TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 18/18 pp207–222
DOI: 10.26599/TST.2021.9010034
Volume 27, Number 1, February 2022

GGC: Gray-Granger Causality Method for Sensor Correlation Network Structure Mining on High-Speed Train

Jie Man, Honghui Dong*, Limin Jia, and Yong Qin

Abstract: Vehicle information on high-speed trains can not only determine whether the various parts of the train are working normally, but also predict the train's future operating status. How to obtain valuable information from massive vehicle data is a difficult point. First, we divide the vehicle data of a high-speed train into 13 subsystem datasets, according to the functions of the collection components. Then, according to the gray theory and the Granger causality test, we propose the Gray-Granger Causality (GGC) model, which can construct a vehicle information network on the basis of the correlation between the collection components. By using the complex network theory to mine vehicle information and its subsystem networks, we find that the vehicle information network and its subsystem networks have the characteristics of a scale-free network. In addition, the vehicle information network is weak against attacks, but the subsystem network is closely connected and strong against attacks.

Key words: vehicle information network; structure mining; gray theory; Granger causality theory; complex network theory; high-speed train

1 Introduction

In recent years, China's railway development has become faster and safer. In addition, China has formed the fastest and largest high-speed railway network in the world[1]. As of December 2018, China's high-speed railway operation mileage reached 29 000 km, accounting for two-thirds of the world's high-speed railway operating mileage. China Railway High-speed (CRH) series Electric Multiple Units (EMU) trains are the primary carrier vehicles independently developed by China. The average driving speed of these trains can exceed 350 km/h. However, under the trend of high-speed and large-scale development, high-speed train accidents occasionally occur. For example, on July 23, 2011, lightning strikes caused detection sensor failure, the train numbered D301 collided with the train numbered D3115, resulting in 40 deaths and 172 injuries. On August 12, 2018, the air conditioner of the train numbered G40 failed and the passengers fainted due to the high temperature.

To improve the safety and comfort of railway transportation, China Railway installs thousands of sensors on the train to detect the operation of each part and establishes a vehicle Wireless Transmission Data System (WTDS) on EMU. WTDS is composed of a vehicle-mounted host, a vehicle-mounted antenna, a multiband combiner, and an antenna extension cable. It collects and processes train operation status and alarm information, stores data, and transmits wirelessly. As shown in Fig. 1, the signal acquisition unit collects train operation status data and then sends these data to the Data Control Center (DCC) through the WTDS. After the data are obtained, the DCC uses the data to predict train operation status.

An eight-section high-speed train is equipped with

• Jie Man, Honghui Dong, Limin Jia, and Yong Qin are with the State Key Laboratory of Rail Traffic Control and Safety, and Beijing Research Center of Urban Traffic Information Sensing and Service Technologies, Beijing Jiaotong University, Beijing 100044, China. E-mail: 18114024@bjtu.edu.cn; hhdong@bjtu.edu.cn; jialm@vip.sina.com; yqin@bjtu.edu.cn.
∗ To whom correspondence should be addressed.

Manuscript received: 2021-04-01; revised: 2021-04-26; accepted: 2021-04-28

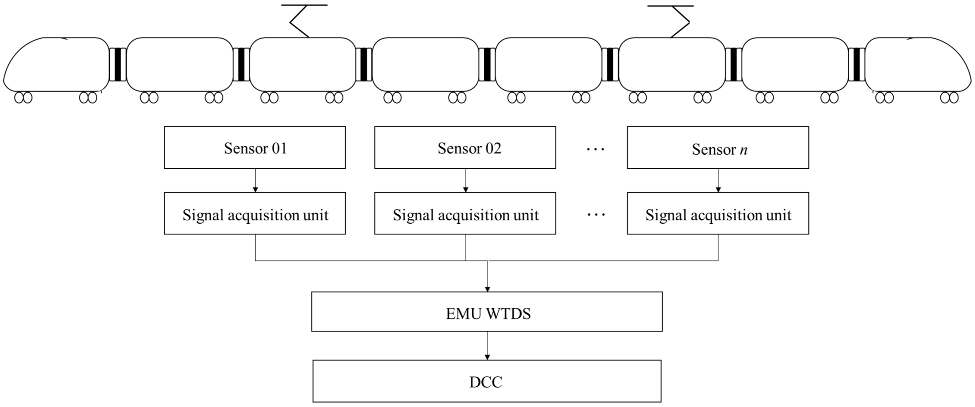


Fig. 1 WTDS architecture of bearing failure prediction in high-speed trains.

1673 sensors with a sampling frequency of 1 s, meaning six million pieces of data are processed in the DCC per hour. Furthermore, substantial vehicle information data are invalid for the running status of the train. In addition, data loss is serious during train operations. Given problems, such as weak transmission network signals and individual sensor failures, further research has become difficult. To solve the problem of a large amount of data and poor data quality, we choose complex network theory to mine vehicle data in advance.

A complex network is a serviceable model for data repair and mining[2]; it can abstract individuals into network nodes and abstract the associations between individuals into network edges. Complex network models include regular networks[3, 4], small-world networks[5, 6], and random networks[7]. Transportation networks[8–10], scientific cooperation networks[11, 12], software communication networks[13, 14], protein networks[15–17], social networks[18, 19], and gene regulatory networks[20–22] can all be abstracted into complex networks. In general, if the network structure of a system is given, then the characteristics of the system can be easily analyzed. However, numerous networks cannot fully understand their topology, even unknown networks, such as brain neural networks[23]. These networks have important implications for human understanding of the nature of a system in reality. How to obtain the network structure in accordance with each individual's behavior data is one critical issue worth studying. Garlaschelli and Loffredo[24] focused on the reciprocity of networks and proposed a new reciprocity measurement method, which uses interconnection to calculate the correlation and reorder the network structure. Tsonis and Swanson[25] constructed a meteorological network based on linear correlation coefficients. Donges et al.[26, 27] used nonlinear mutual information to establish a climate network and compared the linear correlation coefficient with the nonlinear mutual information method; both methods can obtain roughly the same result in many cases. Elzen et al.[28] considered structural and timing factors on the network and provided an extended massive sequence view, which can intuitively analyze the hierarchical structure, node attributes, community structure, network trends, and others. Sun et al.[29] proposed a new method of constructing a three-dimensional network structure and verified that the method is scientific in using the internal dataset of the system. Zhang and Ma[30] used space L and space P to build a high-speed rail train service network, which uses high-speed railway stations as nodes and the connection between stations as edges by applying train service data. The service network can quantitatively evaluate railway stations. Wang[31] calculated the income of each future in the target time interval based on the future income data, and then calculated the correlation coefficient of income between different futures. They constructed a network by using futures as nodes and the income correlation coefficient between futures as edges. The future income network reflects the correlation between future earnings.

Complex networks play a crucial role in data mining. However, research on the mining of vehicle information is scarce. First, we divide the onboard data of high-speed trains into subsystems in accordance with the functions of the collection components. Then, we propose the Gray-Granger Causality (GGC) model based on gray theory and Granger causality theory; this model can construct a vehicle information network based on the correlation between the collection components. We use vehicle information data from April 10 to April 16, 2019, to construct the network and its corresponding 13 subsystem networks. Lastly, we use the complex network theory to mine the vehicle information network and its 13 subsystem networks and find that the vehicle information network and its subsystem networks have the characteristics of a scale-free network. In addition, the vehicle information network is weak against attacks, but the subsystem network is closely connected and is strong against attacks.

The innovations of this study are listed in the following three aspects:

(1) Research object innovation. China Railway established WTDS for EMU in January 2019; thus, not many scholars have conducted research on vehicle information. Therefore, our research is innovative in the industry. In addition, we use vehicle information to ensure that the safety of train operation is of practical importance.

(2) Research perspective innovation. Regarding research on train operation safety, most scholars focus on data prediction, and a few people excavate the relationships between data. We use complex network theory to mine the vehicle data network and find that the network has the characteristics of a scale-free network. At the same time, we can infer the operating status of the entire train on the basis of the data of a few important nodes on the network, thus greatly reducing the amount of calculation.

(3) Research method innovation. We propose the GGC model, which can construct a vehicle information network based on the correlation between data. This method is universal for similar data research.

The rest of the paper is organized as follows: The proposed GGC model for a vehicle information network is presented in Section 2. Section 3 provides the statistic and preprocessing of vehicle information data. Section 4 reports the results, and Section 5 presents the discussions. The conclusions are provided in Section 6.

2 Method

2.1 Vehicle information network structure

We stipulate that vehicle data detection points are represented as nodes of the vehicle information network, the relationship between the detection points is the edge of the network, and the correlation coefficient between the detection points is the weight of the network edge. In addition, the causal relationship between detection points is the direction of the edge of the network. The vehicle information network is defined as follows:

The vehicle information network is composed of a finite set of nonempty vertices and a set of edges between the vertices, usually expressed as

$$\begin{cases} D = (V, E, W); \\ V = \{v_1, v_2, v_3, \dots, v_n\}; \\ E = \{e_{11}, e_{12}, \dots, e_{ij}, \dots, e_{nn}\}, \quad i, j = 1, 2, \dots, n; \\ W = \{w_{11}, w_{12}, \dots, w_{ij}, \dots, w_{nn}\}, \ i, j = 1, 2, \dots, n \end{cases} \tag{1}$$

where D represents a network, V is a set of vertices in network D, E is a set of edges in network D, and e_{ij} represents the direction from node v_i to v_j . Furthermore, W is a set of weight in network D, and w_{ij} represents the weight from node v_i to v_j.

2.2 GGC model

(1) Determining the network framework based on gray theory

The basic idea of the gray theory is to compare the similarity of the system data sequence to show the size, strength, and order of the correlation between various analysis factors by calculating the gray correlation degree. In general, if the changing trends of the two factors are consistent, then the correlation between the two factors is relatively high. We use the gray correlation analysis method to calculate the correlation degree of the nodes and then give a threshold. If the correlation degree of two nodes exceeds this threshold, then the two nodes are considered connected. Lastly, we add two physical factors as auxiliary conditions for whether two nodes are connected, that is, whether they belong to the same subsystem and whether they belong to the same car.

The steps to determine the network framework are as follows:

Step 1: Construct a vehicle information sequence matrix X with the number of detection points n and the time length t,

$$X = (X_1, \dots, X_n) = \begin{pmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ x_1(2) & x_2(2) & \cdots & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t) & x_2(t) & \cdots & x_n(t) \end{pmatrix} \tag{2}$$

Step 2: Calculate the correlation coefficient $\zeta_{ij}(k)$ of the element between each comparison measuring point x_j and reference measuring point x_i at time k,

$$\zeta_{ij}(k) = \frac{\min_j \min_k |x_i(k) - x_j(k)| + \rho \cdot \max_j \max_k |x_i(k) - x_j(k)|}{|x_i(k) - x_j(k)| + \rho \cdot \max_j \max_k |x_i(k) - x_j(k)|}, \quad i \neq j \tag{3}$$

where ρ is the resolution coefficient. The smaller ρ, the stronger the resolution ability. In this study, $\rho = 0.5$.

Step 3: Calculate a weighted average of the correlation coefficients of all the corresponding elements between each comparison measuring point and the reference measuring point, which is also called r'_{ij},

$$r'_{ij} = \frac{1}{t}\sum_{k=1}^{t} W_k \cdot \zeta_{ij}(k),\ \ i \neq j \tag{4}$$

Step 4: Construct a sensor data relationship matrix $R_1 = (r^1_{ij})_{n\times n}, i, j = 1, 2, \ldots, n$, where R_1 meets the following conditions:

$$r^1_{ij} = \begin{cases} r'_{ij}, & i \neq j; \\ 0, & i = j \end{cases} \tag{5}$$

Step 5: Define the constraints of the vehicle information network. The relationship between nodes is related to the subsystem and carriage. Therefore, the factors affecting network connection must be comprehensively considered when describing the network edge. We choose three constraints, namely, data relationship R_1, system ownership R_2, and carriage ownership R_3, and obtain the correlation matrix R of the vehicle information network by the weighted average method. $R_2 = (r^2_{ij})_{n\times n}$, $R_3 = (r^3_{ij})_{n\times n}$, and $R = (r_{ij})_{n\times n}$ meet the following conditions:

$$r^2_{ij} = \begin{cases} 1, & \text{If sensor } i \text{ and sensor } j \text{ belong to the} \\ & \quad \text{same system;} \\ 0, & \text{Others} \end{cases} \tag{6}$$

$$r^3_{ij} = \begin{cases} 1, & \text{If sensor } i \text{ and sensor } j \text{ belong to} \\ & \quad \text{the same carriage;} \\ 0, & \text{Others} \end{cases} \tag{7}$$

$$R = (r_{ij})_{n\times n} = w_1 \cdot R_1 + w_2 \cdot R_2 + w_3 \cdot R_3 \tag{8}$$

where $w_1 + w_2 + w_3 = 1$.

Step 6: Construct a weight matrix $W = (w_{ij})_{n\times n}$, which specifies that the weight of the network is represented by the correlation coefficient between sensors,

$$W = (w_{ij})_{n\times n} = (r_{ij})_{n\times n} \tag{9}$$

Step 7: Determine the relationship threshold T, and change the sensor correlation network matrix $R = (r^*_{ij})_{n\times n}$ to meet the following conditions:

$$r^*_{ij} = \begin{cases} 1, & r_{ij} \geqslant T; \\ 0, & \text{Others} \end{cases} \tag{10}$$

(2) Determining the network direction based on the Granger causality test

After discussing the correlation between nodes, we can obtain a directionless weighted network. The direction of the edge is determined in accordance with the causal relationship between the two measuring points connected by each edge. We stipulate that from the time series, the value of the vehicle information measuring point X_i affects the value of the measuring point X_j within l seconds, that is, the information changes observed on X_j can explain the information changes appearing on X_j after l seconds, and a causality relationship exists between X_i and X_j, and $X_i \rightarrow X_j$. Given that the vehicle information time series is a stable series, we use the Granger causality test to establish the model of measuring points X_i and X_j to determine the direction. The model is shown in the following:

$$X_j(t) = \sum_{k=1}^{l} a_k X_i(t-k) + \sum_{k=1}^{l} b_k X_j(t-k) + \varepsilon_t \tag{11}$$

$$X_i(t) = \sum_{k=1}^{l} c_k X_j(t-k) + \sum_{k=1}^{l} d_k X_i(t-k) + \mu_t \tag{12}$$

where $X_i(t)$ and $X_j(t)$ are the current vehicle values; $X_i(t-k)$ and $X_j(t-k)$ are the vehicle values of the previous period; a_k and c_k are Granger causality coefficients; b_k and d_k are autoregression coefficients; ε_t and μ_t are prediction errors, which are independent of time point and are white noise by default.

The Granger causality test is completed by the constrained F-statistic. If $X_i(t)$ is not the Granger cause of $X_j(t)$, then $a_1 = a_2 = \cdots = a_l = 0$ in Eq. (11). The residual sum of squares of $X_i(t)$ is RSS_U, and the residual sum of squares of $X_j(t)$ is RSS_R; the F-statistic is

$$F = \frac{(RSS_R - RSS_U)/m}{RSS_U/(n-k)} \tag{13}$$

where m is the maximum lag order of $X_i(t)$.

For Eq. (11), if the conclusion of the F test rejects hypothesis $H_0 : a_1 = a_2 = \cdots = a_l = 0$, then $X_i(t)$ is the Granger cause of $X_j(t)$; otherwise, it is not. For Eq. (12), if the conclusion of the F test rejects hypothesis $H_0 : c_1 = c_2 = \cdots = c_l = 0$, then $X_j(t)$ is the Granger cause of $X_i(t)$. Moreover, the lower the probability of the F-statistic, the stronger the Granger causality. A vehicle information network adjacency matrix $E = (e_{ij})_{n\times n}$ is constructed, where E meets the following conditions:

$$e_{ij} = \begin{cases} 1, & \text{If } r^*_{ij} = 1 \text{ and } v_x \rightarrow v_y; \\ 0, & \text{Others} \end{cases} \tag{14}$$

The process of GGC model is shown in Fig. 2.

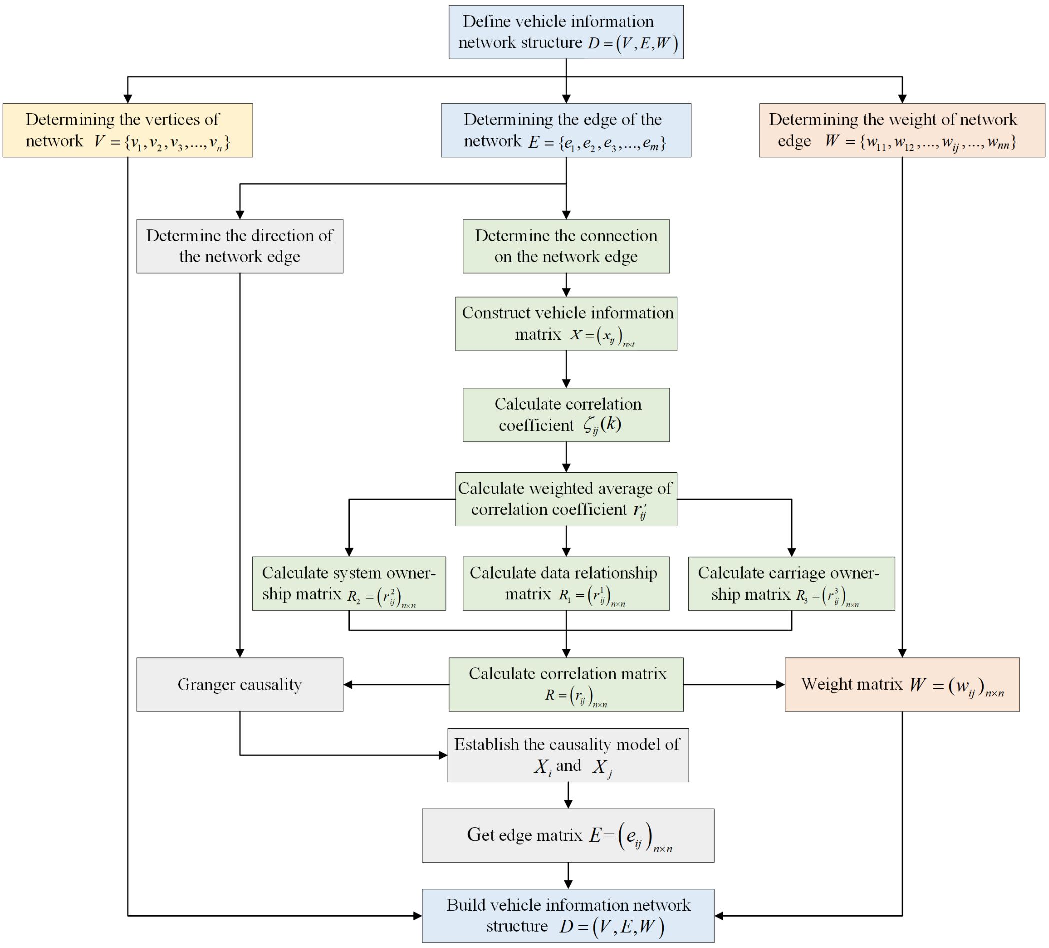

Define vehicle information network structure $D=(V,E,W)$
Determining the vertices of network $V=\{v_1,v_2,v_3,\dots,v_n\}$
Determining the edge of the network $E=\{e_1,e_2,e_3,\dots,e_m\}$
Determining the weight of network edge $W=\{w_{11},w_{12},\dots,w_{ij},\dots,w_{nn}\}$
Determine the direction of the network edge
Determine the connection on the network edge
Construct vehicle information matrix $X=(x_{ij})_{n\times t}$
Calculate correlation coefficient $\zeta_{ij}(k)$
Calculate weighted average of correlation coefficient r_{ij}'
Calculate system ownership matrix $R_2=(r_{ij}^2)_{n\times n}$
Calculate data relationship matrix $R_1=(r_{ij}^1)_{n\times n}$
Calculate carriage ownership matrix $R_3=(r_{ij}^3)_{n\times n}$
Granger causality
Calculate correlation matrix $R=(r_{ij})_{n\times n}$
Weight matrix $W=(w_{ij})_{n\times n}$
Establish the causality model of X_i and X_j
Get edge matrix $E=(e_{ij})_{n\times n}$
Build vehicle information network structure $D=(V,E,W)$

Fig. 2 Process of GGC model.

3 Data Statistic

3.1 Subsystem division

In accordance with the function of the detected component, WTDS can be divided into traction system, braking system, axle temperature system, enabling system, network system, passenger service system, charging system, fire alarm system, auxiliary system, ventilation system, door system, facility system, and linked system. The detection area is divided, as shown in Fig. 3.

In Fig. 3, the functions of the 13 subsystems are as follows:

(1) The traction system provides traction power for the train and detects traction converter and traction motor.

(2) The braking system mainly obtains the brake cylinder, air brake, and emergency brake.

(3) The axle temperature system obtains the temperature of the bogie, gearbox, and traction motor.

(4) The enabling system enables the train to obtain electrical energy and mainly detects the rise and fall of the pantograph.

(5) The network system ensures that the train communication network transmits normally and mainly

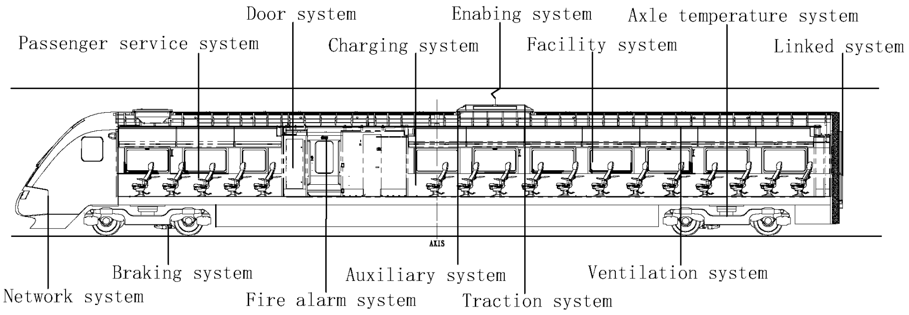


Fig. 3 Subsystem distribution.

detects a line transmission link.

(6) The passenger service system tests whether the electronic display and water heater are abnormal.

(7) The charging system records whether the charger and battery of the train are abnormal.

(8) The fire alarm system consists of a series of smoke detectors, which primarily sense tiny smoke particles.

(9) The auxiliary system obtains the variation values of an auxiliary inverter, auxiliary converter, and auxiliary transformer.

(10) The ventilation system determines whether the air conditioner is turned on and obtains the temperature inside the carriage.

(11) The door system guarantees that the door can be opened and closed automatically.

(12) The facility system detects lighting facilities, sewage disposal facilities, and other common facilities on the train.

(13) The linked system tests the status of the electric and mechanical hooks, as well as the power supply condition during the operation of the electric hook.

3.2 Vehicle data statistic

Vehicle data have two types, namely, continuous and discrete data. We count the number of vehicle data from four aspects: total number, number of different subsystems, number of different carriages, and number of different types, as shown in Fig. 4.

From Fig. 4, we analyze from different subsystems that the number of vehicle data in the ventilation system is the highest, followed by the traction system, the facility system, the braking system, and the axle temperature system. In addition, the types of equipment detected by these five systems are essential factors that affect train operation. We count from different carriages that the number of vehicle data in each carriage is nearly equal, indicating that vehicle data are evenly distributed in carriages. Lastly, we discover from different types that the number of discrete data is far more than the amount of continuous data, thus explaining that continuous data need to be discrete during data processing.

3.3 Vehicle data processing

We intend to use the vehicle data from April 10 to April 16, 2019. The sampling frequency is 1 second. With additional data, such as weather, GPS, and faults, a train of eight carriages can receive approximately 2000 data each second, a total of 172 million data per day. The processing power and I/O performance of a single machine cannot support the operation of such vast amounts of data. To improve the efficiency of data processing, we decide to use the Spark big data platform to process vehicle data. During data transmission, vehicle data are encrypted in accordance with a set of protocol specifications for data security. Therefore, the sensor interpretation table must be decrypted before data mining is performed. Table 1 shows an example of encrypted information and an example of interpreted reference information.

Figure 5 shows the flow of vehicle data preprocessing.

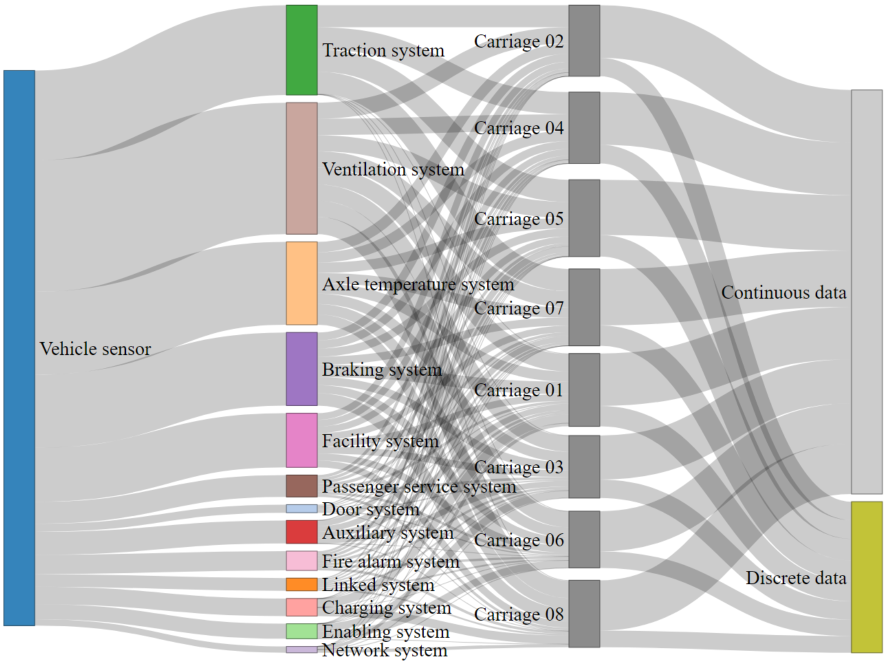


Fig. 4 Vehicle data statistic.

Table 1 Vehicle information example.

(a) Encrypted information

Train_no	Sensor_time	Sensor_code	Sensor_value
A001	2019/4/10 00:00:00	ktlocalautovol	0
A001	2019/4/10 00:00:01	ktfault8814tx	0
A001	2019/4/10 00:00:01	ktfault881C	0
A001	2019/4/10 00:00:01	ktfaultearth22t	0.35
A001	2019/4/10 00:00:01	ktfault8818t	0
A001	2019/4/10 00:00:02	zdwspspd	1
⋮	⋮	⋮	⋮

(b) Interpreted reference information

Sensor_code	Vehicle_system	Sensor_description	Sensor_type
sncompanto	Enabling system	Whether the pantograph gets electricity	Discrete
fzinputvol	Auxiliary system	Auxiliary converter input voltage	Continuous
qyinclude	Traction system	Whether the traction converter is working	Discrete
zdwspspd	Braking system	WSP speed value	Continuous
cdubat	Charging system	Battery voltage	Continuous
tfcool	Ventilation system	Whether the air conditioning cooling mode is on	Discrete
⋮	⋮	⋮	⋮

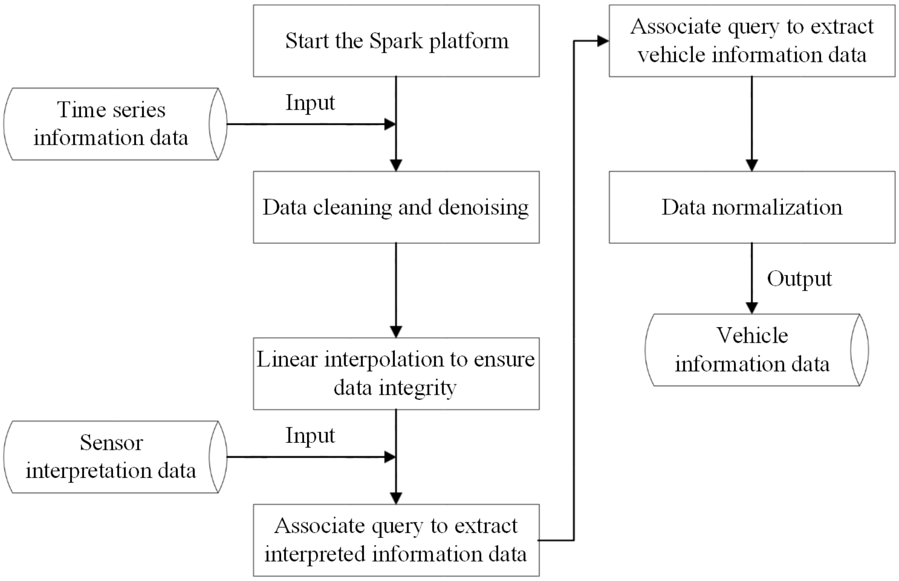

Start the Spark platform
Time series information data
Input
Data cleaning and denoising
Linear interpolation to ensure data integrity
Sensor interpretation data
Input
Associate query to extract interpreted information data
Associate query to extract vehicle information data
Data normalization
Output
Vehicle information data

Fig. 5 Vehicle data preprocessing process.

First, we preprocess the encrypted vehicle data to remove noise and duplicate values. Then, we use Sensor_code in Table 1a to match Sensor_code in Table 1b, and establish the processed vehicle information table, as shown in Table 2.

The first column in Table 2 is a time series, and the remaining columns are divided in accordance with the subsystem, and each subsystem can be divided into multiple columns of vehicle information. Lastly, we normalize the extracted vehicle information in Table 2.

4 Experiment and Result

4.1 Parametric setting

Based on the analysis in Section 2.2, we discover that the GGC model has five pivotal parameters, namely, the weight of data relationship w_1, the weight of system ownership w_2, the weight of carriage ownership w_3, the relationship threshold T, and the length of lag time l. They are divided into two sections to explain the selection of parameters.

Table 2 Processed vehicle information example.

Time	Traction system			Axle temperature system			···
	Continuous value 01	Discrete value 01	···	Continuous value 01	Discrete value 01	···	···
2019/4/10 7:00:00	0	0	···	0	0	···	···
2019/4/10 7:00:01	0	0	···	0	0	···	···
2019/4/10 7:00:02	0	0	···	0	0	···	···
2019/4/10 7:00:03	0	0	···	0	0	···	···
2019/4/10 7:00:04	0.070	0	···	0.1	0	···	···
2019/4/10 7:00:05	0.140	0	···	0	0	···	···
2019/4/10 7:00:06	0.149	0	···	0.053	0	···	···
2019/4/10 7:00:07	0.134	0	···	0	0	···	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

4.1.1 Edge connection parameters

For the connection of edges, we first introduce a concept, namely, the degree of a network. The degree of a node is the number of other nodes it connects and is a fundamental concept in single-node attributes. In the vehicle information network, the degree of a node reflects the number of nodes that have an association with it. If a node X_i in the network has k edges, then the degree of X_i is k.

When we use the GGC model to build a vehicle information network, we need to refer to two principles:

(1) Connect as many points as possible. Given that we are mining the association relationship between various detection points on the train, we need to try our best to ensure that each detection point has a detection point connected to it and reduce isolated points in the network. In other words, the number of points that have a degree greater than 0 needs to be as many as possible. We use C to describe this attribute,

$$C = \sum_{i=1}^{n} c_i,$$
$$c_i = \begin{cases} 1, & k_i > 0; \\ 0, & k_i = 0 \end{cases} \tag{15}$$

where the larger the value of C, the better the vehicle information network.

(2) Degree of a single point is as small as possible. We must control the degree of a single point when constructing the network. If many points in the network have a degree $n-1$, which means that this point is connected to all the remaining points, then research on the network is meaningless. We use the statistic S for description,

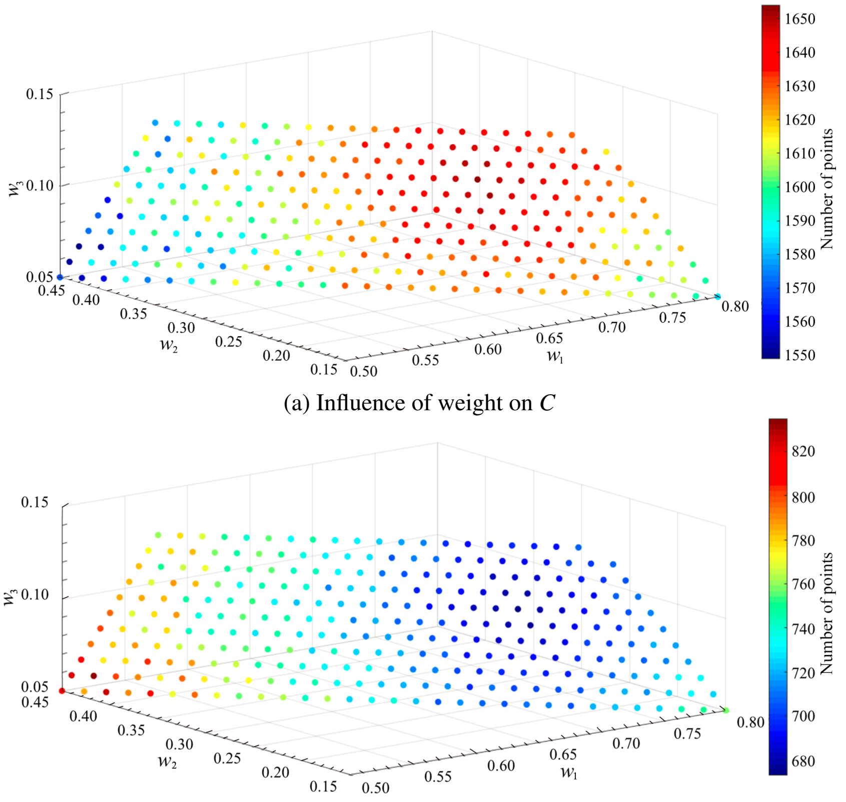
$$S = \sum_{i=1}^{n} s_i,$$
$$s_i = \begin{cases} 1, & k_i < 0.1n; \\ 0, & k_i \geqslant 0.1n \end{cases} \tag{16}$$

For $W = \{w_1, w_2, w_3\}$, we must provide a range of values on the basis of the actual content. The data relationship obtained by calculation is the main reference basis; thus, we set $w_1 \in [0.50, 0.80]$. The same subsystem has a certain effect on the vehicle data network. For example, the bearing temperature in the axle temperature system rises as the temperature of the generator rises; thus, we set $w_2 \in [0.15, 0.45]$. Furthermore, the same carriage has minimal effect on the vehicle data network; thus, we set $w_3 \in [0.05, 0.25]$. Therefore, the constraint condition of W can be obtained,

$$\begin{cases} w_1 + w_2 + w_3 = 1; \\ w_1 \in [0.50, 0.80]; \\ w_2 \in [0.15, 0.45]; \\ w_3 \in [0.05, 0.25]; \\ w_1 > w_2 > w_3 \end{cases} \tag{17}$$

We set the minimum change in weight to 0.01; thus, w_1, w_2, and w_3 have 375 combinations. We provide the threshold $T = 0.85$ and then use the values of C and S as evaluations to determine the values of w_1, w_2, and w_3, as shown in Figs. 6a and 6b. From Fig. 6, we can obtain the best weight $w_1 = 0.67$, $w_2 = 0.21$, and $w_3 = 0.12$.

After setting the weight W, we must determine the threshold T. Similar to the method of determining weights, we use the values of C and S as evaluation indexes to determine the threshold T, as shown in Fig. 7. From Fig. 7, we can obtain the best weight $T = 0.88$.



(a) Influence of weight on C

(b) Influence of weight on S

Fig. 6 Influence of weight on C and S.

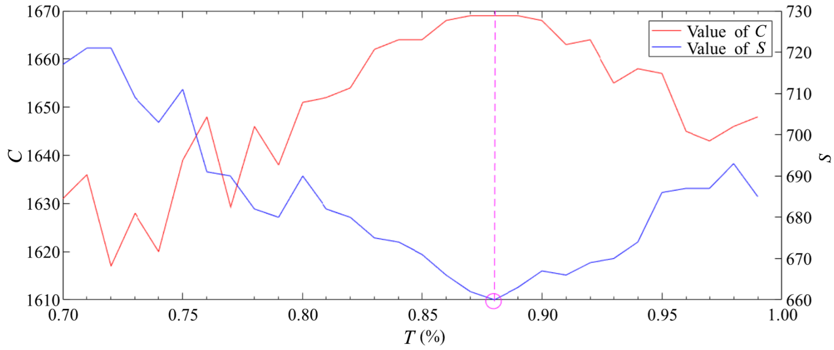


Fig. 7 Influence of the threshold on C and S.

4.1.2 Edge direction parameters

In accordance with the above analysis, the undirected vehicle information network can be obtained. From Section 2.2, we can know whether a causal relationship exists between node X_i and node X_j on the basis of the F statistical value. Meanwhile, in Eq. (13), the parameter to be determined is the length of lag time l. We choose the motor temperature X_1 and the bearing temperature X_2 as the research objects to conduct a two-way causal analysis to determine the value of l. The data of X_1 and X_2 are shown in Fig. 8.

We use EViews to operate X_1 and X_2 , where l is 30, 60, 90, 120, and 150 s. The test results are shown in Table 3 according to the null hypothesis.

From Table 3, we can know that when $l \geqslant 120$ s, $X_1$and X_2 are mutually causal. To save calculation time, we set $l = 120$ s. In addition, we use ten sets of measuring points with known causal relations for verification and prove that $l = 120$ s meets the requirements.

4.2 Vehicle information network

We use the GGC model to build the vehicle information network and use Gephi to draw the network structure, as shown in Fig. 9. In Fig. 9, the connection structure

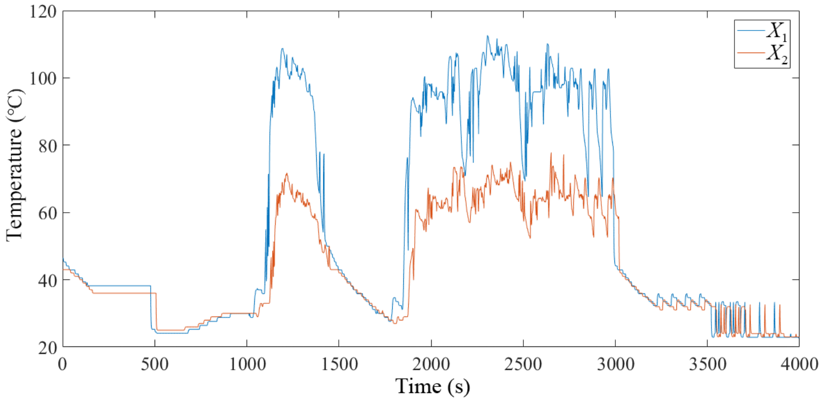


Fig. 8 Temperature values of X_1 and X_2.

Table 3 Granger causality test result.

Lag time (s)	Null hypothesis	F-statistic	Probability	Test result
30	X_1 is not Granger causality of X_2	0.9272	0.5363	Accept
	X_2 is not Granger causality of X_1	0.2260	0.8031	Accept
60	X_1 is not Granger causality of X_2	2.2509	0.2744	Accept
	X_2 is not Granger causality of X_1	0.8836	0.4582	Accept
90	X_1 is not Granger causality of X_2	4.0071	0.0796	Reject
	X_2 is not Granger causality of X_1	2.2686	0.1927	Accept
120	X_1 is not Granger causality of X_2	36.9368	0.0002	Reject
	X_2 is not Granger causality of X_1	10.1845	0.0015	Reject
150	X_1 is not Granger causality of X_2	133.2419	0	Reject
	X_2 is not Granger causality of X_1	13.0561	0.0004	Reject

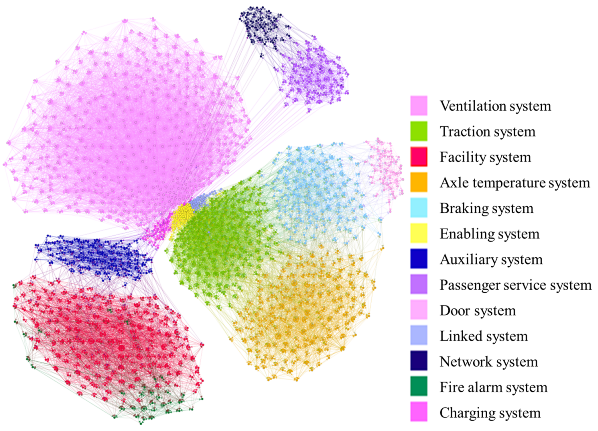


Fig. 9 Connection structure of vehicle information network.

of the 13 subsystems is presented. For example, the auxiliary system is the subsystem with the most vehicle detection points. It is connected to the network system, passenger service system, and enabling system. The auxiliary system is a large subsystem that escorts the safe operation of other subsystems on the train. Almost every subsystem needs an auxiliary system to complete its work, but the enabling system has the closest connection with the auxiliary system.

To verify the rationality of the vehicle information network established by the data, we count the number of system failures occurring every 120 s; the statistical results are shown in Table 4, which shows only the statistical results of the relationship between system pairs. If three or multisystem pairs fail at the same time, then the vehicle information network is divided into three groups or more related systems for statistical purposes. In Table 4, the fourth column indicates the ratio of the number of simultaneous failures of Subsystem 1 and Subsystem 2 to the number of all train failures in a cycle.

From Fig. 9 and Table 4, we can conclude that the failure of each subsystem in Table 4 is causal. For example, if the traction system fails and the train loses traction, then the train brakes urgently, and the axle temperature of the train rises rapidly. Therefore, the braking system and the axle temperature system return the fault information to the data center. Similarly, if the enabling system fails and the train cannot obtain power, the charging system also sends the fault information. This relationship is reflected in Fig. 9. For example, the traction system is closely connected with the axle temperature system and the braking system. We reverse the actual fault situation to prove that the connection structure in Fig. 9 is correct, thus verifying that the

Table 4 Subsystem fault data statistics.

Subsystem 1	Subsystem 2	Number of system failures	Ratio
Traction system	Braking system	83	0.5533
Enabling system	Charging system	82	0.5467
Traction system	Enabling system	81	0.5400
Door system	Braking system	78	0.5200
Axle temperature system	Braking system	77	0.5133
Axle temperature system	Traction system	76	0.5067
Linked system	Traction system	76	0.5067
Charging system	Facilitysystem	75	0.5000
Network system	Passenger service system	75	0.5000
Auxiliary system	Charging system	75	0.5000
Charging system	Network system	74	0.4933
Auxiliary system	Facility system	73	0.4867
Facility system	Fire alarm system	73	0.4867
Charging system	Passenger service system	72	0.4800
Charging system	Ventilation system	72	0.4800
Door system	Traction system	70	0.4667
Traction system	Facility system	67	0.4467
Enabling system	Auxiliary system	65	0.4333
Linked system	Braking system	63	0.4200
Enabling system	Ventilation system	58	0.3867
Network system	Facility system	60	0.4000

vehicle information network is in line with the actual engineering.

5 Discussion

Given that the vehicle information network is a large network with 1763 points, we decompose the network into 13 subsystem information networks to explore the relationship between the nodes. There are two decomposition principles. First, the edges within the same subsystem are retained. Second, the edges between nodes of different subsystems are deleted. The subsystem information network after processing is shown in Fig. 10.

Next, we use the relevant knowledge in the complex network to mine the vehicle information network from the aspects of the degree distribution, average node strength, average aggregation coefficient, and community prediction.

5.1 Distribution of degrees

The definition of degree has been introduced in Section 4.1. We analyze the degree distribution of subsystem networks, as shown in Fig. 11. After calculation, a statistical graph of the degrees of the vehicle network nodes of each subsystem and vehicle system is obtained, as shown in Fig. 11a. From Fig. 11a, in the vehicle information network, 87.89% of the nodes have a degree below 40. The degree of most of the nodes is relatively smaller than the total number of nodes, indicating that only a few nodes in the network play an important role in network failure.

If there are a total of n nodes in the network, among which n_k nodes have degree k, the calculation formula for degree distribution $p(k)$ can be obtained,

$$p(k) = \frac{n_k}{n} \tag{18}$$

The scale-free nature of each network is analyzed, and the degree distribution of each network is drawn in logarithmic coordinates, as shown in Fig. 11b, where the degree distribution of all systems obeys the power-law distribution; thus, the vehicle information network has scale-free characteristics. The scale-free characteristics reflects that the network has strong fault tolerance, but its anti-attack ability is poor. If one or more nodes with a large degree fail, then the vehicle information network may be paralyzed directly. Therefore, we must monitor the vehicle with numerous degrees in each system.

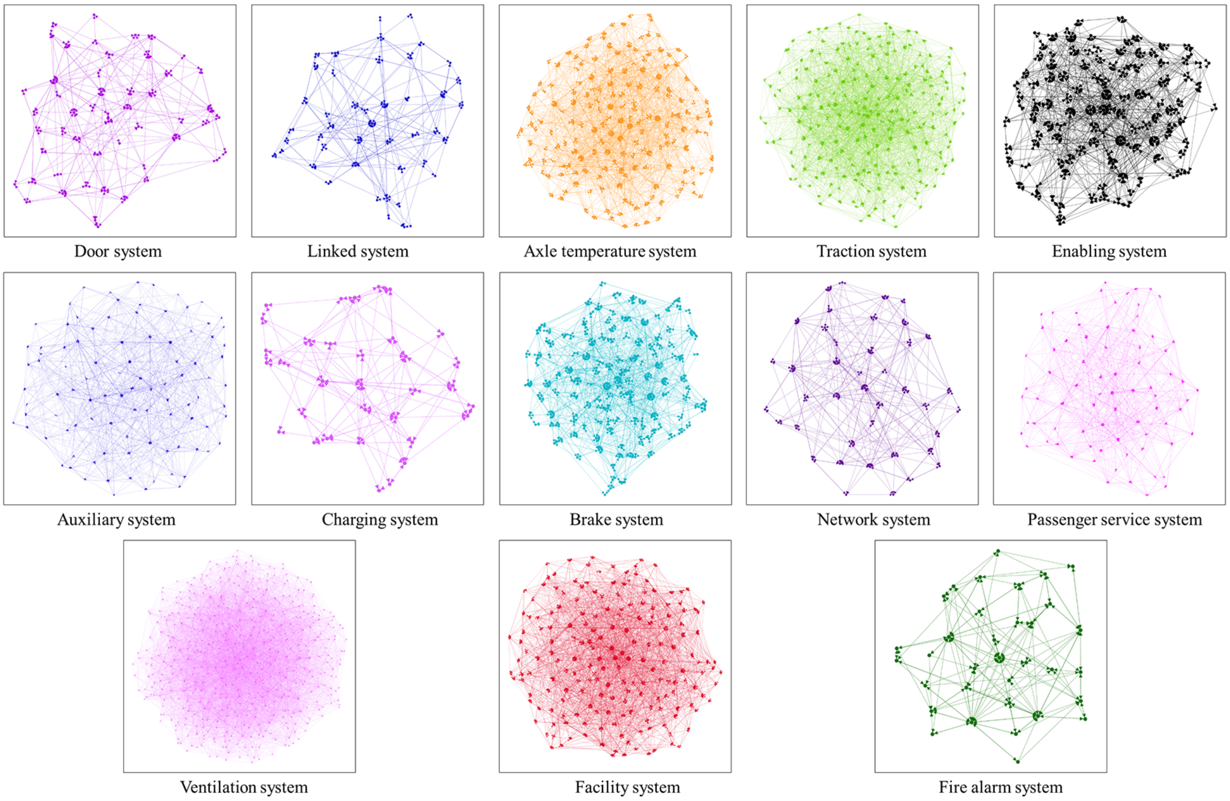
5.2 Average node strength

The node strength of the vehicle information network s_i is the sum of the weights on all the edges of node i. The average node strength $\overline{S}$ is the average of the strengths of all nodes in the network,

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n} s_i = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij}e_{ij}\right) \tag{19}$$

Average node strength can reflect network connectivity; the higher the strength, the better the network connectivity. We calculate the average node strength of 14 networks, and the results are shown in Fig. 12.

As shown in Fig. 12, the average node strength of the overall vehicle information network is 13.216, indicating that a node can be evenly associated with 13 nodes. As for the subsystem network, the average node strength of the door system, linked system, charging system, and fire alarm system is all less than 5, indicating that each node is relatively independent. The average node strength of the ventilation system, traction system, and facility system is more than 13, indicating that the interconnectivity performance of the vehicle information network in these systems is excellent, and the correlations are relatively significant. If one measuring point fails, then other measuring points in


Door system

Linked system

Axle temperature system

Traction system

Enabling system

Auxiliary system

Charging system

Brake system

Network system

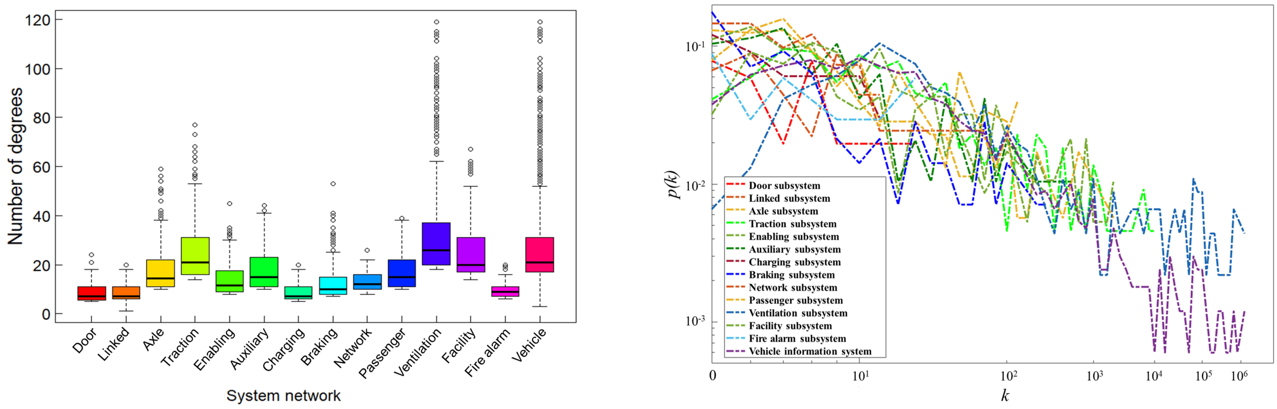
Passenger service system

Ventilation system

Facility system

Fire alarm system

Fig. 10 Connection structure of subsystem networks.



(a)

(b)

Fig. 11 Degree analysis of subsystems and vehicle information system. (a) Degree statistics and (b) degree distribution (logarithmic coordinates).

the network are more likely to fail. Therefore, we should focus more on measuring point failures with high average node strength.

5.3 Average clustering coefficient

Under the assumption that node i in the network has k edges connected to a total of k nodes, the k nodes can be connected to $k(k-1)/2$ edges at most. We use the number of real edges $\sum_{i=1}^{k} k_i$ to divide $k(k-1)/2$, so that we can obtain clustering coefficient c_i. The average clustering coefficient $\overline{C}$ is the average of the clustering coefficient of all nodes in the network,

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} c_i = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2\times\sum\limits_{i=1}^{k} k_i}{k(k-1)}\right) \tag{20}$$

The clustering coefficient can reflect the coincidence degree of the connection edge between any two nodes, that is, the degree to which the connection edge of the vehicle measuring points also connects with other

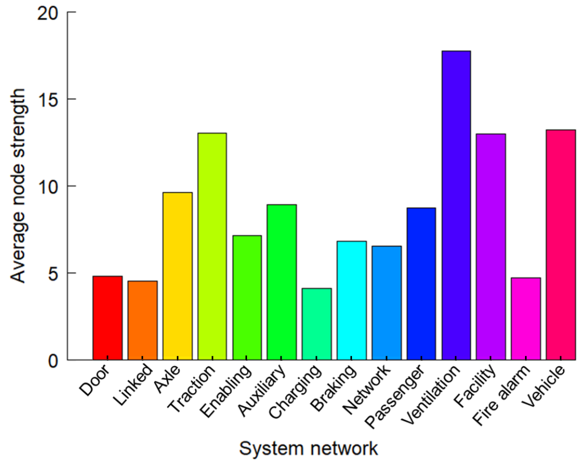


Fig. 12 Statistics on average node strength.

measuring points. Therefore, the average clustering coefficient can reflect the clustering degree of the network. The higher the average clustering coefficient, the higher the clustering degree of the network. We calculate the average clustering coefficients of 14 networks, and the results are shown in Fig. 13.

As shown in Fig. 13, the average clustering coefficient of the overall vehicle information network is 0.116, which is lower than the average clustering coefficient of all subsystems, indicating that the aggregation degree of the overall vehicle information network is weak, and we can cluster the network. The method of clustering is described in Section 5.4. As for the subsystem network, the average aggregation coefficient of the network system, passenger service system, and fire alarm system is higher than 0.25, indicating that the three networks are more aggregated than other networks. The value of the node can be directly derived from its neighboring node. Moreover, the higher the average clustering coefficient value, the higher the accuracy.

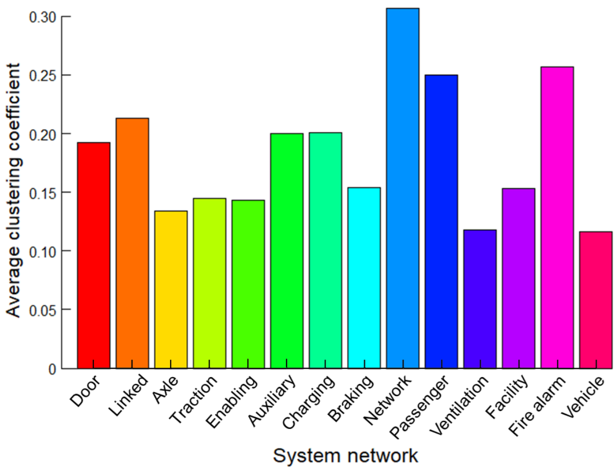
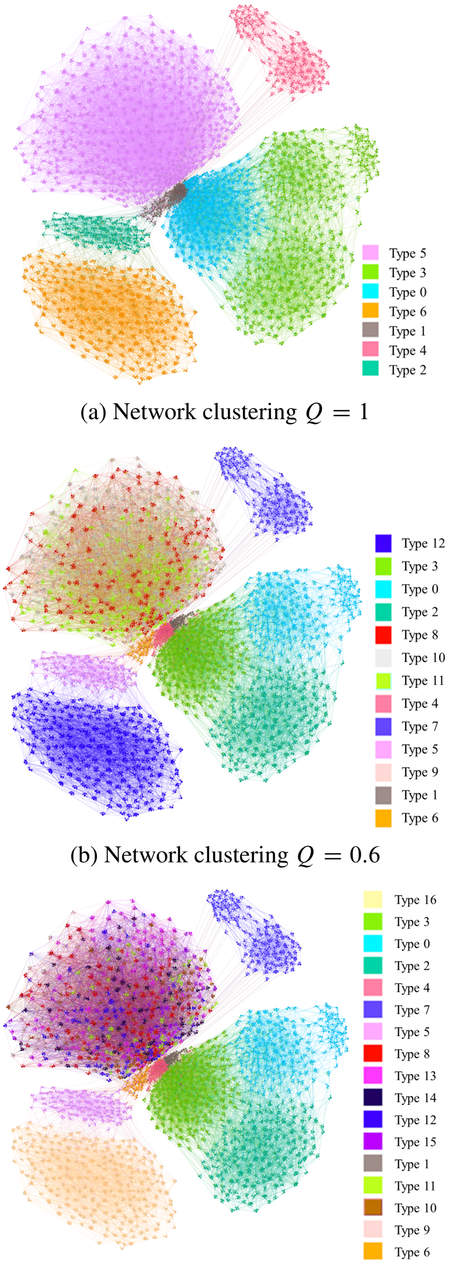


Fig. 13 Statistics on average clustering coefficient.

5.4 Community prediction

Community prediction is a method used to reveal network aggregation behavior. We use the method proposed by Lambiotte et al.[32], i.e., computing module degree Q, to calculate the sensor correlation network and obtain networks under different module degrees. When $Q = 1$, the sensor correlation network can be divided into seven modules, as shown in Fig. 14a. When $Q = 0.6$, it can be divided into 13 modules, as shown in Fig. 14b. When $Q = 0.4$, it can be divided into 17 modules, as shown in Fig. 14c. In Fig. 14, each type



(a) Network clustering $Q = 1$

(b) Network clustering $Q = 0.6$

(c) Network clustering $Q = 0.4$

Fig. 14 Analysis of different network clustering.

represents a classified community.

In the sensor correlation network, we compare the modules divided by community discovery with the modules divided by the system, and the following findings are obtained:

When $Q = 1$, the community discovery method treats some closely related systems as one class, for example, the network system and the passenger service system are not distinguished.

When $Q = 0.6$, the community discovery method divides the entire system into 13 communities, which is consistent with the actual number of systems. However, in two cases, they cannot be distinguished and overly distinguished in the 13 modules. For example, the braking system and door system cannot be divided into two modules, and the ventilation system is incorrectly divided into two modules.

When $Q = 0.4$, the method divides the system into 17 communities, but the ventilation system is divided into five communities. In conjunction with the discussion of the average aggregation coefficient, we find that the ventilation system can be easily segmented because of its low average aggregation coefficient.

6 Conclusion

We study the connection between the measuring points of the train from a new perspective. We first propose the GGC model, which can integrate the time series of different characteristics into a network structure. Then, we classify the 1673 measuring points on the train according to their functions and divide them into 13 subsystems. Next, we discuss the parameters in the GGC model and build a vehicle information network based on actual data. Lastly, we use complex network theory to mine the vehicle information network. Through this research, we find that each node in the vehicle information network is affected by approximately 24 nodes, and each of them affects these nodes simultaneously. In addition, the vehicle information network of each subsystem conforms to the characteristics of a scale-free network. Furthermore, the vehicle information network is weak against attacks, but the subsystem network is closely connected and is strong against attacks.

In the future, we hope to use the vehicle information network to predict the value of nodes. In practical applications, substantial vehicle information is returned to the data center and displays null values due to signal loss and sensor failure issues. To understand the operating status of the train fully, vehicle data must be supplemented. Given that link prediction can compensate for missing vehicle information, we will focus more on this direction.

Acknowledgment

This work was supported by the Graduate Innovation Project of Beijing Jiaotong University (No. 2020YJS098).

References

[1] P. Y. Pan, H. T. Hu, X. W. Yang, F. Blaabjerg, X. F. Wang, and Z. Y. He, Impedance measurement of traction network and electric train for stability analysis in high-speed railways, *IEEE Trans. Power Electron.*, vol. 33, no. 12, pp. 10086–10100, 2018.

[2] M. E. J. Newman, *Networks*: *An Introduction*. Oxford, UK: Oxford University Press, 2010.

[3] R. Albert and A. L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.*, vol. 74, pp. 47–97, 2002.

[4] G. Caldarelli, *Scale-Free Networks*: *Complex Webs in Nature and Technology*. Oxford, UK: Oxford University Press, 2007.

[5] M. E. J. Newman, Models of the small world, *J. Stat. Phys.*, vol. 101, no. 3, pp. 819–841, 2000.

[6] B. Hayes, Computing science: Graph theory in practice: Part 1, *Am. Sci.*, vol. 88, no. 1, pp. 9–13, 2000.

[7] G. P. Liu, Y. Q. Xia, J. Chen, D. Rees, and W. S. Hu, Networked predictive control of systems with random network delays in both forward and feedback channels, *IEEE Trans. Ind. Electron.*, vol. 54, no. 3, pp. 1282–1297, 2007.

[8] W. B. Du, M. Y. Zhang, Y. Zhang, X. B. Cao, and J. Zhang, Delay causality network in air transport systems, *Transp. Res. Part E*: *Logist. Transp. Rev.*, vol. 118, pp. 466–476, 2018.

[9] J. Du, J. Song, Y. Ren, and J. T. Wang, Convergence of broadband and broadcast/multicast in maritime information networks, *Tsinghua Science and Technolology*, vol. 26, no. 5, pp. 592–607, 2021.

[10] A. D. F. Santos, D. Valério, J. A. T. Machado, and A. M. Lopes, A fractional perspective to the modelling of Lisbon's public transportation network, *Transportation*, vol. 46, no. 5, pp. 1893–1913, 2019.

[11] F. T. S. Chan and H. J. Qi, An innovative performance measurement method for supply chain management, *Supply Chain Manage.*, vol. 8, no. 3, pp. 209–223, 2003.

[12] Y. Ma, G. Q. Cheng, Z. Liu, and X. X. Liang, Clustering-based link prediction in scientific coauthorship networks, *Int. J. Mod. Phys. C*, vol. 28, no. 6, p. 1750082, 2017.

[13] A. S. Khanna, P. Schumm, and J. A. Schneider, Facebook network structure and awareness of preexposure prophylaxis among young men who have sex with men, *Ann. Epidemiol.*, vol. 27, no. 3, pp. 176–180, 2017.

[14] X. Fei, Y. H. Zhang, and W. M. Zheng, XB-SIM: A simulation framework for modeling and exploration of ReRAM-based CNN acceleration design, *Tsinghua Science and Technolology*, vol. 26, no. 3, pp. 322–334, 2021.

[15] B. de Chassey, V. Navratil, L. Tafforeau, M. S. Hiet, A. Aublin-Gex, S. Agaugué, G. Meiffren, F. Pradezynski, B. F. Faria, and T. Chantier, et al., Hepatitis C virus infection protein network, *Mol. Syst. Biol.*, vol. 4, p. 230, 2008.

[16] F. Pelisch, T. Tammsalu, B. Wang, E. G. Jaffray, A. Gartner, and R. T. Hay, A SUMO-dependent protein network regulates chromosome congression during oocyte meiosis, *Mol. Cell*, vol. 65, no. 1, pp. 66–77, 2017.

[17] K. Osman, J. H. Yang, E. Roitinger, C. Lambing, S. Heckmann, E. Howell, M. Cuacos, R. Imre, G. Dürnberger, K. Mechtler, et al., Affinity proteomics reveals extensive phosphorylation of the Brassica chromosome axis protein ASY1 and a network of associated proteins at prophase I of meiosis, *Plant J.*, vol. 93, no. 1, pp. 17–33, 2018.

[18] P. H. Reingen, B. L. Foster, J. J. Brown, and S. B. Seidman, Brand congruence in interpersonal relations: A social network analysis, *J. Consum. Res.*, vol. 11, no. 3, pp. 771–783, 1984.

[19] W. M. Bowler and D. J. Brass, Relational correlates of interpersonal citizenship behavior: A social network perspective, *J. Appl. Psychol.*, vol. 91, no. 1, pp. 70–82, 2006.

[20] S. Pillai, K. Szekeres, N. J. Lawrence, S. P. Chellappan, and G. Blanck, Regulation of interlocking gene regulatory network subcircuits by a small molecule inhibitor of retinoblastoma protein (RB) phosphorylation: Cancer cell expression of HLA-DR, *Gene*, vol. 512, no. 2, pp. 403–407, 2013.

[21] M. Shibata and K. Sugimoto, A gene regulatory network for root hair development, *J. Plant Res.*, vo. 132, no. 3, pp. 301–309, 2019.

[22] C. G. Toenhake, S. A. K. Fraschkaet, M. S. Vijayabaskar, D. R. Westhead, S. J. van Heeringen, and R. Bártfai, Chromatin accessibility-based characterization of the gene regulatory network underlying Plasmodium falciparum blood-stage development, *Cell Host Microbe*, vol. 23, no. 4, pp. 557–569, 2018.

[23] P. C. Ma, B. Jiang, Z. G. Lu, N. Li, and Z. W. Jiang, Cybersecurity named entity recognition using bidirectional long short-term memory with conditional random fields, *Tsinghua Science and Technolology*, vol. 26, no. 3, pp. 259–265, 2021.

[24] D. Garlaschelli and M. I. Loffredo, Patterns of link reciprocity in directed networks, *Phys. Rev. Lett.*, vol. 93, no. 26, p. 268701, 2004.

[25] A. A. Tsonis and K. L. Swanson, Topology and predictability of El Niño and La Niña networks, *Phys. Rev. Lett.*, vol. 100, no. 22, p. 228502, 2008.

[26] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, The backbone of the climate network, *Europhys. Lett.*, vol. 87, no. 4, p. 48007, 2009.

[27] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, Complex networks in climate dynamics, *Eur. Phys. J. Spec. Top.*, vol. 174, no. 1, pp. 157–179, 2009.

[28] S. van den Elzen, D. Holten, J. Blaas, and J. J. van Wijk, Dynamic network visualization with extended massive sequence views, *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 8, pp. 1087–1099, 2014.

[29] Y. H. Sun, W. C. Duan, Y. Q. Li, and H. Xie, A construction method of spatial network of ANP-BOCR for complex systems, (in Chinese), *Chin. J. Manage. Sci.*, vol. 24, no. 2, pp. 144–152, 2016.

[30] Q. Zhang and Y. Ma, High-speed train service network features based on complex network, (in Chinese), *J. Railway Sci. Eng.*, vol. 15, no. 3, pp. 559–566, 2018.

[31] N. Wang, Futures time series characteristics of complex network and portfolio strategy research, (in Chinese), PhD dissertation, China University of Geosciences (Beijing), Beijing, China, 2016.

[32] R. Lambiotte, J. C. Delvenne, and M. Barahona, Laplacian dynamics and multiscale modular structure in networks, *Physics*, vol. 10, pp. 812–841, 2009.

Honghui Dong received the BEng degree from Xi'an Jiaotong University, Xi'an, China in 1999, the MEng degree from China Academy of Railway Sciences, Beijing, China in 2002, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, China in 2007. He is currently a professor at the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China. His current research interests include pattern recognition, intelligent detection technology, and transportation science and engineering.

Jie Man received the BEng degree from Beijing Jiaotong University, China in 2016. She is currently a PhD candidate in control system and engineering at the School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China. Her current research interests include artificial intelligence and rail transit safety early warning.

Limin Jia received the BEng degree from Tongji University, Shanghai, China in 1984, and the MEng and PhD degrees from China Academy of Railway Sciences, Beijing, China in 1987 and 1991, respectively. He is currently a professor at the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China. His research interests include transportation science and engineering, machine learning, and fault diagnosis.

Yong Qin received the BEng and MEng degrees from Tongji University, Shanghai, China in 1993 and 1996, respectively, and the PhD degree from China Academy of Railway Sciences, Beijing, China in 1999. He is currently a professor at the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China. His research interests include intelligent transportation diagnosis and health assessment.