)}\}} - 1 \tag{3.1}$$

where $b = M(\triangle t)$. Since the expression "$\frac{1}{1+exp\{-\frac{|f(e,\triangle t)-\mu(e,b)|}{\sigma(e,b)}\}}$" above is a sigmoid function ranging from 0.5 to 1 for any non-negative real number $(e, \triangle t)$, $A(e, \triangle t)$ ranges from 0 to 1. If the difference between the traffic (i.e., $f(e, \triangle t)$) and its expected traffic (i.e., $\mu(e, b)$) is larger, then $A(e, \triangle t)$ will be larger. Otherwise, it will be smaller. For example, if the difference between $f(e, \triangle t)$ and $\mu(e, b)$ is equal to $3\sigma$, then $A(e, \triangle t)$ is equal to 0.905. Thus, Equation (3.1) captures the extent of the

deviation of the traffic from the expected traffic.

Now, we introduce a user parameter called the *anomaly threshold* $a_o$, which is a non-negative real number from 0 to 1, to determine whether a road segment $e$ in a time interval $\triangle t$ is an *anomaly*. Given a road segment $e$ and a time interval $\triangle t$, the road segment $e$ in the time interval $\triangle t$ is said to be *anomaly* if $A(e, \triangle t)$ is at least $a_o$.

Thus, we propose our deviation-based method which is to determine all road segments with their anomaly values at least $a_o$.

**3.3 Diffusion Method** In this section, we propose a *diffusion method* for the second goal. Specifically, we want to find all major anomaly causes of the anomalies found in the first goal. We first introduce a model called the *heat diffusion model* (Section 3.3.1) and then describe how we use this model to find all major anomaly causes (Section 3.3.2).

**3.3.1 Heat Diffusion Model** We observe that the phenomenon that the abnormal traffic of a road segment affects the abnormal traffic of road segments "near" to this road segment *progressively* is similar to the heat diffusion phenomenon on an object that the heat energy spreads from a single source to other places in the object *progressively*. Motivated by this observation, we propose a *heat diffusion model* for the second goal where we regard the major causes of abnormal traffic in our problem as the sources of heat energy in the heat diffusion model.

Before we introduce our heat diffusion model, we define the concept of "adjacency". Given two edges $e$ and $e'$ in $E$, $e$ is said to be *adjacent* to $e'$ if $e$ and $e'$ shares a vertex. $e$ is also called a *neighbor* of $e'$. Given an edge $e \in E$, we denote the set of all possible neighbors of $e$ by $N(e)$. Given a time interval $\triangle t$, the *duration* of the time interval, denoted by $|\triangle t|$, is defined to how long the time interval is. For example, the duration of the time interval from "Jul 1, 2012 11:00am" to "Jul 1, 2012 11:30am" is 30 minutes.

Now, we present our heat diffusion model as follows. Each edge $e \in E$ is associated with *heat energy*. We denote the *amount* of the heat energy for $e$ at a timestamp $t$ by $E(e, t)$ which is a non-negative real number. At timestamp $t$, each edge $e$ receives an amount of heat from its neighbor $e'$ during the time interval $\triangle t$ whose start timestamp is equal to $t$. This amount is denoted by $Q(e, e', t, \triangle t)$. It should be proportion to the duration of the time interval (i.e., $|\triangle t|$) and the difference in the amount of the heat energy between $e$ and $e'$ (i.e., $E(e', t) - E(e, t)$). Besides, the heat energy flows from $e'$ to $e$ via the vertex shared by these two edges. Thus, we have

$$Q(e, e', t, \triangle t) = \alpha \cdot |\triangle t| \cdot (E(e', t) - E(e, t))$$

where $\alpha$ is a non-negative real number called the *thermal conductivity*-the heat diffusion coefficient. Thus, the total amount of heat energy that an edge $e$ receives between timestamp $t$ and timestamp $t + |\triangle t|$, denoted by $\delta(e, [t, t + |\triangle t|])$, is equal to the sum of the amount of heat energy that $e$ receives from all of its neighbors, which is equal is

$$\alpha \cdot |\triangle t| \cdot \sum_{e' \in N(e)} (E(e', t) - E(e, t))$$

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 heat diffusion model, we model this external mechanism by an exponential decay process in which the amount of heat energy of an edge decreases with time exponentially. Thus, the total amount of heat energy of $e$ which can be *kept* at timestamp $t + |\triangle t|$ due to its original heat energy at timestamp $t$, denoted by $R(e, t + |\triangle t|)$, is

$$exp(-\lambda|\triangle t|) \cdot E(e, t)$$

where $\lambda$ is the *decay factor* of the exponential decay process. Thus, the total amount of heat energy of $e$ at timestamp $t + |\triangle t|$ (i.e., $E(e, t + |\triangle t|)$) is equal to the following.

$$E(e, t + |\triangle t|) = \delta(e, [t, t + |\triangle t|]) + R(e, t + |\triangle t|)$$

That is,

$$E(e, t + |\triangle t|) = \alpha \cdot |\triangle t| \cdot \sum_{e' \in N(e)} (E(e', t) - E(e, t)) + exp(-\lambda|\triangle t|) \cdot E(e, t)$$

We re-write it as follows.

$$\frac{E(e, t + |\triangle t|) - exp(-\lambda|\triangle t|) \cdot E(e, t)}{|\triangle t|} = \alpha \cdot \sum_{e' \in N(e)} (E(e', t) - E(e, t)) \tag{3.2}$$

Suppose that $e_1$, $e_2$, ..., $e_m$ correspond to the road segments in $E$. Let $\mathbf{E}(t)$ be a vector with $m$ entries where the $i^{th}$ entry is equal to $E(e_i, t)$, i.e., $\mathbf{E}(t) = (E(e_1, t), E(e_2, t), ..., E(e_m, t))^T$. We can further express Equation (3.2) as follows.

$$\frac{\mathbf{E}(t + |\triangle t|) - exp(-\lambda|\triangle t|)\mathbf{E}(t)}{|\triangle t|} = \alpha \mathbf{H}\mathbf{E}(t) \tag{3.3}$$

where $\mathbf{H}$ is an $m \times m$ matrix with

$$H_{ij} = \begin{cases} 1 & \text{if } e_i \text{ is a neighbor of } e_j \\ -|N(e_i)| & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

We approximate $exp(-\lambda|\triangle t|)$ with the first two terms of its *Taylor series*, that is,

$$exp(-\lambda|\triangle t|) \approx 1 + (-\lambda|\triangle t|) \tag{3.4}$$

Substituting Equation (3.4) in Equation (3.3), we obtain

$$\frac{\mathbf{E}(t+|\triangle t|)-\mathbf{E}(t)+\lambda|\triangle t|\mathbf{E}(t)}{|\triangle t|}=\alpha\mathbf{H}\mathbf{E}(t) \tag{3.5}$$

That is,

$$\frac{\mathbf{E}(t+|\triangle t|)-\mathbf{E}(t)}{|\triangle t|}=\alpha\mathbf{H}\mathbf{E}(t)-\lambda\mathbf{E}(t) \tag{3.6}$$

With $\triangle t$ approaching 0, we further deduce that

$$\frac{d}{dt}\mathbf{E}(t)=\mathbf{H}'\mathbf{E}(t) \tag{3.7}$$

where

$$\mathbf{H}'=\alpha\mathbf{H}-\lambda\mathbf{I}$$

By solving Equation (3.7), we obtain

$$\mathbf{E}(t)=exp(\mathbf{H}'t)\mathbf{E}(0) \tag{3.8}$$

where

$$exp(\mathbf{H}'t)=\mathbf{I}+t\mathbf{H}'+\frac{t^2}{2!}\mathbf{H}'^2+...$$

We call the matrix $exp(\mathbf{H}'t)$ the *diffusion kernel*.

**3.3.2 How to Find Major Anomaly Causes** Next, we present our method which finds major causes of anomalies.

As we described before, we regard the major causes of abnormal traffic in our problem as the sources of heat energy in the heat diffusion model. Consider a road segment $e \in E$ and a time interval $\triangle t$. If the anomaly value of $e$ in $\triangle t$ is large, then the amount of heat energy of $e$ is large. Otherwise, the amount of heat energy of $e$ is small. Thus, we assume that the "expected" anomaly value of $e$ in $\triangle t$ can be regarded as the amount of heat energy of $e$ at the starting timestamp of $\triangle t$ in our model. In order to make this assumption holds, the *initial* amount of heat energy of each edge should be set to the anomaly value of $e$ calculated based on the trajectory data (by Equation (3.1)).

The major idea of finding the major anomaly causes is to find all road segments such that their "observed" anomaly values deviates a lot from their "expected" anomaly values in a given time interval. The "observed" anomaly value of a road segment in a given time interval can be calculated based on the trajectory data (by Equation (3.1)). As described before, the "expected" anomaly value of a road segment in a given time interval can be regarded as the amount of the heat energy of $e$ at the starting timestamp of $\triangle t$ in the heat diffusion model. In order to determine whether the deviation is large, we introduce a user parameter $\epsilon$ called the *major cause threshold* (which is a non-negative real number). If the difference between the "observed" anomaly value of a road segment $e$ in a time interval $\triangle t$ and its "expected" anomaly value is at least $\epsilon$, we say that the traffic of $e$ in $\triangle t$ is a *major anomaly cause*. Whenever we find a major anomaly cause of a road segment $e$, we can find out a heat source in the heat diffusion model. In this case, we re-start the heat diffusion model by re-initializing the initial state/value of the heat energy of each road segment $e$ which is found to be a major anomaly cause (representing the "expected" anomaly value) to the current "observed" anomaly value so that $e$ can be regarded as a heat source in the proceeding diffusion process.

Now, we are ready to introduce our diffusion-based method which is presented in Algorithm 1. Algorithm 1 maintains three variables, namely $S, t$ and $t_{offset}$. $S$ is a variable which stores the set of all major anomaly causes each in the form of $(e, \triangle t)$ meaning that the traffic of the road segment $e$ in $\triangle t$ is a major anomaly cause. $t_{offset}$ 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.

Algorithm 1 involves two steps. The first step is the *initialization step*. $S$ is first initialized to $\emptyset$ (line 2). Since the starting timestamp is 0 and we assume that the first initial state starts at this timestamp, $t_{offset}$ is set to 0 (line 3). Since the current 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 $\triangle t$ to [0, 30 mins) (line 5). Then, we initialize $\mathbf{E}(0)$ by setting the $i$-th entry of $\mathbf{E}(0)$ to $A(e_i, \triangle t)$ for each $i \in [1, m]$ where $\triangle t$ = [0,30 mins) (lines 6-7).

The second step is the *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 *heat diffusion* step. In this sub-step, we compute $\mathbf{E}(t)$ (denoting the "expected" anomaly values of all road segments) according to Equation (3.8) (line 12). The second sub-step called the *major anomaly cause finding* is to find all road segments which are major anomaly causes. In this sub-step, we do 3 things as follows. Firstly, we find the "expected" anomaly value of road segment $e_i$ (i.e., the $i$-th entry of $\mathbf{E}(t)$) and set variable $E_i$ to this value. (line 15). Secondly, we set $\triangle t$ to $[t + t_{offset}, t + t_{offset}$+30 mins) if $t + t_{offset}$+30 mins is at most the greatest time stamp of the trajectory dataset (line 16-17); otherwise, we terminate the algorithm (line 18). Thirdly, we check whether the difference between the "observed" anomaly value of $e_i$ and its "expected" anomaly value is at least $\epsilon$ (i.e., $|A(e_i, \triangle t) - E_i| \geq \epsilon$). We have two cases. Case 1: there exists no $i$ such that $|A(e_i, \triangle t) - E_i| \geq \epsilon$. In this case, we find no major anomaly causes and thus do nothing. Case 2: there exists an $i$ such that $|A(e_i, \triangle t) - E_i| \geq \epsilon$ (line 19). In this case, we find some major anomaly causes. Thus, we start a new heat diffusion process and re-set the initial state by assigning the content of $\mathbf{E}(t)$ to $\mathbf{E}(0)$ (line 20). Besides, for such major

**Algorithm 1** Algorithm for finding all major anomaly causes

**Input:** a major cause threshold $\epsilon$
**Output:** A set $S$ of major anomaly causes each in the form of $(e, \triangle t)$ 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: $\triangle t \leftarrow$ [0,30 mins)
6: **for** each $i \in [1, m]$ **do**
7: the $i$-th entry of $\mathbf{E}(0) \leftarrow A(e_i, \triangle t)$
8: // Step 2 (Iterative Step)
9: **while true do**
10: $t \leftarrow t + 30$ mins
11: // Step 2(a) (Heat Diffusion)
12: compute $\mathbf{E}(t)$ according to Equation (3.8)
13: // Step 2(b) (Major Anomaly Cause Finding)
14: **for** each $i \in [1, m]$ **do**
15: $E_i \leftarrow$ the $i$-th entry of $\mathbf{E}(t)$
16: **if** $t + t_{offset}$+30 mins is at most the greatest time stamp of the trajectory dataset **then**
17: $\triangle t \leftarrow [t + t_{offset}, t + t_{offset}$+30 mins)
18: **else break**
19: **if** there exists an $i$ such that $|A(e_i, \triangle t) - E_i| \geq \epsilon$ **then**
20: $\mathbf{E}(0) \leftarrow \mathbf{E}(t)$
21: **for all** $i$ such that $|A(e_i, \triangle t) - E_i| \geq \epsilon$ **do**
22: $S \leftarrow S \cup \{(e_i, \triangle t)\}$
23: **output** $(e_i, \triangle t)$
24: the $i$-th entry of $\mathbf{E}(0) \leftarrow A(e_i, \triangle t)$
25: $t_{offset} \leftarrow t_{offset} + t$
26: $t \leftarrow 0$

anomaly cause $(e_i, \triangle t)$ (line 21), we include it into $S$ (line 22), output it (line 23) and update the $i$-th entry of $\mathbf{E}(0)$ with $A(e_i, \triangle t)$ (line 24) since $(e_i, \triangle t)$ corresponds to a new heat source. At the end, since the initial state has been re-set at this timestamp, we update $t_{offset}$ by increasing it by $t$ (line 25)and reset $t$ to 0 (line 26).

## 4 Experiments

**4.1 Data and Set-up** We used a taxi trajectory dataset which contains the trajectories of 23,876 taxis in Shenzhen City within a period of 8 months (from January 1, 2012 to August 28, 2012). In this dataset, the spatial location of a taxi is sampled with an average sampling rate of 33.3 seconds and the average distance between two consecutive sampled positions is 316.9 meters.

We constructed the road network of Shenzhen from OpenStreeMap (OSM) (www.openstreetmap.org). We have about 300 road segments in the road network.

We compared our algorithms, namely *Detect* and *Diffuse*, with the two state-of-the-art algorithms, namely *minDistort* [5] and PCA [6]. *Detect* is our deviation-based algorithm used to detect anomalies. *Diffuse* is our diffusion algorithm used to find the major causes of anomalies.

**4.2 Performance Study** In this section, we conducted experiments to compare our proposed algorithms (i.e., *Detect* and *Diffuse*) with the two state-of-the-art algorithms (i.e., *minDistort* and *PCA*).

In order to measure the goodness of the algorithms, we have to know the ground truth. In our experiments, we consider two types of events as the ground truth of traffic anomalies, namely a *traffic event*, which corresponds to either an accident or a traffic control event, and a *holiday event*, which corresponds a traffic anomaly due to public holidays. We collected the events from three sources: (1) Shenzhen News online edition (http://dtzbd.sznews.com/), (2) Wikipedia (http://zh.wikipedia.org/zh-cn/) and (3) Shenzhen Traffic Police on MicroBlog (http://e.weibo.com/shenzhenjiaojing). In total, we collected 30 traffic events and 50 holiday events.

For each event type, we partition the datasets into 3 parts for cross evaluation. Specifically, we use two of the parts as training data and the remaining part as testing data. Thus, our cross evaluation is three-fold.

We adopt *F1 score*, a common accuracy measure of a test, to measure the goodness of each algorithm. Specifically, F1 score is defined as follows.

$$\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4.9}$$

The experimental results are shown in Figure 4. We have the following observations. First, our algorithm *Diffuse* consistently performs the best for all event types in all folds of cross evaluation. This essentially shows that our new framework of detecting traffic anomalies and at the same time, finding the major causes for the detected traffic anomalies is superior over the existing methods. Second, our algorithm *Detect* (the one without being augmented with the "heat diffusion" technique) performs better than the existing methods for the holiday event in all folds of cross evaluation. This give us a clue that *Detect*, though simply a deviation-based method, is quite useful in practice. Third, there is a clear effectiveness gap between *Diffuse* and *Detect*. This shows clear justice of our novel idea of capturing the "diffusion phenomenon" of traffic anomalies with the "heat diffusion" model. We note here that the F1 scores of all algorithms are relatively low which is mainly due to the fact that our collected events correspond to the a small portion of real anomaly events only (In this case, the precision of each algorithm is usually quite small).

We also observe that our algorithm *Diffuse* usually performs better than the existing methods *PCA* and *minDistort* in terms of both precision and recall. For example, for the traffic event type, the average precisions (recalls) of *Diffuse*, *PCA* and *minDistort* are 0.095 (0.260), 0.057 (0.316) and 0.065 (0.150), respectively.

**4.3 Case study** In this section, we have two case studies about real world events. The following shows two prominent examples of known events.

**Event 1.** The first event is the "International Children's Day" (1 June of each year). Children's Day is recognized on different days in different countries to honor children.

**Event 2.** The second event is an "car accident in Huanggang flyover". This accident happened in Huanggang

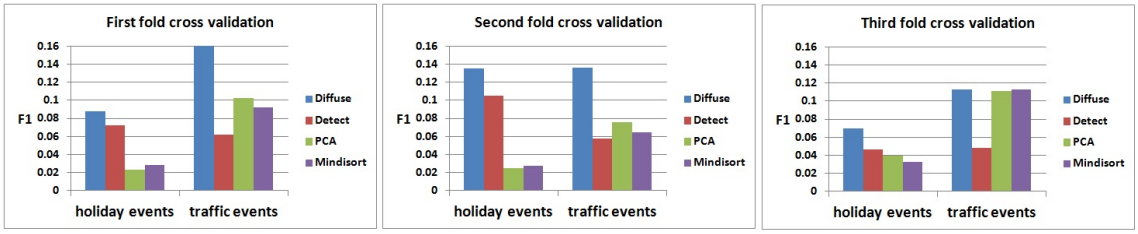
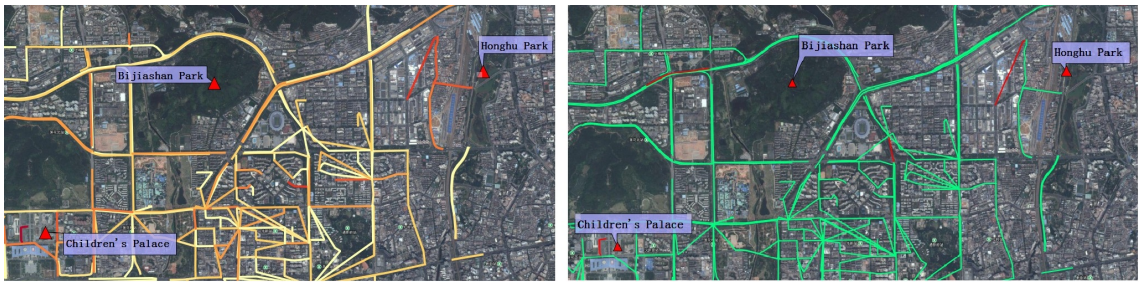



Figure 4: The test F1 of different algorithms using 3-fold cross validation for different event types




(a) (b)

Figure 5: Case Study 1: Results on Road Network between 9:00am to 9:30am, On June 1, 2012

flyover, North Ring of Shengzhen from 8:00am to 8:30am on January 16, 2012.

Figure 5 shows the results of for Event 1 (the International Children's Day), where the color of a road segment indicates the anomaly value of this segment. A darker color indicates that the corresponding road segment has a higher anomaly value. Figure 5(a) shows the result returned by *Detect*, which shows the real anomaly values of the road segments, and Figure 5(b) shows the result returned by *Diffuse*, which shows the major anomaly causes. As could be noticed, in Figure 5(b), the traffic of the road segments near to some children's parks on the International Children's Day were abnormal, which might be explained by the phenomenon that on the International Children's Day, many parents guided their children to parks, thus resulting in some abnormal traffic patterns.

Figure 6 shows the results for Event 2 (a car accident in Huanggang flyover). In the figure, we observe that on January 16, the traffic of "Huanggang flyover" are identified as abnormal, which can give us the information where the car accident happened (Figure 6(b)) and the information how the car accident affected the traffic of other road segments (Figure 6(a)).

## 5 Related Work

**5.1 Traffic Anomaly Detection** The increasing capability to track moving objects and a huge volume of spatio-temporal data leads to more studies on trajectory mining, especially outlier detection among traffic data. Sun et al. [13] proposed a measure, spatial local outlier measure (SLOM), to capture the local behaviour of trajectory data in their spatial neighbourhood. This measure takes into account the local stability around a spatial point and reports outliers in highly *unstable* areas. There are also studies about temporal outlier detection. One example is [14], which proposed Temporal Outlier Discovery (TOD) framework and utilizes agglomerated temporal information of the entire dataset to detect outliers which are determined based on drastic changes in the trends.

There are three closely related studies about traffic anomaly detection, namely [5, 6, 8]. In Section 1, we described the two-step framework for the first two related studies [5, 6]. The two algorithms in [5, 6], the two state-of-art algorithms, are *region-based approaches*. That is, they both first need to partition the whole space into a number of regions via all major road segments, and then find "unexpected" movements from regions to regions. In particular, [5] proposed an algorithm which detects Spatial-temporal outlier (STO) and infers its causal interaction, and [6] proposed a two-step mining and an optimization framework for interring the root cause of anomalies. The third closely related study is [8], which was an older study compared with [5, 6]. The algorithm in [8] is an *grid-based approach*. Instead of partitioning the whole space into regions via road segments used in [5, 6], [8] partitions the whole space into a number of *regular grids* in order to find "unexpected" abnormal traffic patterns in the data space. However, same as [5, 6], the algorithm in [8], focusing on finding abnormal traffic patterns based on grids, does not return abnormal traffic patterns on road segments, studied in this paper. Returning abnormal traffic patterns on road segments is more reasonable since all taxi trajectories studied in [5, 6, 8] must lie on road segments and thus the abnormal traffic patterns should be defined based on road


(a) (b)

Figure 6: Case Study 2: Results on Road Network between 8:00am to 8:30am, on January 16, 2012

segments.

**5.2 Heat Diffusion Model** To the best of our knowledge, we are the first to introduce the Heat Diffusion Model to simulate the diffusion processes of traffic anomalies. We briefly introduce the heat diffusion model and its applications as follows. Heat diffusion is a physical phenomenon such that in a medium, heat always flows a position with high temperature to a position with low temperature. Recently, heat diffusion-based approaches have been successfully applied in various domains such as classification and dimensionality reduction problems [15, 16, 17]. In [17], Ma et al. used the diffusion model to simulate the diffusion of product adoption for viral marketing in social networks. In this paper, we model the diffusion of anomalies as a process of heat diffusion.

## 6 Conclusion

In this paper, we propose a novel framework to detect and analyze the traffic anomaly. The proposed framework consists of two steps: 1) detecting the road segments with abnormal traffic, and 2) inferring the major causes of all anomalies on the road network. A deviation-based method and a diffusion-based model were introduced to finish the two steps respectively. With a large amount of GPS data collected from about 20 thousands taxi in Shenzhen, China over eight months, we conduct comprehensive experiments to validate our methods. The experimental results demonstrate that our methods are better than the state-of-the-art algorithms. In the future, we plan to discover more abnormal patterns from the trajectory data.

## Acknowledgements

We thank the support of Natural Science Foundation of China 91224006, 61003138 and 41371386, the Strategic Priority Research Program of the Chinese Academy of Sciences XDA06010202 and XDA05050601, 12th Five-Year Plan for Science & Technology Support 2012BAK17B01 and 2013BAD15B02. The research of Cheng Long and Raymond Chi-Wing Wong is supported by grant FSGRF14EG34.

## References

[1] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye, "Mining periodic behaviors for moving objects," in *KDD*, 2010, pp. 1099–1108.

[2] Z. Li, J. Han, B. Ding, and R. Kays, "Mining periodic behaviors of object movements for animal and biological sustainability studies," in *DMKD*, 2012.

[3] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Semantic trajectory mining for location prediction," in *GIS*, 2011, pp. 34–43.

[4] Y. Liu, L. Chen, J. Pei, and Q. Chen, "Mining frequent trajectory patterns for activity monitoring using radio frequency tag arrays," in *PerCom*, 2007.

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

[6] S. Chawla, Y. Zheng, and J. Hu, "Inferring the root cause in road traffic anomalies," in *ICDM*, 2012.

[7] Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou, "A taxi driving fraud detection system," in *ICDM*, 2011.

[8] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng, "On mining anomalous patterns in road traffic streams," in *ADMA (2)*, 2011, pp. 237–251.

[9] Y. Bu, L. Chen, A. W.-C. Fu, and D. Liu, "Efficient anomaly monitoring over moving object trajectory streams," in *KDD*, 2009, pp. 159–168.

[10] J.-G. Lee, J. Han, and X. Li, "Trajectory outlier detection: A partition-and-detect framework," in *ICDE*, 2008, pp. 140–149.

[11] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," in *VLDB*, 2006, pp. 187–198.

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

[13] P. Sun and S. Chawla, "On local spatial outliers," in *ICDM*, 2004, pp. 209–216.

[14] X. Li, Z. Li, J. Han, and J.-G. Lee, "Temporal outlier detection in vehicle traffic data," in *ICDE*, 2009, pp. 1319–1322.

[15] J. D. Lafferty and G. Lebanon, "Diffusion kernels on statistical manifolds," *Journal of Machine Learning Research*, vol. 6, pp. 129–163, 2005.

[16] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, 2003.

[17] H. Ma, H. Yang, M. R. Lyu, and I. King, "Mining social networks using heat diffusion processes for marketing candidates selection," in *CIKM*, 2008, pp. 233–242.