SafeDrive: Online Driving Anomaly Detection From Large-Scale Vehicle Data

Mingming Zhang, Chao Chen, *Member, IEEE*, Tianyu Wo, *Member, IEEE*, Tao Xie, *Senior Member, IEEE*, Md Zakirul Alam Bhuiyan, *Member, IEEE*, and Xuelian Lin

***Abstract*—Identifying driving anomalies is of great significance for improving driving safety. The development of the Internet-of-Vehicle (IoV) technology has made it feasible to acquire big data from multiple vehicle sensors, and such big data play a fundamental role in identifying driving anomalies. Existing approaches are mainly based on either rules or supervised learning. However, such approaches often require labeled data, which are typically not available in big data scenarios. In addition, because driving behaviors differ under vehicle statuses (e.g., speed and gear position), to precisely model driving behaviors needs to fuse multiple sources of sensor data. To address these issues, in this paper, we propose *SafeDrive*, an *online* and *status-aware* approach, which does not require labeled data. From a historical dataset, *SafeDrive* statistically offline derives a state graph (SG) as a behavior model. Then, *SafeDrive* splits the online data stream into segments and compares each segment with the SG. *SafeDrive* identifies a segment that significantly deviates from the SG as an anomaly. We evaluate *SafeDrive* on a cloud-based IoV platform with over 29 000 real connected vehicles. The evaluation results demonstrate that *SafeDrive* is capable of identifying a variety of driving anomalies effectively from a large-scale vehicle data stream with an overall accuracy of 93%; such identified driving anomalies can be used to timely alert drivers to correct their driving behaviors.**

***Index Terms*—Anomaly, big data, data stream, driving behavior, Internet-of-Vehicles, on-board diagnostics (OBD), state graph (SG).**

I. Introduction

IDENTIFYING abnormaldriving behaviors is known to be an important research focus due to its significant influence on people's daily life. Apart from the impact on fuel consumption [1], driving behaviors also play a key role in transportation safety [2]. With in-vehicle sensing and Internet-of-Vehicle (IoV) technologies, we are capable of collecting abundant driving data, such as speed and engine parameters, from a large number of vehicles. Such data are characterized as large volume, multifrequency, and multisource, which largely reflect the vehicle status and thereby are widely used to evaluate driving behaviors. For example, insurance companies provide a new "pay-as-you-drive" service to customers by collecting their driving data and judging their driving behaviors [3]. With the collected data, fleet-operating companies regulate their drivers to behave more properly, lowering the accidental risk and fuel consumption.

These applications have inspired previous research on identifying driving anomalies. Rule-based techniques are often adopted to extract abnormal driving behaviors due to its simplicity and high efficiency [4], [5]. Supervised-learning-based techniques are another kind of popular solution. With predefined abnormal patterns and manually labeled training data, a classifier can be trained and further used to identify similar patterns [6], which are marked as anomalies. Such techniques basically require manually labeled training data, where fixed behavioral definitions such as patterns and rules (e.g., fast acceleration) need to be predefined.

However, the prior research cannot effectively identify abnormal driving behaviors for three main reasons. First, in IoV, the volume of data is huge. They are collected from multiple sensors and are with complicated relations, making it infeasible to label normal and abnormal behaviors. Second, the process of manually labeling the huge volume of the data stream can be difficult and biased because abnormal driving behaviors can be uncertain and human perceptions can be error-prone. Third, whether a driving behavior is abnormal or not is heavily dependent on the current vehicle status (e.g., speed and gear position). For instance, Fig. 1 shows the relationship between acceleration behaviors and vehicle speed statuses. It can be observed that drivers would normally accelerate more slowly when driving at high speed. Such behavior is a kind of contextual-status-related behaviors. As a comparison, another behavior can be observed in the relations between different types of data, and such behavior is a kind of correlational-status-related behaviors.

Manuscript received December 24, 2016; accepted January 12, 2017. Date of publication February 24, 2017; date of current version August 1, 2017. This work was supported by the China Key Research and Development Program under Grant 2016YFB0100902, the National Natural Science Foundation of China under Grant 61602067 and Grant 61421003, Fundamental Research Funds for the Central Universities under Grant 106112015CDJXY180001. The work of T. Xie was supported in part by the Natural Science Foundation under Grant CCF-1409423, Grant CNS-1434582, Grant CNS-1513939, and Grant CNS-1564274. Paper no. TII-16-1574. *(Corresponding author: T. Wo.)*

M. Zhang, T. Wo, and X. Lin are with Beihang University, Beijing 100191, China (e-mail: zhangmm@act.buaa.edu.cn; woty@act.buaa.edu.cn; linxl@act.buaa.edu.cn).

C. Chen is with the College of Computer Science, Chongqing University, Chongqing 400030, China (e-mail: cschaochen@cqu.edu.cn).

T. Xie is with the Department of Computer Science, University of Illinois at Urbana–Champaign, Champaign, IL 61801 USA (e-mail: taoxie@illinois.edu).

M. Z. A. Bhuiyan is with the Department of Computer and Information Sciences, Fordham University, New York, NY 10023 USA (e-mail: mbhuiyan3@fordham.edu)

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

Digital Object Identifier 10.1109/TII.2017.2674661

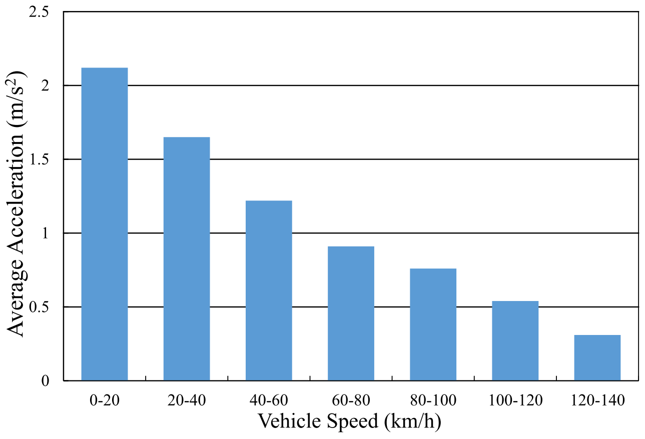


Fig. 1. Accelerations when driving at different speed ranges. It can be observed that acceleration of 1˜m/s^2 when the vehicle speed exceeds 100˜km/h would be abnormal, while the same acceleration would be normal when the speed is lower than 60˜km/h.

To effectively identify driving anomalies requires considering such detailed status-related characteristics. Moreover, the evaluation criterion for anomaly detection should be based on objective data instead of subjective judgment. To address the preceding issues, in this paper, we propose *SafeDrive*, an online and status-aware approach for detecting driving anomalies. *SafeDrive* does not require costly labeled data, by employing a state graph (SG). *SafeDrive* fuses data on both the vehicle-sensor level and the fleet level; such data precisely reflect the normal driving styles. For the online detection, *SafeDrive* compares the real-time driving data stream with the SG to detect anomalies. *SafeDrive* includes novel techniques to address two main challenges: 1) uniformly modeling a variety of vehicle statuses represented by complex data relations; and 2) capturing how people normally drive based on the modeled relations between statuses.

In particular, *SafeDrive* includes an SG to model 1) contextual relations between statuses of the same type of data, such as speed, at different timings and 2) correlational relations between statuses of different types of data, such as the vehicle revolutions per minute (RPM) and gear position, at the same timing. *SafeDrive* represents all the statuses as states and connected with edges in the graph. To construct an objective driving model, *SafeDrive* fuses different vehicles' historical data together and statistically calculates the structure of the SG. In the online setting, *SafeDrive* identifies driving anomalies by splitting the data stream into segments and comparing each segment with the SG. *SafeDrive* considers as abnormal those segments that largely deviate from the SG.

We implement *SafeDrive* on a real-world cloud-based IoV platform, which connects over 29 000 vehicles from 60 cities. Each vehicle is equipped with an on-board diagnostics (OBD) connector to collect the vehicle's parameter values and send the data to the server through the mobile wireless network. The evaluation results demonstrate that *SafeDrive* is able to effectively identify driving anomalies including aggressive acceleration, sudden braking, fast turn, and even mismatching of RPM with speed.

In summary, this paper makes the following main contributions.

1) A status-aware behavior model, which is able to combine multisensor data of different vehicles, to characterize normal driving behaviors quantitatively.
2) A lightweight online anomaly detector for detecting a variety of abnormal driving behaviors from large-scale vehicle data.
3) An implementation of *SafeDrive* upon a large-scale cloud-based IoV platform, and an evaluation of *SafeDrive* with a huge volume of real-world driving data.

The remainder of this paper is organized as follows. Section II summarizes related work. Section III presents an overview of the proposed *SafeDrive* approach. Section IV illustrates *SafeDrive*'s online detection of driving anomalies. Section V presents our evaluation of *SafeDrive*. Section VI concludes this paper.

II. Related Work

Safe driving is one of the major public concerns, and identifying abnormal driving behaviors is an indispensable part of improving driving safety [2], [7]. In recent years, various techniques have been proposed to detect driving anomalies.

Rule-based techniques employ thresholds to filter out data of specific ranges and mark these data as driving events. These techniques are often adopted in previous work as system solutions and basic behavior-evaluation mechanisms due to their simplicity and efficiency [4], [5]. For example, using sensor data collected from smartphones, Zhao *et al.* [8] detect aggressive driving events based on thresholds. To detect drunk driving, Dai *et al.* [9] propose a pattern-matching algorithm that compares acceleration with predefined drunk driving thresholds. Fazeen *et al.* [10] combine rule-based behavior analysis with road-condition evaluation to construct a smartphone-based safe driving system. Moreover, Taha and Nasser [11] propose a threshold-based framework to evaluate the driving behaviors from controller area network (CAN-bus) data collected by OBD connectors.

Supervised-learning-based techniques are another kind of widely adopted techniques. By using labeled data, a classifier can be trained and further used to predict unlabeled data. Specifically, Chen *et al.* [4] propose a classifier based on a support vector machine (SVM) to recognize abnormal driving styles, such as swerving and fast U-turn, from smartphone sensors in real time. Quintero *et al.* [12] propose a technique based on an artificial neural network to detect erratic driving from OBD and GPS data, and the model is evaluated on a driving simulator. Hong *et al.* [6] propose a technique of data fusion by combining accelerometer data with OBD data and using a naive Bayes classifier to identify aggressive driving behaviors. Jaramillo and Narvez [13] propose an online monitoring system based on a fuzzy clustering algorithm.

In addition, Johnson and Trivedi [14] use dynamic time warping to detect aggressive driving using smartphone sensor data. Li *et al.* [15] construct a driving analysis system via operation-mode classification. Miyajima *et al.* [19] propose a Gaussian mixture model to model driving behaviors and further to identify drivers. As foundational research, Constantinescu *et al.* [17]

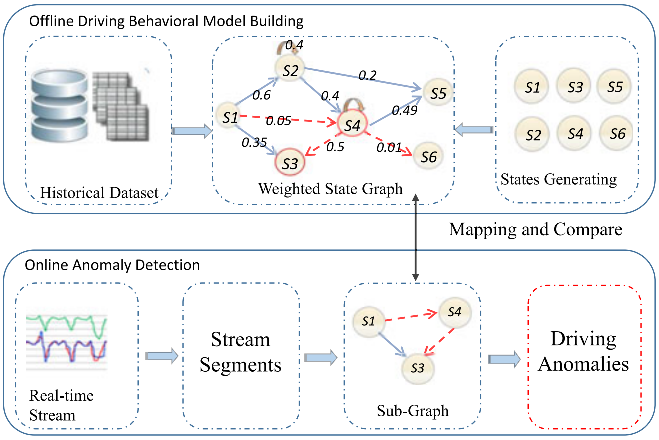


Fig. 2. Overview of *SafeDrive*.

investigate driving-style categories with a clustering algorithm. Bolovinou *et al.* [16] survey techniques of driving-style recognition for cooperative driving. Banerjee *et al.* [22] propose an algorithm named skill-aggression-quantifier (SAQ) to evaluate driving behaviors. They also implement a tool named *MyDrive* based on the SAQ algorithm. Lei [23] designs a framework to detect anomalies in trajectory behaviors, which can be considered as another kind of driving behaviors. There are also various other techniques [20], [21] in this research area.

In summary, although there exist various previous techniques [24]–[26] of unsupervised or semisupervised anomaly detection, in driving-analysis scenarios, previous work is mainly based on the definition of driving behaviors, such as rules or patterns. To identify a specific type of driving anomalies, it is usually necessary to define anomaly patterns and prepare labeled data first. However, in a real-world IoV scenario, such labeled data are not available because the data are collected automatically, and thus, manually labeling driving styles is not applicable. In addition, driving behaviors are status aware, and the behavioral model should be able to reflect detailed characteristics of driving. Therefore, based on an IoV system and the huge volume of collected data, in this paper, we propose *SafeDrive*, an online, data-driven, and status-aware approach for driving-anomaly detection.

III. Overview of *SafeDrive*

Modeling of driving behaviors plays a fundamental role for detecting driving anomalies but is quite challenging. As stated earlier, to detect driving anomalies, a behavioral model should be able to 1) cover variable relationships of driving data and 2) reflect driving styles quantifiably. From the data perspective, the model should reflect the relationships of different types of data and their patterns. To that end, we propose an SG-based behavioral model, as discussed in detail in this section.

The overview of *SafeDrive* is shown in Fig. 2. The model contains two main parts: the offline building of a driving behavioral model and the online detection of driving anomalies. Overall, we use the model built offline based on the historical data to online identify the newly arrived data stream.

A. Offline Building of a Driving Behavioral Model

Our basic idea is to uniformly model the status relations of streamed vehicle data in a weighted SG, in which the state is a term used to represent the value (or its range) of data attributes. Specifically, we adopt discrete states to quantify status (data values) and employ weighted edges (connections between states) to measure the relationship between states. The states are generated from different sensor data and connected with each other via weighted edges, in which manner the model can combine multiple data, even those with different frequencies. The structure of the graph is constructed based on statistics of the historical data; as a result, the graph becomes a detailed behavioral model for fusing different vehicles' data. In this way, the graph structure can objectively reflect how people usually drive under different conditions or statuses since the weights are generated from real-world data. The formal definition of an SG is given in Definition 1.

Definition 1 (SG): An SG $=< S, E >$ is a weighted directed graph, where S is a set of states and E is a set of weighted edges. A weighted edge $e \in E$ corresponds to a kind of relation between states, where weight $w \in (0, 1]$.

B. Online Anomaly Detection

In many cloud-based applications, the collected driving data are often organized as a stream or data sequence where each data instance contains different attributes. The stream, as stated earlier, reflects driving behaviors (related to contextual and correlational statuses) where anomalies may occur. To detect anomalies online, we first split the newly arrived data instances in data stream into segments and then map each segment as a temporal subgraph (TS-SG), which is further evaluated by being compared with the offline-generated SG model. The subgraph (or segment) that significantly deviates from the SG model is considered as an anomaly. The formal definition of a temporal subgraph is given in Definition 2.

Definition 2 (TS-SG): A TS-SG $=< S^*, E^* >$ is a temporal subgraph of an SG, generated by a data subsequence, where $S^* \in S$ and $E^* \in E$. A specific state may repeatedly, with different time stamps, appear in a TS-SG.

Formally, in *SafeDrive*, a state $s \in S$ represents a category or a set of data instances. For numerical data, states can be acquired by discretization, while for categorical data, the category itself can be used as a state. In our evaluation (described in Section V), for example, the vehicle speed is generated into 100 states, each of which covers a speed range of ± 1 km/h. Acceleration and deceleration behaviors can be reflected by the transitions between those speed states. Therefore, the abnormal level of acceleration behaviors can be evaluated by the connection weight from speed states with a smaller value to states with higher values. We regard the edge between the same type of states as the contextual edge, which models the contextual driving behaviors reflected by the same kind of OBD data parameters. Note that for numerical data attributes, a potential risk of discretization is state explosion: a huge number of states might be generated and thus cause an extremely large graph. However, in most of the real-

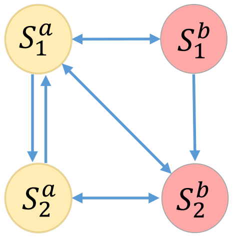


Fig. 3. Example of an SG, in which bidirectional arrows represent correlational edges and unidirectional arrows represent contextual edges.

world cases, only a limited scope of ranges or data sources are used in practice, largely limiting the scale of the graph.

In *SafeDrive*, different parameters are separately generated into different types of states. The co-occurrence relations between different types of data reflect the correlational-state-related behaviors. For instance, the speed has a correlation with RPM, e.g., a high RPM value usually implies high vehicle speed. To strengthen the expression ability of *SafeDrive*, we use correlational weighted edges to represent the co-occurrence relationships between different types of states. As a result, two kinds of edges are enclosed in the SG, i.e., contextual edges and correlational edges. Fig. 3 illustrates an abstract example of a SG with two types of states, S^a and S^b. Note that in *SafeDrive*, the SG may contain cycles since the data in a stream can be repeatable.

As discussed earlier, the SG behavioral model is built by two steps: state generation and graph construction. The first step uses discretization to transfer data ranges of OBD parameters into states, while the second step scans the historical dataset statistically, from which the edges between states and their weights are derived and calculated. The value of the weight of a contextual edge is computed by (1), where t is a time stamp referring to the relative temporal relationship between states. The value of connection weight denotes the conditional probability of s_2 appearing at time $(t+1)$ when s_1 appears at t

$$w(s_1, s_2) = p(s_2(t+1)|s_1(t)). \tag{1}$$

For a correlational relationship between two different types of states s^a and s^b, with the objective of presenting a detailed reflection of the correlation, we implement it with two conditional edges separately, i.e., from s^a to s^b and from s^b to s^a. Their weights can be calculated according to (2). As can been seen in (2), the values of connection weights denote the probability of s^b appearing at time *t* given the condition of s^a appears at t, and the probability of s^a appearing at time t given the condition of s^b appears at t. Note that the correlational edges between s^a and s^b are asymmetric since $w(s^a, s^b)$ is usually not equal to $w(s^b, s^a)$

$$\begin{aligned} w(s^a, s^b) &= p(s^b(t)|s^a(t)) \\ w(s^b, s^a) &= p(s^a(t)|s^b(t)). \end{aligned} \tag{2}$$

The finally realized model includes 308 states: 100 states for driving speed from 0 to 200 km/h, 100 states for engine RPM from 0 to 5000, 100 states for swing angle from 0° to 360°, and the remaining eight states for gear positions.

Building an SG model that combines different types of data makes it feasible to model a variety of driving behaviors. For example, the speed states and their connections reflect acceleration and deceleration behaviors, while the connections between states of speed and gear position reflect the combined control behaviors of vehicle speed and gear position.

IV. Online Abnormal Detection With *SafeDrive*

Typically, we evaluate driving behaviors for each short period of time. To that end, after the SG is built, we use it to measure the online stream data. As shown in Fig. 2, the stream is split into segments, each of which is a behavior unit and being mapped to a TS-SG. The segmentation length is based on the time duration for completing a behavior. Our empirical investigation suggests that an interval of 10–15 s is suitable to represent a driving behavior.

Unlike other subgraphs, a TS-SG contains contextual information of data stream. In such a graph, a state is allowed to appear repeatedly given that specific data are likely to be generated repeatedly. For example, when driving in a stable status, many of the sampled speed data in the uploading stream might be the same; hence, the TS-SG may contain some recurring states with different time stamps.

The states of TS-SG are generated in the same way that the states of the SG are generated. The edges in TS-SG also have weight values assigned according to their counterpart edges in the original SG. For example, if there is an edge from s^a to s^b in a TS-SG, then its weight value equals to the value of $w(s^b, s^a)$ in the SG. Specifically, if no such edge exists in the SG, the graph would be updated by *SafeDrive* automatically

$$f(\text{TS-SG}) = \frac{1}{m} \sum_{s_i, s_j \in S^*} w(s_i, s_j)^{-1}. \tag{3}$$

After the TS-SG of each segment is generated, according to (3), we compute an anomaly score for the subgraph TS-SG, marked as f(TS-SG). Given that the aim of the score is to filter out those state connections with low probabilities, we hereby employ an inverse proportional function to construct f(TS-SG). In this manner, we are able to amplify low probabilities and filter them out. Note that m is the number of edges in the subgraph. Basically, we consider the TS-SG with low-probability edges, which usually cause a high value of f(TS-SG), as an anomaly. The score is compared with δ, a threshold defined manually. If $f(\text{TS-SG}) > \delta$, then the subgraph is marked as an anomaly. In practice, it is suggested to choose the threshold δ according to the distribution of score f(TS-SG).

Fig. 4 shows two abstract examples of the subgraph with different types of anomalies caused by contextual and correlational relationships, respectively. In real-world driving scenarios, contextual and correlational driving anomalies may occur simultaneously because the data are generated by vehicle components closely working together, and a specific abnormal driving behavior or operation may cause various anomalies.

By analyzing the structure of the abnormal TS-SG and evaluating which kind of edge causes a high score value, data analysts are able to understand the detailed reason for this anomaly. For example, if the cause of high f(TS-SG) is vehicle-speed state transition, it signifies that the driver behaves not so well in accelerating or decelerating. While if

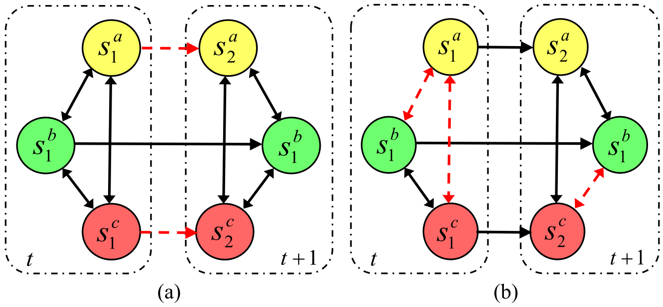


Fig. 4. Temporal subgraph that contains an anomaly, where a red-dashed line represents an abnormal connection with low probability. (a) Contextual anomaly. (b) Correlational anomaly.

Algorithm 1: Online Anomaly Detection Algorithm.

```
Input: (1) $DS$;(2) $SG$.
Output: (1) $Anomaly$.
$SegmentSet \leftarrow$ Split{$DS$}
//split the data sequence into segments
for each Segment $Seg$ in $SegmentSet$ do
    $TS$-$SG \leftarrow$ Match{$Seg$, $SG$}
    matching the $Seg$ with $SG$, generate a $TS$-$SG$
    calculate $f(TS - SG)$
    if $(f(TS - SG) > \delta)$
    Output $Seg$ and $TS$-$SG$ as an anomaly
end for
```

RPM states cause the high score, it signifies that the driver does not take a smooth control of accelerator pedal or gear position, suggesting that the driver drives either aggressively or unskillfully.

Note that due to the limitation of historical data, change of environment, or people's driving styles, the structure of the graph may need to be able to evolve over time. Such a characteristic is known as concept drift in anomaly detection for streaming data [18]. In this scenario, for example, given two states, s_i and s_j, when measured in different times, $w(s_i, s_j)$ could be different. As a result, the abnormal level of sequence s_i, s_j changes over time. Failing to sense or account for such change could lower the performance of the detector by causing many false alarms. We address this problem by designing a module in the cloud named *SG-Maintainer* to maintain and update the SG. In practice, the *SG-Maintainer* maintains an array that records the connection number between states. It updates the array when each data arrives and then periodically calculates the connection weights of the *SG* according to the array.

Given a data sequence DS, the online detection is described in Algorithm 1.

V. Evaluation

We comprehensively evaluate the effectiveness and efficiency of *SafeDrive*. In this section, we first present the data description used in the evaluation, followed by a detailed category analysis of the detected abnormal driving behaviors. We quantitatively evaluate the detection accuracy and computational performance of *SafeDrive* and compare *SafeDrive* with other related approaches.

TABLE I
Data Specification

Name	Type	Range	Description
Speed	Numerical	0–200 km/h	Vehicle speed
RPM	Numerical	0–8000	Engine round per minute
Swerving	Numerical	0–270°	The change of vehicle direction
Gear position	Enum	Eight positions	Gear position

A. System and Data Description

We evaluate the performance of *SafeDrive* on a real-world IoV system. The system is designed as a cloud-based IoV architecture, in which driving data are collected with OBD devices plugged in the vehicles. Each OBD device has integrated a wireless communication module to maintain connections with the back-end server and send the collected data to the server with an adjustable time interval. Over 29 000 real vehicles from 60 cities have been connected to the system. This system collects around 0.2 billion data instances daily.

Table I lists the details of the data attributes used in the evaluation, including speed, RPM, swerve angle, and gear position. The OBD connector is capable of sampling various attributes from vehicles, such as the door status or brake pedal status, and we are aware that employing more parameters may improve the analysis performance. However, due to the fact that different vehicle manufacturers use different CAN-Bus protocols, it is not easy to collect all the parameters from all the vehicles in the fleet. Hence, to assure the applicability of the learned model, we construct it based on attributes that could be collected from almost any kind of vehicles. Also, such attributes are considered as directly influenced by driving behaviors, and they can reflect driving behaviors to a large extent. We use these attributes to evaluate the lateral and longitudinal dynamic of a vehicle. Note that the swerve angle is not directly collected but is calculated based on position data collected from the GPS module embedded in the OBD connector.

The vehicles in this system belong to a chauffeur company and the drivers are hired after a strict selection, most of whom are experienced and well trained. Thus, we assume that the behaviors of most of the drivers, under most of the situations, are normal. Therefore, it makes sense to use the SG generated in this system as a benchmark for evaluating abnormal driving. In the training phase, the data collected in the first month are used to construct the graph, and then, the generated SG is deployed to analyze the data of subsequent months.

However, note that in other driving scenarios, it is possible that *SafeDrive* may ignore some unsafe driving styles if many drivers perform unsafe behaviors habitually. For example, according to a report [27], nearly 35% American drivers are aggressive. Therefore, some aggressive driving styles might be identified as normal by our learned model. In such a situation, the training data should be collected in a manner such that the training data represent only nonaggressive driving styles. Hence, it is suggested to collect the training data with selected drivers driving under safe instructions.

SafeDrive is initially implemented as a cloud service because the infrastructure of the system is designed and implemented with the cloud so that data can be collected. Therefore, the analysis of driving behaviors is conducted in the cloud. However, it is known that driving alert is safety critical and the application may suffer from network delay; thus, to acquire a faster reaction, the model is further implemented into a smartphone application. The smartphone is supposed to be in the vehicle and maintains communication with the OBD connector to sample data in a higher frequency, and with the cloud to update its model. In this manner, the back-end server takes charge of collecting driving data from the fleet and updating the SG model, which is updated to the smartphone application periodically.

The system is running on a cluster with 25 computation servers, each of which is equipped with 16-GB memory and a quad-core processor. The data stream is uploaded from each vehicle and collected by two TCP servers. Then, the data are loaded into a distributed data bus system. The streaming computing system takes five of the servers. The information from the cluster and application monitoring shows that the real-time data uploaded by the vehicles are easily handled by the system, and the CPU and memory use ratio maintains lower than 20%. To further assess the computational cost of *SafeDrive*, we replay the historical data with a much higher ratio on a personal computer with quad-core Intel processor and 8-GB memory, and the data are processed by *SafeDrive* implemented with Java. The simulation result suggests that the model is able to process millions of data instances per second on that single computer, indicating that *SafeDrive* has a potential to be employed to deal with large-scale IoV scenarios.

B. Anomaly-Category Analysis

The detection results are classified and analyzed based on driving behavioral semantics. As previously stated, *SafeDrive* calculates an anomaly score for each sequence segmentation based on its inner data relations and then identifies anomalies by comparing the score with a threshold. The evaluation itself does not provide a semantic description for the detected abnormal behaviors. Therefore, to better understand what kind of behaviors *SafeDrive* is capable of identifying, we manually classify the results according to the structure of the abnormal TS-SG. Table II lists the categories of abnormal driving behaviors identified by *SafeDrive*. Seven kinds of abnormal driving behaviors can be detected.

SafeDrive is a status-aware anomaly detector in that it evaluates driving behaviors under a specific status. The connection weight of the SG is actually the condition probability. The detected anomalies are classified into two categories: contextual anomalies and correlational anomalies. This classification is based on the type of edge that causes a high value of $f(\text{TS-SG})$. If a contextual edge causes a high score, then the anomaly is a contextual anomaly. If a correlational edge causes a high score, then the anomaly is a correlational anomaly.

TABLE II
ABNORMAL DRIVING BEHAVIORS DETECTED BY *SafeDrive*

Anomaly Behavior	Corresponding Anomaly in the TS-SG
Rapid Acceleration	Speed states with small value connect to states with large value and the connection weights are low.
Sudden Braking	Speed states with large value connect to states with small value and the connection weights are low.
RPM-Speed Mismatching	RPM states with large value connect to speed states with small value and the connection weights are low.
Over speed	Speed states with extremely high value occurs. These “rare states” often have very low connection weight With Other states.
RPM Anomaly	RPM states with extremely high value occurs (which are rare) and have very low connection weight with other states.
Rapid Swerving	Swerving states with large value connect to speed states with large value and the connection weights are low.
Neutral Taxiing	Neutral gear position states connect to speed states with value larger than 0 km/h.

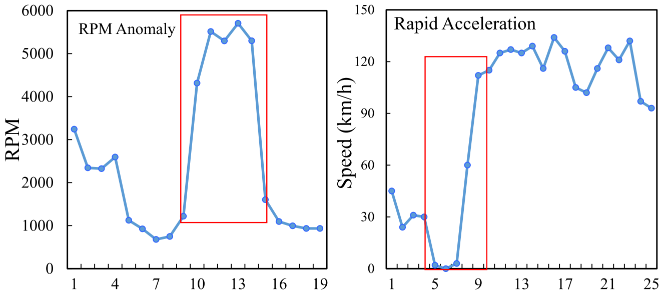


Fig. 5. Examples of contextual driving anomalies.

SafeDrive learns an unsupervised model and automatically detects anomalies, but the detection results provide no semantic description beyond true or false, limiting the practical use of the system. Therefore, to provide readable warning information to drivers, based on Table II and manual analysis, we use a rule set to translate the detected anomalies into their corresponding semantic explanations. For example, a rapid acceleration warning is given if an anomaly is caused by the transition from a speed state with a small value to a speed state with a large value. This section provides a number of anomaly examples and discusses how they are detected. Note that the horizontal axis in Figs. 5 and 7 (for showing the anomaly examples) represents the relative index of the data instances, reflecting their temporal relationships in the stream.

1) Contextual Abnormal Behaviors: In the uploaded data stream, the value of RPM or speed is a behavioral attribute and the time is a contextual attribute. Driving anomalies are identified by evaluating the behavioral attributes under a specific context. For driving evaluation, the behavioral attributes under different contexts are quantified and expressed in an SG, and hence, it is reasonable to apply this model to identify contextual anomalies.

Fig. 5 illustrates two sequences with detected RPM and acceleration anomalies, as marked by the red boxes. Rapid acceleration or deceleration (sudden braking) is the most straightforward

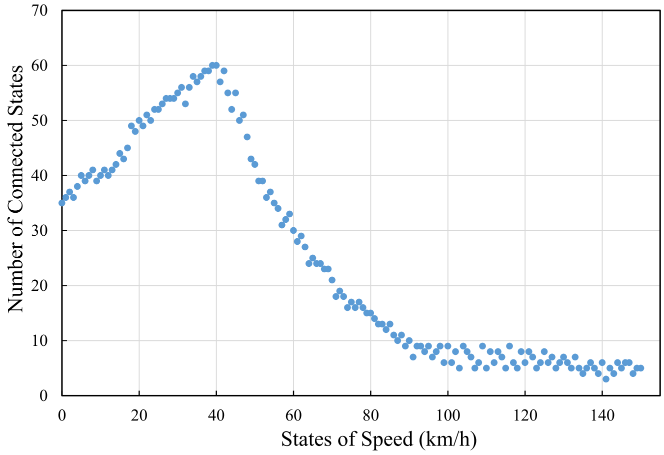


Fig. 6. Connection number of each speed state.

contextual anomaly. They are detected by *SafeDrive* for the reason that, given a current speed state, its connection weight with a much higher speed state is small, causing a higher anomaly score. These driving styles are considered as anomalies and are not advocated because they could cause more fuel consumption and increase vehicle-component wear. Most drivers usually do not adopt such driving styles.

The RPM anomaly shown on the left-hand side of Fig. 5 suggests that the engine raves are extremely high, being unusual and considered as an anomaly. These high values correspond to "rare states" in the SG because the probability of incurring such states (values) is extremely low, meaning that only a few states are connected with those states and the edges to them have low connection weights. As discussed earlier, the SG has 100 initialized RPM states and does not contain a value exceeding 5000. For the implementation, when these rare states occur, they are inserted into the graph by the *SG-Maintainer* module in the case of concept drift. Although they are inserted into the graph, still these rare states do not have a close relationship with other states. The rare RPM states with high values are sensitive to contextual attributes, further causing them to be identified as an anomaly. *SafeDrive* detects several overspeed anomalies exceeding 150 km/h for the same reason. Fig. 6 shows the number of connections of each speed state. It can be seen that states of higher speed tend to have sparse connections with other states in the graph. This sparsity characteristic is the main reason why the anomalies are detected.

SafeDrive is able to detect this kind of driving anomalies, whereas other previous approaches have to rely on a comprehensive rule set to accomplish such detection. Still, future work, such as introducing information of location and road network into *SafeDrive*, is required to improve detection of anomalies such as over speed, because the dangerous level of such anomalies varies depending on the specific road condition.

2) Correlational Abnormal Behaviors: Correlational behaviors exist between coevolving or correlational sequences, such as (speed, RPM) and (speed, gear position). A correlational anomaly can be detected when the data deviate the relationships between the data sequences. As listed in Table II, there are three kinds of correlational anomalies detected by *SafeDrive*, including rapid swerving, RPM-speed mismatching, and neutral taxiing. *SafeDrive* is also able to identify correlational abnormal behaviors by jointly evaluating two or more types of data attributes.

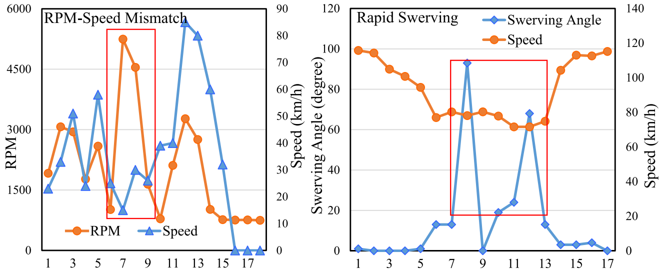


Fig. 7. Examples of correlational driving anomaly.

A representative correlational driving anomaly in this application is rapid swerving, which could be identified by speed and swerving angle. For swerving data, due to a high possibility of making a turn for a vehicle running on an urban road network, the extent of swerving angle change by itself would not provide much value in anomaly detection. But by combining such information with vehicle speed, *SafeDrive* can detect a rapid serving anomaly.

Fig. 7 shows examples of detected RPM-speed mismatching and rapid swerving anomalies. It can be seen from the mismatching anomaly that the vehicle RPM is too high for the corresponding vehicle speed. RPM-speed mismatching occurs when the vehicle speed is low and the driver pushes the gas pedal aggressively. Fast acceleration usually happens with this behavior; however, depending on specific road conditions or gear positions, the vehicle speed does not always dramatically increase, causing this RPM-speed mismatching. For the rapid swerving anomaly, the vehicle continuously takes two turns; before that, the vehicle was running at a fast speed of nearly 100 km/h. Although the driver slows down the vehicle in advance, the speed is still too fast for turning over 100°. In the SG model, most of the high-speed states are connected with lower swerve angle states, causing *SafeDrive* to identify those fast turn anomalies.

For a neutral taxiing anomaly, the gear is put in a neutral position, while the vehicle is still running at a relatively high speed. Such an anomaly is another representative correlational driving anomaly that can be detected by *SafeDrive*. This behavior usually occurs on the downhill path or straight road when the vehicle speed is high. Some drivers' driving exhibits such behavior because they think it is a fuel-efficient way of driving. But, in fact, neutral taxiing is not fuel efficient for most of the automatic transmission vehicles, and it could also be dangerous because it increases the braking distance when danger happens. This behavior can hardly be detected only by speed data; however, by combining speed data with gear position, the detection would be much easier. In the SG model, the gear position states with higher values tend to connect with vehicle speed states with higher value, while normally neutral position states only connect with the speed states with values lower than 10 km/h.

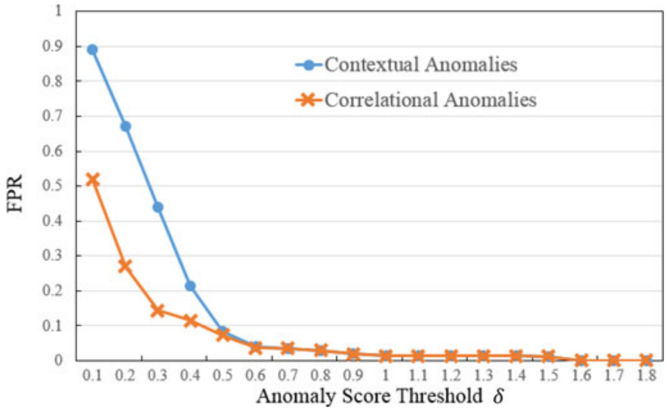


Fig. 8. FPR under different threshold values.

In the evaluation, note that some abnormal behaviors, such as fast acceleration and RPM anomalies, could occur simultaneously because the data are generated by highly correlated vehicle components. Due to the complexity of driving environments, some behaviors could also occur alone.

C. Quantitative Evaluation

To understand the applicability of *SafeDrive*, we manually analyze the detection results of an arbitrarily chosen week. Fig. 8 shows the false positive rate (FPR) evaluation of the model on detecting contextual and correlational anomalies under different score threshold δ. FPR is the probability that a normal behavior being detected as abnormal. A long tail effect can be observed from the curves as the FPR decreases fast with the increase of threshold δ. When setting a smaller threshold δ, many false alarms would be raised by *SafeDrive* due to the fact that most of the behaviors are normal with a low score. As shown in the curve, when the threshold is relatively small, the FPR of correlational anomaly detection is lower than contextual anomaly detection. Such a result shows that the discrimination between normal and abnormal correlational behavior is higher than that of contextual behavior.

The trend of the FPR curve with threshold also shows the distribution characteristic of the anomaly score. Therefore, in practice, it is suggested to set the threshold according to statistical principles. For example, let $\delta > \mu + 3\sigma$, where μ is the mean value of anomaly score and σ is the standard deviation of the score.

Many industrial solutions use rule-based techniques to address the problem of driving-behavior detection. Therefore, in our evaluation, we first compare the performance of *SafeDrive* with a rule-based approach, which is frequently being used as an industrial solution. In the rule-based approach, the used rule set contains a number of rules that define the outlier threshold such as acceleration $> 1.5\,\text{m/s}^2$.

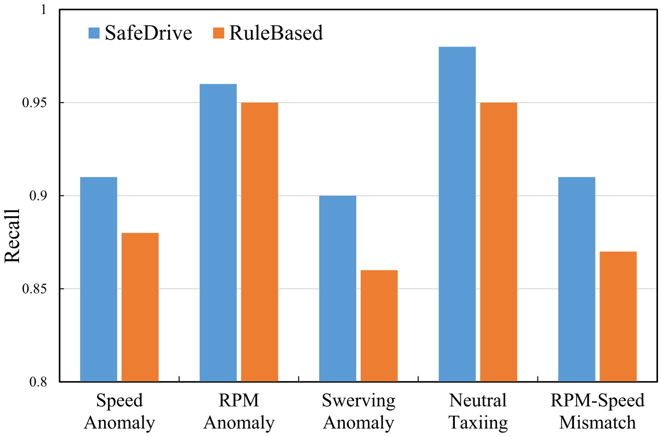


Fig. 9. Recall evaluation for different types of anomalies.

TABLE III
PRECISION OF HMM-BASED DETECTOR AND *SafeDrive*

Model	Contextual	Correlational	Average Precision
SafeDrive	94.0%	92.0%	93.0%
HMM	90.0%	88.0%	89.0%
HMM	85.0%	84.0%	84.5%

Fig. 9 shows the recall evaluation and the comparison with the rule-based approach. The speed anomaly in the figure refers to fast acceleration, deceleration, and overspeed. In the evaluation, it is found that the rule-based approach may have ignored anomalies in several specific situations, such as the fast acceleration when driving at a high speed requiring a smaller threshold for the rules, and also the swerving anomaly, which might need a more comprehensive rule set. It might be possible to build a comprehensive rule set to perform much better. However, in some cases, such as driving monitoring, it might be hard to build such a reasonable rule set manually because the data might have many attributes and have complex relationships with each other. *SafeDrive* fills this gap by automatically extracting complex relationships from the data set and representing such relationships with an SG model.

Recent research uses the hidden Markov model (HMM) and the SVM to detect driving anomalies. Thus, we further compare *SafeDrive* with an HMM-based detector and an SVM-based detector. As shown in Table III, with our dataset, *SafeDrive* outperforms both SVM-based and HMM-based detectors in the detection accuracy. Given that the dataset of *SafeDrive* is too huge to label them all, to train the model, parts of the training data are labeled as a training data set for the HMM and the SVM. The testing result is even lower than 86%. The performance is improved as we increase the labeled training set. However, in IoV scenarios, it is hard to prepare sufficient labeled data to train the model because the collected data volume is too huge to label. Our *SafeDrive* approach outperforms those popular supervised algorithms because *SafeDrive* does not require labeled data, and thus, it can largely utilize the huge training set to improve the performance. It is believed that supervised learning algorithms can also produce a better result, but the high cost in large-scale industrial scenarios might not be desirable.

By fusing different data in an SG, *SafeDrive* is able to detect detailed anomalies from streamed driving data. However, the fusion of multiple data in one model may also have negative effects. As the results suggest, the combination may cover several anomalies because the data may interfere with each other and thus lower the sensitivity of the model. In practice, for different data attributes, it is suggested to do a correlation analysis for different data attributes and to regulate the SG by removing the connections between attributes with no obvious correlations.

Compared with other approaches, *SafeDrive* has several main advantages. First, it uses an SG to represent normal driving behavior. The graph is derived from a large dataset, and thus, this metric is more objective. Second, *SafeDrive* does not require labeled data to train the detector. Such a factor is important for IoV scenarios because labeled abnormal data are often hard to acquire. Third, the SG can be updated with new data arriving, making *SafeDrive* sensitive and adaptive to the change of the environment. Finally, *SafeDrive* has low computation cost. Its major cost is matching the newly arrived data as a temporal subgraph, and such matching can be very efficiently implemented by indexing. These advantages make *SafeDrive* applicable for large-scale IoV scenarios.

However, *SafeDrive* also has its limitations as driving is not only status aware but also environment aware. Since driving behaviors are affected by many environment factors such as traffic and road conditions, and they can be also affected by other drivers' behaviors. Therefore, several detected unsafe behaviors may actually correspond to safe behaviors depending on specific environments. For example, a hard brake to avoid a collision may be considered as safe behaviors when the vehicle driving in the front suddenly brakes or there are some obstacles on the way. Unfortunately, this problem can hardly be solved solely with vehicle data. Other data types such as video or radar data should be introduced to sense environment conditions. In fact, part of our ongoing work is to develop a mobile-phone-based application to collect and analyze driving video and g-sensor data, aiming at extracting information about road conditions and thus to enhance the detection of traffic conditions from driving video data. But, for the time being, this application has not been widely deployed. Nevertheless, to construct a comprehensive driving sensing platform, it is necessary to fuse different types of driving data. Such direction is becoming an important research trend. Another potential limitation of this work could be not taking personal characteristics into consideration. This model addresses the problem of the driver's behavior evolution in general but does not consider the change of personal behaviors as one might change his or her driving style over time. The personalized behavioral analysis should be carefully addressed because a driver may have a habitual bad driving behavior, which might be inappropriately considered as normal and thus be ignored in our solution.

VI. Conclusion

We have proposed an *online, unsupervised*, and *status-aware* approach, named *SafeDrive*, to detect abnormal driving behaviors from large-scale vehicle data. Compared with other approaches, *SafeDrive* uses normal behaviors, which are represented by an SG extracted from a large dataset, as benchmarks for identifying abnormal behaviors. The real-time-uploaded driving data stream is split into segments by *SafeDrive* and each segment is mapped as a TS-SG, which is further compared with the SG. A real-world IoV system with over 29 000 real vehicles connected is used to evaluate the performance of *SafeDrive*. The results suggest that our model performs well in detecting various driving anomalies without using labeled training data. The computational cost of *SafeDrive* is very low, and a single PC is capable of dealing with millions of data instances per second, enabling *SafeDrive* as an ideal option to detect driving anomalies from large-scale vehicle data. Still, future work on analyzing driving behavior patterns by fusing vehicle data and video, and even road network, needs to be conducted to provide comprehensive understanding of driving behaviors.

Acknowledgment

The authors would like to thank UCAR Inc. for the collaboration of the data that used in this paper.

References

[1] K. Jakobsen, S. C. Mouritsen, and K. Torp, "Evaluating eco-driving advice using GPS/CANBus data," in *Proc. 21st ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2013, pp. 44–53.

[2] F. Xiaoqiu, J. Jinzhang, and Z. Guoqiang, "Impact of driving behavior on the traffic safety of highway intersection," in *Proc. 3rd Int. Conf. Measuring Technol. Mechatronics Autom.*, 2011, vol. 2, pp. 370–373.

[3] X. Gao *et al.*, "Elastic pathing: Your speed is enough to track you," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2014, pp. 975–986.

[4] Z. Chen, J. Yu, Y. Zhu, Y. Chen, and M. Li, "D3: Abnormal driving behaviors detection and identification using smartphone sensors," in *Proc. 12th Annu. IEEE Int. Conf. Sens., Commun., Netw.*, 2015, pp. 524–532.

[5] T. Chakravarty, A. Ghose, C. Bhaumik, and A. Chowdhury, "MobiDriveScore—A system for mobile sensor based driving analysis: A risk assessment model for improving one's driving," in *Proc. 7th Int. Conf. Sens. Technol.*, 2013, pp. 338–344.

[6] J. H. Hong, B. Margines, and A. K. Dey, "A smartphone-based sensing platform to model aggressive driving behaviors," in *Proc. 32nd Annu. ACM Conf. Human Factors Comput. Syst.*, 2014, pp. 4047–4056.

[7] M. Amarasinghe *et al.*, "Cloud-based driver monitoring and vehicle diagnostic with OBD2 telematics," in *Proc. 15th Int. Conf. Adv. ICT Emerg. Regions*, 2015, pp. 243–249.

[8] H. Zhao, H. Zhou, C. Chen, and J. Chen, "Join driving: A smart phone-based driving behavior evaluation system," in *Proc. IEEE Global Commun. Conf.*, Dec. 2013, pp. 48–53.

[9] J. Dai, J. Teng, X. Bai, Z. Shen, and D. Xuan, "Mobile phone based drunk driving detection," in *Proc. 4th Int. Conf. Pervasive Comput. Technol. Healthcare*, Mar. 2010, pp. 1–8.

[10] M. Fazeen, B. Gozick, R. Dantu, M. Bhukhiya, and M. C. Gonzlez, "Safe driving using mobile phones," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1462–1468, Sep. 2012.

[11] A. E. M. Taha and N. Nasser, "Utilizing CAN-Bus and smartphones to enforce safe and responsible driving," in *Proc. IEEE Symp. Comput. Commun.*, Jul. 2015, pp. 111–115.

[12] G. C. M. Quintero, J. A. O. López, and J. M. P. Rúa, "Intelligent erratic driving diagnosis based on artificial neural networks," in *Proc. Andean Council Int. Conf.*, Sep. 2010, pp. 1–6.

[13] D. Jaramillo and C. Narvez, "Vehicle online monitoring system based on fuzzy classifier," in *Proc. 3rd Int. Conf. Adv. Veh. Syst., Technol. Appl.*, Jun. 2014, pp. 33–38.

[14] D. A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *Proc. 14th Int. Conf. Intell. Transp. Syst.*, Oct. 2011, pp. 1609–1615.

[15] K. Li, M. Lu, F. Lu, Q. Lv, L. Shang, and D. Maksimovic, "Personalized driving behavior monitoring and analysis for emerging hybrid vehicles," in *Proc. 10th Int. Conf. Pervasive Comput.*, Jun. 2012, pp. 1–19.

[16] A. Bolovinou, A. Amditis, F. Bellotti, and M. Tarkiainen, "Driving style recognition for co-operative driving: A survey," in *Proc. 6th Int. Conf. Adaptive Self-Adaptive Syst. Appl.*, 2014, pp. 73–78.

[17] Z. Constantinescu, C. Marinoiu, and M. Vladoiu, "Driving style analysis using data mining techniques," *Int. J. Comput. Commun. Control*, vol. 5, pp. 654–663, 2010.

[18] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surveys*, vol. 41, no. 3, 2009, Art. no. 15.

[19] C. Miyajima *et al.*, "Driver modeling based on driving behavior and its evaluation in driver identification," *Proc. IEEE*, vol. 95, no. 2, pp. 427–437, Feb. 2007.

[20] C. Karatas *et al.*, "Leveraging wearables for steering and driver tracking," in *Proc. IEEE Int. Conf. Comput. Commun.*, 2016, pp. 1–9.

[21] A. Burton *et al.*, "Driver identification and authentication with active behavior modeling," in *Proc. Int. Workshop Green ICT Smart Netw.*, Montreal, QC, Canada, 2016.

[22] T. Banerjee, A. Chowdhury, and T. Chakravarty, "MyDrive: Drive behavior analytics method and platform," in *Proc. 3rd Int. Workshop Phys. Analytics*, Jun. 2016, pp. 7–12.

[23] P. R. Lei, "A framework for anomaly detection in maritime trajectory behavior," *Knowl. Inf. Syst.*, vol. 47, no. 1, pp. 189–214., 2016.

[24] G. A. Susto, A. Schirru, S. Pampuri, and S. McLoone, "Supervised aggregative feature extraction for big data time series regression," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1243–1252, Jun. 2016.

[25] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection," in *Applications of Data Mining in Computer Security*. New York, NY, USA: Springer, 2002, pp. 77–101.

[26] A. Zimek, E. Schubert, and H. P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statist. Anal. Data Mining*, vol. 5, no. 5, pp. 363–387, 2012.

[27] American Automobile Association, *Aggressive Driving: Research Update*. Washington, DC, USA: Amer. Automobile Assoc. Found. Traffic Safety, 2009.

Mingming Zhang is working toward the Ph.D. degree in computer science at Beihang University, Beijing, China. He is also a visiting Ph.D. student at the University of Illinois at Urbana–Champaign, Champaign, IL, USA.

His research interests include pervaseive computing, Internet of Vehicles, and distributed systems.

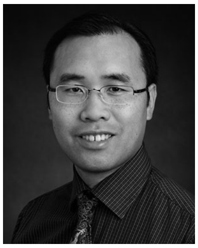
Chao Chen (M'15) received the B.Sc. and M.Sc. degrees in control science and control engineering from Northwestern Polytechnical University, Xi'an, China, in 2007 and 2010, respectively, and the Ph.D. degree in computer science and technology from Pierre and Marie Curie University and Institut Mines-TELECOM/TELECOM SudParis, Paris, France, in 2014.

He is an Associate Professor at the College of Computer Science, Chongqing University, Chongqing, China. His research interests include pervasive computing, social network analysis, urban logistics, data mining from large-scale taxi data, and big data analytics for smart cities.

Dr. Chen received the Best Paper Runner-Up Award at MobiQuitous 2011.

Tianyu Wo (M'06) received the B.S. and Ph.D. degrees in computer science from Beihang University, Beijing, China in 2001 and 2008, respectively.

He is an Associate Professor at Beihang University. His research interests include system software and applications in distributed systems.



Tao Xie (SM'12) received the M.S. degree from Peking University, Beijing, China, and the Ph.D. degree from the University of Washington at Seattle, WA, USA, in 2000 and 2005, respectively, both in computer science.

He is an Associate Professor and Willett Faculty Scholar with the Department of Computer Science, University of Illinois at Urbana–Champaign, Champaign, IL, USA. His research interests include software testing, program analysis, software analytics, software security, and educational software engineering.

Md Zakirul Alam Bhuiyan (M'13) received the B.Sc. degree from International Islamic University, Chittagong, Bangladesh, in 2005, and the M.Eng. and Ph.D. degrees from Central South University, Changsha, China, in 2009 and 2013, respectively, all in computer science and technology.

He is an Assistant Professor with the Department of Computer and Information Sciences, Fordham University. Earlier, he was an Assistant Professor at Temple University and a Postdoctoral Fellow at Central South University, a Research Assistant at the Hong Kong Polytechnic University, and a Software Engineer in industries. His research interests include dependable cyber physical systems, wireless sensor network applications, big data, cloud computing, and cyber security.

Dr. Bhuiyan served as a Lead Guest Editor of the IEEE TRANSACTIONS ON BIG DATA, *ACM Transactions on Cyber-Physical Systems*, and *Information Sciences*. He is a member of the ACM.

Xuelian Lin received the B.E. and Ph.D. degrees in computer science and engineering from Beihang University, Beijing, China, in 1999 and 2013, respectively.

He is a Research Scientist with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University. His research interests include time-series data management, data processing, and Internet of Vehicles.