Validity Analysis of Network Big Data

Peng Wang[1], Huaxia Lv[1], Xiaojing Zheng[1,*], Wenhui Ma[1] and Weijin Wang[2,*]

[1]*School of Economics and Management, Weifang University of Science and Technology, Shandong, China*
[2]*School of International Business, Shandong College of Economics and Business, Shandong, China*
E-mail: Tkwiloi_75@126.com; Zhengxj@wfust.edu.cn
[*]*Corresponding Author*

Received 08 March 2023; Accepted 07 April 2023;
Publication 04 July 2023

Abstract

False data in network big data has led to considerable ineffectiveness in perceiving the property of fact. Correct conclusions can be drawn only by accurately identifying and eliminating these false data. In other words, analysis is the premise to reaching a correct conclusion. This paper develops a big data network dissemination model based on the properties of the network. We also analyze the attributes of the big data random complex network based on the revised F-J model. Then, based on the scale-free nature of network big data, the evolution law of connected giant components and Bayesian inference, we propose an identification method of effective data in networks. Finally, after obtaining the real data, we analyze the dissemination and evolution characteristics of the network big data. The results show that if some online users intentionally spread false data on a large-scale website, the entire network data becomes false, despite a minimal probability of choosing these dissemination sources.

Keywords: Network big data, system dynamics, critical point, scale-free network, F-J model.

Journal of Web Engineering, Vol. 22_3, 465–496.
doi: 10.13052/jwe1540-9589.2234

1 Introduction

In the 21st century, technology and science have developed explosively, which has led traditional data technology being incapable of managing the large volume of data collection [1]. Since 2012, information technology has broadened its focus on big data and its validity, leading to the remarkable development of storing and processing big data technology. In this way, researchers have developed numerous methods to acquire, record, retrieve, share, and interpret big data and its validity. A considerable increase in the importance of big data networks has caused a substantial increase in the significance of their validity and accuracy [15]. Inaccurate big data networks are more detrimental than lack of data [24]. Therefore, big data networks and their validity is becoming increasingly important in human life.

Network big data refers to the big data generated by the interaction and integration of the ternary world of "human, machine, and things" in cyberspace, available on the Internet and called network data in brief [26]. These big data come from the data generated during network dissemination, which reflects people's ideology and network communication process. In other words, we can track and analyze the behaviors of the subjects (hereinafter referred to as big data participants) that generate the data through these data. However, in the process of network data generation and dissemination, the big data representing the personal opinions for certain events describe the irrational behaviors of corresponding people, such as their emotions, ideas, and preferences [14, 25]. Among these massive data, some reflect the big data participants' real intentions, some are trying to cover up the participants' real intention, and some are big data participants' simply copying others' intentions [2]. The latter two cases are called invalid data, which are seriously inconsistent with the event's true nature. The invalid data makes incorrect results, undermining the validity of the analysis, which is the reason why Google failed to predict influenza in 2013 [17]. The premise to fully utilize the opportunities and advantages brought by big network data is reliable, accurate, time-varied, and high-quality data, that is, effective data. Extracting implicit and valuable data from high-quality big network data paves the way for accurate decision-making; otherwise, the advantages of big data come to naught.

In addition to redundant invalid data, a large amount of unstructured data is in the massive network big data, such as pictures, text, video, and audio. Like structured data, unstructured data plays a crucial role in analysis, but it is challenging because of not only the challenges of unstructured data analysis,

but also the great uncertainty in the dissemination of network big data. Manual methods can generally transform unstructured data into structured data but only in small-sized cases [26]. Actually, analyzing the unstructured data case by case is highly demanding, requiring deep learning methods due to the massive nature of network data.

The network data process from generation to transmission can be abstracted into a complex system driven by human irrational behaviors. It is an open system, where each participant can fully express his or her own opinion or views [21]. These ideas or opinions are disseminated by other participants, which makes it a common phenomenon for multiple information to coexist on a certain network platform and share specific information from multiple platforms [8]. Scientists show that if participants highly support a viewpoint, they firmly deny other viewpoints, and vice versa [11]. For their highly affirmative views, they will actively disseminate them; otherwise, they will modify them and disseminate them or not. In other words, the degree to which an individual affirms the truth or falseness of interdependence and whether it spreads is controlled by the preferential connection mechanism of their interpersonal network [27]. In other words, an interpersonal network mechanism determines the certainty of individual' beliefs in interdependent true and false viewpoints [27]. With the elapse of time, individual views (nodes) in the above process will change, and participants (other nodes) who affirm or deny these views will constantly change, leading to changes in the transmission process (connections). Therefore, network big data is essentially a random complex network.

In recent years, many scholars have deeply studied this issue, and revealed the scale-free properties in the degree distribution, aggregation coefficient, and other parameters of the random complex network formed by network big data [3–6, 8]. This kind of scale-free random complex network displays robustness under random attack but vulnerability under intentional attack [6, 10]. This property shows that during the generation of network big data, rarely can part of the data fully reflect the real situation of the event, while the vast majority of the data are invalid [23]. This characteristic implies that identifying valid data in network big data can be abstracted as a critical problem by determining an appropriate critical probability to screening out the valid data [20]. Thus, the critical properties of complex networks are of great value in this paper.

Scientific research has proved that the distribution characteristics of the connected giant components in complex networks play an essential role in the operation of the system. When the rank of the connected giant components

of invalid data exceeds a certain number, invalid data spreads rapidly and occupy a dominant position in the network. Scientists insist that the big data in the core of the networks can describe the property of the event. To obtain the valuable big data, we could delete corresponding data in the periphery of the big data complex networks. There are two kinds of delete method, one is random attack, and the other is intentional attack. The former means that several big data deleted randomly with probability $1 - p$, and the latter means that several big data with strongest connected would be deleted with probability c. According to this method, we can judge whether the deleted data are essential to the event according to the rank of the giant component after attacking randomly or intentionally. This basis paves the way for further judging the authenticity of the data and finally identifying the valid data. It can be seen that the structure of network big data represented by connected giant components is crucial for system functions and valid data identification.

The analysis logic framework of this paper is as follows: when the network big data is deleted, see if the big data in the system can still guarantee whether this event can be followed in the network; if it can be followed, it indicates that there are still some key data not deleted, otherwise it means that all these key data are deleted. Those deleted data are critical for this event. Therefore we need to know the strength of the deletion; Theorem 2 shows the necessary condition for the complex network of big data to ensure the normal communication of big data systems under attack. This necessary condition refers to the maximum rate of deleting each big data from high to low resulting in this event not being noticed from the network after knowing the frequency of each big data being cited. Percolation theory proves that this approach is the most important way to find the real big data. As proved by Tai et al. [24] and Zheng et al. [27], we know that network big data satisfies the scale-free property, and this particular structure makes network big data critical. In this paper, we formally exploit its nature of robustness and vulnerability coexistence to get which part of the network big data is valid through intentional attacks and then separate the exact data from the kernel of network big data for analyzing a specific problem by the technique of Bayesian network inference with the help of some specific constraints. This is the biggest innovation of this paper.

To sum up, the structure is vital to the impact of the complexity of network big data, and the structure can change by random or intentional attacks. For network big data, this conclusion is also applicable. Intentionally deleting some data destroys the network function, revealing the critical role of deleted data in forming and disseminating network big data. The next issue

is whether these data are valid or not to identify the effectiveness of network big data. Based on this idea, we revise the Friedkin–Johnsen model to analyze the propagation and random complex network characteristics of network big data. Based on the characteristics of its random complex network and the percolation theory, we construct a simulation model to further analyze the evolution characteristics (structural changes) of network big data. Then, we propose a valid data recognition method based on the evolution characteristics of network big data and Bayesian inference. Finally, we deduce the effectiveness of network big data, considering the Huang Haibo incident as an example.

2 Friedkin–Johnsen (F-J) Model of Network Big Data

2.1 Generation and Dissemination of Network Big Data

As mentioned above, the dissemination of network big data is actually the evolution process of a specific topic from one person to another. The process is shown as follows. First, a person puts forward an idea on the Internet, which others will see on the network. If they are interested in the idea, they will judge the big data; there are two resulted judgments: right and wrong. If the big data participant accepts the idea, they will spread the idea with a certain probability (e.g., publishing in Weibo or WeChat moments), leading to the spread of the big data in the network. If they disagrees with this view, they will sniff at it in some probability, or refute and correct it in another probability, and then spread it. The above behaviors of many big data participants form network big data. In either case, the network big data will change dynamically in terms of quantity, attribute, and structure, forming a random complex system. Figure 1 displays the decision-making process of the big data participants.

With regard to Figure 1, each participant has their own thoughts due to the person's subjective initiative and irrationality. The participants will judge, revise, and spread the network big data according to their own values, world perspective, and life outlook. Thus, a specific event generates data with varying properties. After these data are spread in the network for a while, they will form a complex networks. Figure 2 illustrates this complex network as a definite graph at a particular time.

In Figure 2, the left side is a simulation diagram of the formation of network big data after two opinions spread in the network for some time, and the right side describes the interaction among individuals (called an

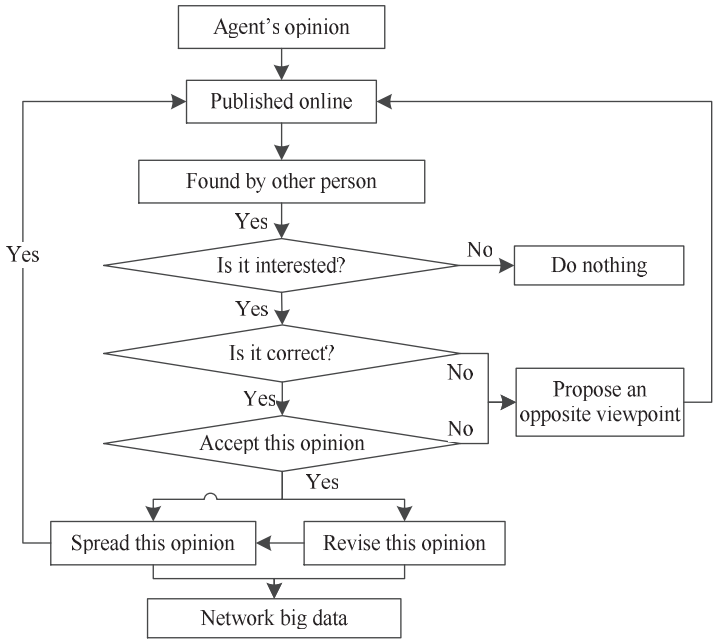

Figure 1 Generation process of network big data.

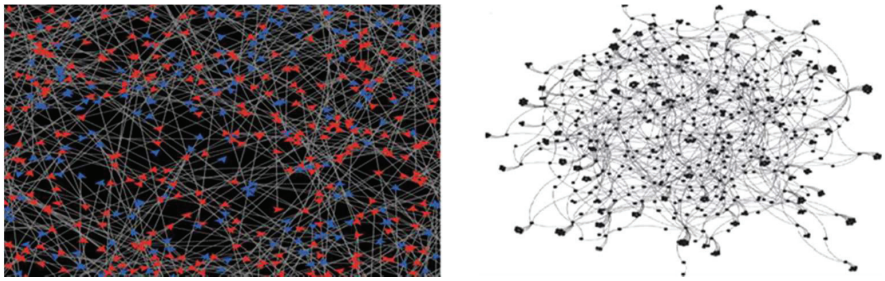

Figure 2 Two different viewpoints in the network: simulation and schematic diagram.

agent). During the interaction process, because of the difference in the agent's values, world view, and life outlook, agents will generally produce network big data consistent with their characteristics after analyzing an event, and also can identity, learn, copy, spread, oppose, and modify the network big data, including their own opinions. However, the interaction among agents is different, which depends on the local spatial topology of the interaction between agents. The unequal interaction among agents makes the network

big data show the characteristics of diversity and difference. Hence, some big data can truly reflect the characteristics of the event, while some hide the attributes of the event, or even wholly subvert the real nature of the event. Therefore, studying the evolution characteristics of network big data becomes very critical for identifying valid data. This paper adopts the F-J model to describe the evolution characteristics.

2.2 F-J Model of Network Big Data Dissemination

As mentioned above, network big data is a complex network describing the propagation of different views through different agents in the network. The most classic dynamic model to describe its spread is the Friedkin-Johnsen model (hereinafter referred to as the F-J model). We consider big data in the network to refer to the collection of individual views on an event, each of which is different and each of which changes dynamically. This process can be accurately described by the classical F-J model. However, the F-J model incorporates the characteristics of the network structure and the propagation of viewpoints among people. This model supposes that each agent can form their initial opinion on a particular event to some extent, and this opinion changes dynamically and randomly with the elapse of time, restructuring the system and the neighbors' views [11]. Furthermore, the F-J model also discusses a situation where a causal relationship exists among a series of dependent viewpoints and then analyzes the evolution laws of these viewpoints in the system, which makes this research more universal. Equation (1) represents the classical F-J model.

$$X(t+1) = AWX(k)C^T + (I - A)X(0),\ k = 0, 1, \ldots \tag{1}$$

where $X(k)$ is a $n \times m$ matrix representing the viewpoints of n agents on m interdependent subjects. W is a $n \times n$ matrix describing the degree of the direct influence of agents, while disconnected agents have no influence. Matrix W determines the diagonal matrix A, and $a_{ii} = 1 - w_{ii}, \forall i = 1, 2, \ldots, n$. The matrix describes the "surplus" influence in addition to the agents' own influence. Matrix C is a $m \times m$ matrix describing the relationship among m interdependent subjects and I is an identity matrix.

To further explain this model, it is assumed that there are n agents in the network. Due to different local topological structures of each agent, there are complex nonlinear effects among agents, which reflect the transmission intensity of views, denoted as w called weight. w_{ij} is assumedly the weight of agent i passing their viewpoint to agent j, and w_{ii} is that of agent i passing

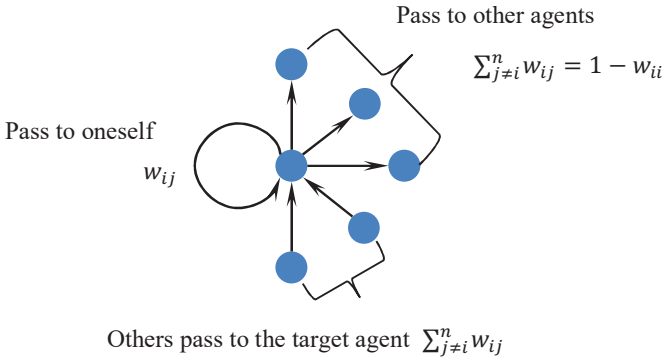

Figure 3 Properties of w_{ij}, where $0 \leq w_{ij} \leq 1, \sum_{j=1}^{n} w_{ij} = 1$.

their viewpoint to themselves as feedback. Figure 3 shows the dissemination characteristics.

According to Figure 3, network big data is actually a directed one. If agent i passes their viewpoint to agent j according to the proportion w_{ij}, this distribution of past viewpoints is $j, i \xrightarrow{w_{ij}>0} j$. If agent i passes the viewpoint to themself for feedback (i.e., keeping it for self-reflection and providing conditions for future revisions), node i forms a loop, denoted as $i \xrightarrow{w_{ii}>0} i$. The influence that other agents pass their viewpoints to agent i is $\sum_{\mathrm{j}\neq\mathrm{i}}^{\mathrm{n}} \mathrm{w_{ji}}$. Obviously, the influence w_{ij} changes randomly over time, resulting in a random matrix W, expressed as follows:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}$$

Its diagonal elements $\{w_{11}, w_{22}, \ldots, w_{\mathrm{nn}}\}$ describe the self-feedback ability of agents, developing Equation (2).

$$w_{ii} = 1 - \sum_{j\neq i} w_{ij} \tag{2}$$

where $1 - w_{\mathrm{ii}}$ is the total relative influence of an agent passing viewpoints to others.

Generally, there are several different subjects in the network, related or independent. Suppose m subjects in the network, whose relationships with

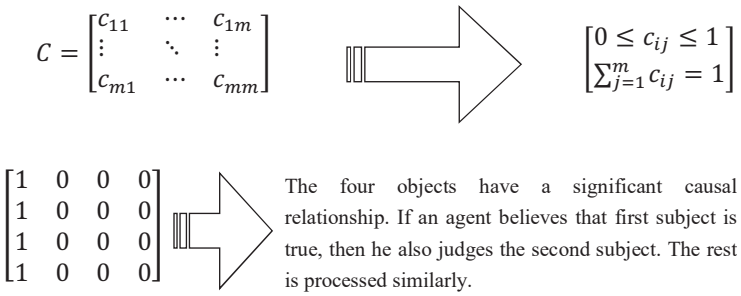

Figure 4 Logical relationships among subjects.

logical constraints rely on the matrix C.

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1m} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mm} \end{bmatrix}$$

where c_{uv} represents the logical relationship between subject u and subject v. $\sum_{v=1}^{m} c_{wv} = 1 \forall v = 1, 2, \ldots, m$ for all $u = 1, 2, \ldots, m$. This matrix is random. If these subjects are unrelated, then $c_{uu} = 1, c_{uv} = 0, u, v = 1, \ldots, m, \forall u \neq v$. If they form a complete causal relationship: $c_{u1} = 1$, $\forall u = 1, 2, \ldots, m$. Therefore, the property of matrix C represents the logical relationships among the subjects, as shown in Figure 4.

Figure 4 represents the logical relationships among subjects in the matrix. According to this logical relationship, we can determine the dependence between the two subjects. In other words, c_{uv} depicts the possibility of mutual support between the viewpoints of corresponding subjects. $c_{uv} = 1$ means that subject u is the cause of subject v. Obviously, because of the asymmetry of the logical relationship between subjects, that is, $c_{uv} \neq c_{vu}$ for any two subjects.

Moreover, this logical relationship is certain under specific constraints. For any subject j, agent i completely believes, denies, or doubts the authenticity of a viewpoint at moment k, represented by $w_{ij}(k) = 1, w_{ij}(k) = 0$, and $\cdot w_{ij}(k) = 0.5$, respectively. However, time is of high significance due to the dynamic property of this viewpoint. In other words, any agent's viewpoint randomly changes over time, which is a random process in essence, as presented in Figure 5.

Therefore, as time goes by and new subjects enter, the logical relationships change between subjects and their structure. When the environment

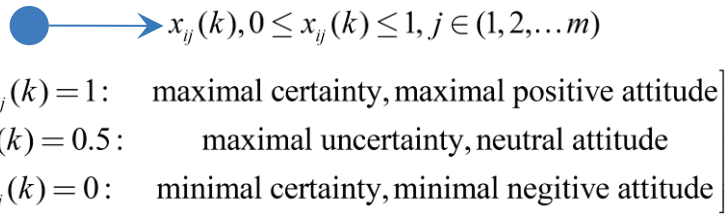

Figure 5 The value regarding the judgment of agent i on the subject j at a particular moment k.

changes, the constraints will change, and so will the logical relationships. These changes make this research extremely difficult because of such complex and changeable subjects and their independent relationships.

2.3 Revision of the F-J Model Based on Network Big Data

The classical F-J model presented above has a considerable limitation in studying network big data. First, network big data should be an open system, with the participants' numbers and structure changing simultaneously. Second, the interaction among agents is selective and determined by the priority connection mechanism of the network [3], rather than the completely random selection described by the F-J model [12]. Moreover, agents' perception regarding a particular attribute behavior may change suddenly, not a fixed value described by the F-J model, which affects the generation of corresponding big data. These factors lead to the gaps between the constraints of the classical F-J model and reality [8]. Hence, this paper optimizes the F-J model, where the orders of matrix *W*, *X*, *A* are no longer constant, but randomly change with time, which makes it closer to reality and more accurately explains the dissemination process of network big data. Equation (3) represents the random time-varying form of Equation (1) [27].

$$\begin{aligned} \|X(t+1)_{n(t)\times m(t)}\| = {} & \|A_{n(t)\times n(t)} W_{n(t)\times n(t)} X(k)_{n(t)\times m(t)} C^{T}_{m(t)\times m(t)}\| \\ & + \|(I_{n(t)\times n(t)} - A_{n(t)\times n(t)}) X(0)_{n(0)\times m(0)}\| \\ & + \|\varepsilon(t)_{n(t)\times m(t)}\| \end{aligned} \tag{3}$$

where $k = 0, 1, \ldots$. Equation (3) is an F-J model with a random time-varying structure. Both the number and structure of agents and the subjects in network big data change with time, but the time required for the changes is different. Furthermore, the issues are changed dynamically with time, and people who

scan, use, revise, and innovate are also changed. Some people receive more than one issue at simultaneously, and they can transfer their belief to the next one dynamically and randomly, which makes $n(t)$ and $m(t)$ the mapping of time *t*. In this process, population of big data, issues, connection between each big data are all random. According to the characteristics of network big data, any subject needs a certain amount of time for changing. Thus, we assume that the changes in network big data and subject are slowly time-varying. Specifically, $m(t)$ is constant for a long time. Beyond the critical time point, new subjects appear in the system. In contrast, $n(t)$ changes faster, describing that the total number of agents discussing these subjects is changing all the time in a specific subject structure. Obviously, model (3) can describe the reality more precisely than model (1), but is more complex to calculate. This is the innovation of the F-J model. This implication shows that a certain period can be compressed into a time point to coarse-grain the time and form a random sequence in the structure. If *t* and *L* are, respectively, the length of time for changing the subject network big data can generally be viewed as a number of processes and, in each of them, the properties of the data are consistent. However, when the evolution of big data exceeds a certain critical point, its properties change. Therefore, we can ignore its data if it changes microscopically, and abstract the evolutionary process of this big data from a macroscopic level into a more coarse-grained temporal change process by coarse-graining techniques; a more coarse-grained temporal model can be used to implement this process, we vary the time according to the changing characteristics of the data as follows:

$$\underbrace{0,1,\ldots,L_1}_{t_1},\underbrace{L_i+1,\ldots,L_2}_{t_2},\underbrace{\ldots\ldots}_{\ldots\ldots},\underbrace{(L_{i-1}+1),\ldots,L_i}_{t_i},\underbrace{\ldots\ldots}_{\ldots\ldots},i\in 1,2,\ldots. \tag{4}$$

At coarse-grained time scales, some changes in the state, structure and properties of the data will occur. However, the unequal time interval L_i at each microscopic granularity leads to a difficult analysis of this problem. This problem is tough to solve, exceeding the research object described by Friedkin [12]. Although this model is very close to reality, the conventional methods create inconclusive results. Reanalysis of this problem can be abstracted as a complex adaptive system, of which the function, structure, and property not only change with the variations of the environment, but also affect the attributes of the environment. Therefore, we introduce the analysis method of the complex adaptive system to explore this problem.

3 Dissemination Process and Law of Network Big Data

As illustrated previously, data reflects the propagation of a thought, which exhibits characteristics such as randomness, diversity, and dynamics, showing people's learning process. Therefore, a key issue is how people learn and improve data and spread it in the network during the process, namely the evolution law of network big data. Obtaining the evolution law of the system helps people identify which big data is valid and which is not. This part analyzes the evolution law of network big data.

3.1 Scale-free Characteristics of Network Big Data

Studies claim that in most networks few nodes occupy most connections of the whole network while most nodes have few connections, and the distribution of the number of nodes and connections accords with the power-law distribution. Scientists discovered many network structures satisfying the power-law distribution in scale-free networks. The power law is ubiquitous, especially in complex networks [3]. Studying the frequency of English words shows that people use only a few words frequently, but the vast majority of words rarely. The economist Alberto Pareto studied the statistical distribution of personal income and found that a minority earns far more than the majority. He devised the famous 80/20 rule, in which 20% of the population has 80% of society's wealth. In other words, the Pareto principle emphasizes the influential minority and the trivial majority, confirming that the world is full of imbalances. Thus, the above scaling-free network generation mechanism considers two assumptions. The first assumption is the growth mechanism, i.e., networks constantly generate new nodes over time. The second assumption is the priority connection mechanism, implying that newly added nodes prefer to attach to nodes of more connections [3]. Similar rules have been rediscovered in the Internet era. For example, the number of fans of all users on the WeChat public account, Weibo, and other online platforms roughly conforms to the power-law distribution, indicating that a small number of users have a large number of fans. According to the previous studies, a scale-free complex network has been formed among the participants of network big data, in which agents generate and modify big data, forming the propagation of big data in the scale-free network [3, 4, 18] and ultimately leading to the emergence of different types of big data in the network [16]. Figure 6 presents the evolution results of network big data after a while.

Figure 6(a) represents the nature of the topological structure formed by big data, and 6(b) depicts the power law distribution features obtained

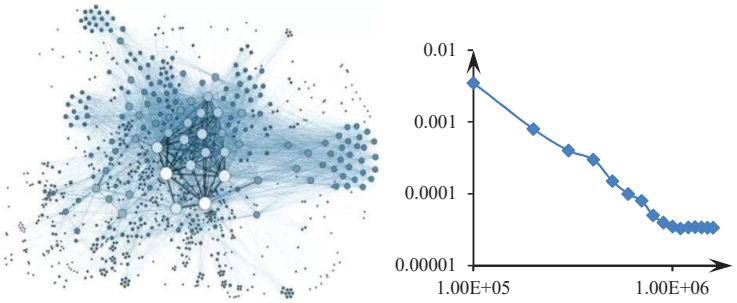

Figure 6 Power law characteristics formed during the evolution process of network big data.

through statistics of connections in big data. In Figure 6(b), the horizontal axis represents the logarithm of degree, and the vertical axis represents the logarithm of distribution probability. According to Figure 6, scale-free propagation of big data exists in the network, i.e., a node with high connections is more likely to obtain new connections than a node with fewer connections.

3.2 Connected Giant Components of Network Big Data

In order to determine which of the big data corresponding to the total events is valid, we need to identify some key data. The idea is as follows: according to the evolution process of the complex network of big data and its characteristics, the kernel data represent the overall viewpoint of the event. Then we can determine the distribution characteristics of several different views and, according to these distribution characteristics, we can eliminate a part of the data and determine whether the statistical properties of big data change significantly according to Bayes' theorem, and then we can determine which data is true, collect such data inside the nucleus for analysis, and dig out the true properties and evolutionary laws of the events. Furthermore, A large component refers to a connected subgraph formed by the same viewpoint data, indicating the distribution of the same viewpoint in space. If the statistical distribution of data of different nature is indicated from a statistical point of view, we can take Bayesian inference to determine which data is true when and only when this statistical distribution is especially determined. The property of a connected giant component of homogeneous big data helps describe the property of network big data in a system. Under general circumstances, when an event occurs a large amount of big data will inevitably be formed in the network. These big data, roughly divided

into several limited kinds, contradict or have causal relationships. However, the others quote, refute, revise, and disseminate each type of big data [13]. The network evolution divides the big data with a particular property into several connected sub-graphs, namely connected giant components, whose properties play an essential role in determining the dominant viewpoint in the system [19]. Generally, even if a viewpoint is invalid, it soon becomes dominant in the whole system when its proportion exceeds a certain threshold [28]. When invalid data in the network form connected giant components during the evolution process, especially when the order of such connected giant components reaches a certain degree, invalid data rapidly diffuse in the whole network and bury valid data to create a result that "mix the spurious with the genuine". This phenomenon happens commonly in the fields affected dramatically by the media, such as the entertainment and medical industries. Therefore, it is essential to analyze the distribution characteristics of connected giant components in homogeneous big data during data evolution for data screening. According to percolation theory [28], we construct a corresponding simulation model. The simulation results show that the group agents always follow the principle "birds of a feather flock together" to form a connected giant component during the evolution process of viewpoints in a system, due to the interactions among agents. If and only if the order of the giant component exceeds a specific value, the number of agents in the system agreeing with the viewpoint increases immediately. Figure 7 displays the variation process of a giant component during the propagation of a viewpoint in a scale-free network.

Figure 7 illustrates that initially accepting a specific viewpoint is hard, but is propagated in the system as all agents communicate with each other. According to the vertical line in Figure 7, the vast majority of people may quickly recognize the minority opinion in the network, forming a "reversal" phenomenon. This phenomenon is a critical point, representing the minimum amount of effort to diffuse a specific viewpoint. Furthermore, the minimum effort actually corresponds to the key population at the critical point. By contrast, to control some rumors, the proportion of rumor-holders should be under this critical point [28].

3.3 Identification of Valid Network Big Data Based on Connected Giant Components

According to the previous analysis, in the process of network big data transmission, data of whatever nature will form different connected giant

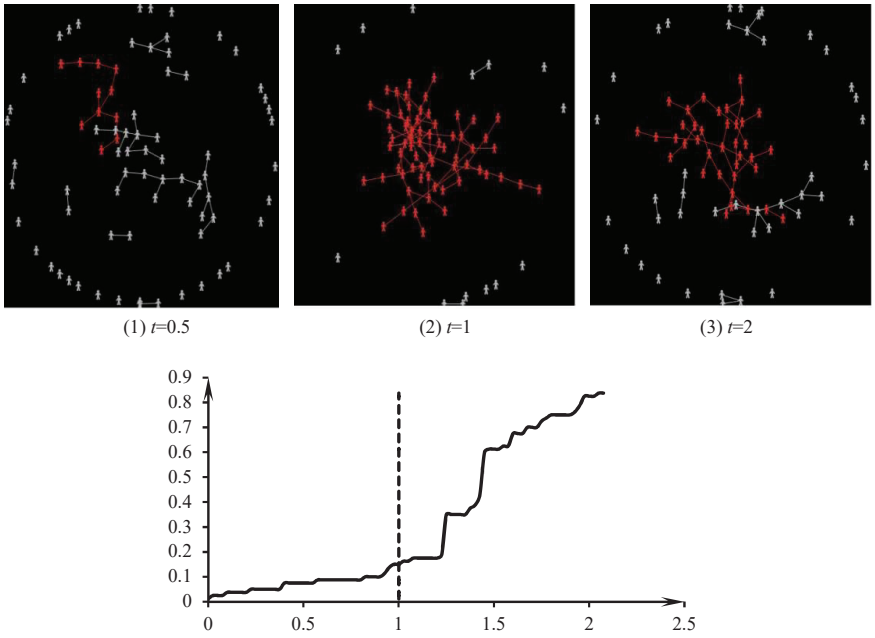

Figure 7 Variation of a giant component during a specific viewpoint's propagation in a system.

components. Those big data in a "free state" excluded from the connected giant components can be ignored. Even if the data in the connected giant components cannot necessarily be used, only a part of the data within the connected giant components is valid, while the rest can be deleted. We analyze how to select this part of valid data.

A hypothesis claims that some data are invalid in the system. In this sense, some big data should be deleted from the population. Then, which data should be deleted is the question. In some sense, the valid big data can be obtained by deleting the invalid ones according to percolation method. Firstly, we use the evolution characteristics of network big data to abstract it into a complex network, as defined above.

To do thus, following equation defines the priority connection probability between data, mainly considering the strength.

$$p_i = \sum_{j \in \overline{\mathcal{N}}_i \backslash \{\mathrm{i}\}} \pi_j$$

where π_j represents the strength of citing and being cited of data j. If the data is disseminated to an authoritative website, its value has a greater weight; otherwise, it has a lower weight. The citation numbers of data measure its strength. $\overline{\mathcal{N}}_i$ represents the set consisting of data i and the data directly connected to it. $\overline{\mathcal{N}}_i/\{i\}$ is the set of neighbors of data i. Based on this priority connection, network big data reflects significant "rich club" characteristics.

Suppose a specific moment when network big data has formed several connected giant components with different opinions, intertwined and distributed in the social network platforms. This research tests whether the network big data system can remain connected by deleting some data. In a positive case, the remaining data can fully describe the network unstructured big data system corresponding to the event and reflect the event's nature. In a negative case, the remaining data cannot describe the properties of the event. Therefore, the question becomes: how and with what probability to delete the data, can we find these essential data? There are two ways to delete: random attack and intentional attack. Random attack means that some data is randomly selected and retained with probability p and deleted with the rest of the big data. Intentional attack sorts these unstructured big data from large to small according to their strength, then deletes the strongest big data with probability c. The actual operation needs to constantly change the probabilities of p and c to judge the properties of the system, a process which has been analyzed in previous studies [22]. The research results prove that under random attack, network big data is always connected, but under intentional attack, a critical probability c_0 exists that can make the system disconnected. Based on the results, finding valid big data with random attack is difficult, while finding them with intentional attack is easy. Thus, the key is to determine this critical probability. Further research shows that, under random attack, the agents in the network are deleted with a significant probability of $1 - p$, while the rest are kept with a probability of p.

By invoking the result of Zheng [28], if the system of big data is deleted randomly, there are at least two connected giant components that still exist in the network. The order of the largest connected giant components is $\mathrm{L}_1(p)$. and that of the order of the second largest connected giant component is $\mathrm{L}_2(p)$. Furthermore, $\mathrm{L}_1(p)$ and $\mathrm{L}_2(p)$ satisfy the Theorem 1.

Theorem 1. As for $0 < p < 1$, there exists a function $\lambda(p) > 0$ to obtain $L_1(G_p) = (\lambda(p)+o(1))n$ and $L_2(G_p) = o(n)$ with a probability of $1-O(1)$.

As for any small probability $p > 0$, a huge component guarantees the system's robustness. The property $\lambda(p)$ confirms the order size of the huge

component. Further, if $p \to 0$, then,

$$\exp(-\Theta(1-p^2)) \leq \lambda(p) \leq \left(1+5d\sup_{\alpha\in\mathcal{A}} E[\pi]p/8\right)\exp\left(-1/\sup_{\alpha\in\mathcal{A}} E[\pi]p\right)$$

where d is the average number of times that big data is spread, and $\mathrm{g(x)} = \Theta(\mathrm{f(x)})$ means both functions $g(x)$ and $f(x)$ satisfy the formulas $g(x) = O(f(x))$ and $f(x) = O(g(x))$.

Based on Theorem 1, two giant components maintain their connectivity in the network big data, despite reserving only a small minority of agents when the network big data is under random attack. The minimum diameter of the largest one is $\mathrm{nexp}(-\Theta(1-p^2))$, and the $\cdot$ maximum diameter is $(1+5\mathrm{dsup}_{\alpha\in\mathcal{A}}\mathrm{E}[\pi]\mathrm{p}/8)\exp(-1/2\mathrm{sup}_{\alpha\in\mathcal{A}}\mathrm{E}[\pi]\mathrm{p})$, while the diameter of the other one is $o(n)$.

According to Theorem 1, it is concluded that, even if almost all big data can be retained, the function of this system would have taken effect. That is to say, the strategy of random attack of big data makes no sense, so, we consider intentionally deleting the big data. In the case of intentionally deleting the network big data (i.e., deleting a total of $c_0 n$ agents with the most connections in the network based on a critical probability c_0), according to the result of Zheng [28], the system shows robust criticality. The corresponding conclusion should be described as shown in Theorem 2.

Theorem 2. Let $0 < c < 1$, there exists a constant $c_0 = \frac{q_3-q_4}{1+\delta_{in}(q_1+q_2-q_5)} \frac{inf_{a\in\mathcal{A}}E[\pi]-1}{sup_{\alpha\in\mathcal{A}}E[\pi]+1}$. If $c \geq c_0$, the formula $L_1(G_c) = o(n)$ has a probability of $1-o(1)$. Moreover, if $\mathrm{c} < \mathrm{c}_0, \theta(\mathrm{c})$ is a positive constant in the formula $L_1(G_c) = (\theta(c)+o(1))n$ and $L_2(G_c) = o(n)$ with a probability of $1-o(1)$.

Where q_1 represents the probability of selecting a big data on the same social platform, q_2 is the probability of disconnecting with big data on the same social platform, q_3 denotes the probability of selecting big data across platforms, q_4 indicates the probability of deleting the cross platform big data connection, and q_5 shows the probability of deleting big data from the platform. The summation of these q is equal to one, $\sum_{\mathrm{i}=1}^{5} q_i = 1$. Furthermore, some big data is actively connected because agents in the social network platform recognize the data and are willing to accept the benefits brought by the data. Some big data is abandoned or deleted because the meaning expressed by the data goes against the concept of agents (including the platform).

Theorem 2 assumes that if the deletion probability is less than the critical probability c_0, two giant components keep most agents connected in the network, of which the larger one has an order of $(\theta(c)+o(1))n$, and the other has an order of $o(n)$, resulting in solid robustness in the network big data. If the deletion probability is higher than c_0, then the network big data leaves only one giant component, whose order is $o(n)$, eventually damaging the complex system's functions. Moreover, two determinants of critical probability are the average number of times that the corresponding data of the node spread and the local topological structure. Using the network big data characteristics and Theorem 2, first calculate the critical probability of network big data percolating, then confirm the strength of each data according to the flow and degree, and sequence the data by strength from large to small. Therefore, $c_0 n$ data are valid network big data. According to Theorem 2, it is easy to see that even critical probability c is very small, even close to 0.2, and the most important big data in this system could be found, i.e., a very small number of big data can describe the property of the system, and most big data can be regarded as the disturbance.

By Theorem 2, we get the most valid data. The network forms different homogeneous giant components after a while because these core valid data are in different homogeneous giant components, and all the data in these homogeneous giant components generally express the same viewpoint. These homogeneous giant components are intertwined in the network, forming the whole network big data. In this process, the data with different properties form the connected components in turn to shape a random process of their orders. Thus, some connected giant components with specific properties change randomly in the network. Note that the complex stochastic networks were proved in some detail in our previous paper [28] and are correctly cited in this paper, and the procedure is omitted in view of the limited space and the subject matter.

Furthermore, as a particular viewpoint evolved, the connected giant component is finally formed one by one under the premise of "birds of a feather flock together". For example, Figure 8 displays the formation of connected giant components in the network at a specific moment.

Figure 8 illustrates two different viewpoints that appeared in the network, the red and blue giant components. At moment *t*, the component lengths of the red and blue big data are 5 and 3, respectively. Nevertheless, at the moment $t + 1$, the connected giant component of the red big data splits into two components, each with a length of 2, while the blue big data has one component with a length of 2. The variation is the result of agents' behaviors

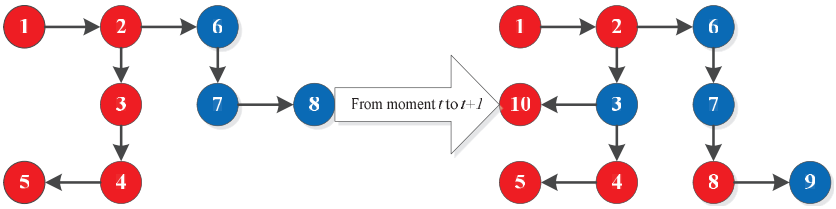

Figure 8 Giant components of network big data and their changes.

in the network. Under this circumstance, the red data is true (valid). Thus, this variation develops a random process model, in which the observation subject is the giant component to analyze the distribution characteristics of the giant component.

In the design of this paper, our idea is as follows: first find the set X of data inside the kernel according to Theorem 2 and get the form of statistical distribution of the data, then delete a few key suspected wrong data and then determine its statistical distribution. If the two statistical distributions are completely different, then this indicates that the data deleted are very important, i.e., the one that was deleted was correct. By processing data of different nature in this way, and comparing them according to the magnitude of significant differences in probability changes, it is possible to determine which data are accurate.

Despite the importance of the resulting data, they are not necessarily valid. The method of Bayesian inference examines their truth or falseness. In general, the initial formation of the data can describe the truth of the event, forming causality in the dimensions of time and space between the original and derived data [7]. Hume insists that the relationship between two similar objects in temporal precedence and spatial proximity are called causality. Based on this causality, we can adopt causal inference to find valid big data consistent with the nature of the event and separate invalid data [9]. This study employs the Bayesian inference method expressed as follows:

$$\mathbb{P}[Y(t+1) \in \mathcal{A}|\mathcal{I}(t)] \neq \mathbb{P}[Y(t+1) \in \mathcal{A}|\mathcal{I}_{-X}(t)]. \tag{5}$$

The Bayesian inference method analyzes the conditional probability of historical data based on their evolution characteristics. If the distribution is different from that in the next stage, then invalid data is generated. Equation (5) shows the Bayesian inference method for real data analysis. The main idea is as follows: first find the data set X inside the kernel according to Theorem 2 and get the form of statistical distribution of the data, then delete a

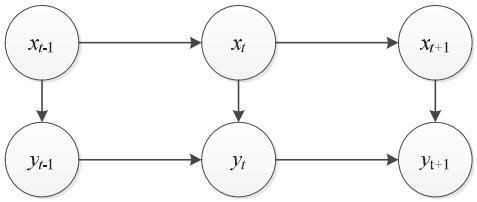

Figure 9 The property of stochastic process of giant component of complex networks of big data.

few key suspected wrong data and then determine their statistical distribution. If the two statistical distributions are completely different, then this indicates that the data deleted are very important, i.e., the one that was deleted was correct. By processing data of different nature in this way, and comparing them according to the magnitude of significant differences in probability changes, it is possible to determine which data is accurate.

Specifically, each connected giant component is a random process, and the authenticity of these data satisfies the following characteristics.

In Figure 9 the authenticity of data *x* satisfies the equation

$$P(x_t|x_1, x_2, \ldots, x_{t-1}) = P(x_t|x_{t-1}).$$

The following conditions are met to determine whether x is a real event.

$$P(y_t|x_1, x_2, \ldots, x_{t-1}, y_1, y_2, \ldots, y_{t-1}) = P(y_t|x_t)$$

Under the action of parameter *w*, the relationship between data *x* and *y* is as follows.

$$y_i = f(x_i|w) + N(0, \sigma^2).$$

Parameter *w* has the following properties

$$P(w|x, y) = \frac{P(x, y, w)}{P(x, y)} = \frac{P(y|w, x)P(w, x)}{P(y|x)P(x)} = \frac{P(y|w, x)p(w|x)p(x)}{P(y|x)p(x)}$$

$$= \frac{P(y|w, x)p(w)}{P(y|x)} = \frac{P(y|w, x)p(w)}{\int_w P(y|w, x)p(w)}$$

Therefore, the following Bayes' theorem obtains the truth-falsity of the data corresponding to the connected giant component.

$$\underbrace{P(\theta|x)}_{\text{posterior}} = \frac{\overbrace{P(x|\theta)}^{\text{likelihood}}\ \overbrace{P(\theta)}^{\text{norior}}}{\underbrace{P(x|\theta)}_{\text{normalization constant}}} = \int_\theta \frac{P(x|\theta)P(\theta)}{P(x|\theta)P(\theta)}$$

Furthermore, we can obtain a posteriori inference.

$$\begin{aligned}
&\inf_{q(\theta)\in Q}\{KL(q(\theta)\|P(\theta)) - E_{\theta\sim q(\theta)}\ln(P(x|\theta))\} \\
&\quad = \inf_{q(\theta)\in Q}\left\{\int_{\theta}\ln\frac{q(\theta)}{P(\theta)}q(\theta) - \int_{\theta}\ln P(\theta)q(\theta)\right\} \\
&\quad = \frac{1}{P(x)}\inf_{q(\theta)\in Q}\left\{\int_{\theta}\ln\frac{q(\theta)}{P(\theta)}q(\theta)\right\} \\
&\quad = \inf_{q(\theta)\in Q}\{KL((q(\theta)\|P(\theta|x)))\}.
\end{aligned}$$

The above procedure judges the authenticity of the data.

4 Example Analysis: Valid Data Identification of Actor Huang Haibo's Prostitute Scandal

In 2014, Chinese actor Huang Haibo was involved in a scandal. At first, the Internet suddenly spread rumors that Huang Haibo had visited a sex worker, then he apologized. Then, the Internet spread rumors that Huang Haibo had disappeared from the public eye, and later the Internet talked about Huang Haibo being wronged. During the course of this event, the following four themes gradually emerged:

(1) Huang Haibo visiting a sex worker is well-known in the performing arts circle and a manifestation of the chaos in the performing arts circle.
(2) There is a big difference between apologizing in public and trying to hide, which reflects the actor's moral quality.
(3) Is there such a regulation that sex workers are detained for six months?
(4) Huang Haibo was framed for visiting a sex worker.

In the whole course of the event, there was only the first subject at the beginning, and finally, there were four subjects intertwined with each other. The Internet has two camps of statements for different subjects (i.e., yes and no), and it is difficult to determine which are true or false on the surface. In addition, a special relationship exists among the four subjects. The development process of the event shows that some subjects (topics) constantly appear with the elapse of time. The structure of the relationship matrix C changes dynamically because of the constant change in these subjects. Weibo gave some corresponding data using a data crawler, and determined corresponding parameters according to these data. We further analyze these

data to obtain the logical relation matrix *C* of subjects. Matrix *C* has the following properties.

$$C(t=1)=\begin{bmatrix}1 & 0\\ 1 & 0\end{bmatrix},\ C(t=2)=\begin{bmatrix}1 & 0 & 0\\ 1 & 0 & 0\\ 1 & 0 & 0\end{bmatrix},$$

$$C(t=3)=\begin{bmatrix}1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0\end{bmatrix}$$

or

$$C(t=1)=\begin{bmatrix}0.8 & 0.2\\ 0.5 & 0.5\end{bmatrix},\ C(t=2)=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.3 & 0.4\\ 0.3 & 0.3 & 0.4\end{bmatrix},$$

$$C(t=3)=\begin{bmatrix}0.2 & 0.2 & 0.3 & 0.3\\ 0.2 & 0.2 & 0.3 & 0.3\\ 0.2 & 0.2 & 0.3 & 0.3\\ 0.2 & 0.2 & 0.3 & 0.3\end{bmatrix}$$

At the same time, we also get the behavioral dynamics of each agent. For this event, the parameters of the modified F-J model, namely Equation (4), are not fixed values. Analyzing this data shows that the parameters of the model are actually a ϕ-mixing process, and parameter $A(k,j)\varsigma_{k-j}$ satisfies the following properties:

$$\phi_k=\sum_{j=0}^{\infty}A(k,j)\varsigma_{k-j}+\xi_k,\sum_{j=0}^{\infty}\sup_k\|A(k,j)\|<\infty$$

This process satisfies the following conditions.

(a) $\mathrm{E}[\varepsilon_k|\mathcal{F}_k]=0$; $\mathrm{E}[\Delta_{k+1}|\mathcal{F}_k]=\mathrm{E}[\Delta_{k+1}\varepsilon_k|\mathcal{F}_k]=0$
(b) $E[\varepsilon_k^2|\mathcal{F}_k]=\mathcal{R}_\varepsilon(k)$; $E[\Delta_k\Delta_k^\tau]=\mathcal{Q}(k)$
(c) $sup_kE[|\varepsilon_k|^r|\mathcal{F}_k]\leq M,\gamma\triangleq sup_k\|\Delta_k\|_r<\infty$.

Then, we simulated the subject that "Huang Haibo was framed" was true, as represented in Figure 10.

In Figure 10a–c, the red represents the people believing that Huang Haibo was framed, and the blue represents the people believing that "Huang Haibo visited sex workers" is real. Figures 10a to 10c show that, at the beginning,

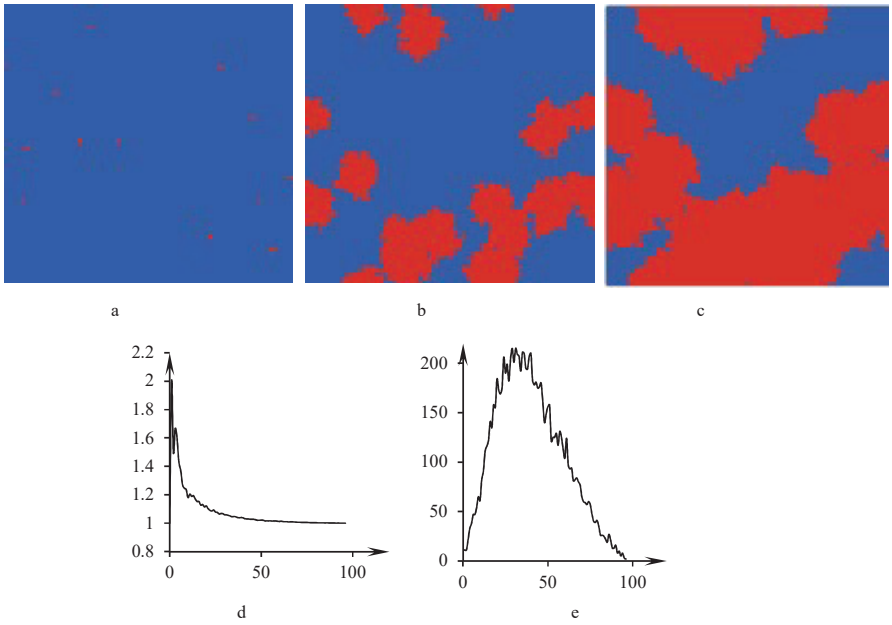

Figure 10 Transmission process of network big data in the Huang Haibo event.

most people believed that Huang Haibo visiting sex workers was true, but after a while, many people thought that he was framed. In a moment, we suddenly found that most people believed that Huang Haibo was framed. The main reason for this sudden change is that the people believing that Huang Haibo had been framed formed a connected giant component. Once this component is formed, the agents in the whole Internet will accelerate the recognition of this viewpoint. To describe this problem more clearly, Figure 10d introduces an order parameter $(x(t+1)-x(t))/x(t+1)$ describing the marginal transmission capacity. This parameter shows that the propagation acceleration of this network's big data increases from slow to fast and then to slow. Moreover, there is a characteristic of critical phase transition in this transmission process, which is the sudden formation of connected giant components in data transmission. At this critical point, the order of the corresponding largest connected giant component is $n\exp(p^2-1)$, where n is the total number of agents participating in the discussion at that moment, and p is the transmission probability of viewpoint. Figure 10e depicts how the success rate of this transmission kind changes over time. The value of this new order parameter is to explain the change in the number of people

believing that Huang Haibo visiting a sex worker was true. The number of people who spread this viewpoint to their neighbors shows a trend of first increasing and then decreasing with the increase of transmission. In addition, the success rate of transmission first increases sharply, reaches the maximum value, and then slowly decreases to a stable value. Based on this fluctuation, at first, most people thought that Huang Haibo was guilty, but later, more and more people believed that this matter was false. The main reason is that those agents believing that Huang Haibo was framed are connected with each other, forming a cycle called a critical state. Once this cycle occurs, the public accepts this viewpoint.

Because of the emergence of giant components, the critical data are determined according to Theorem 2, and the key giant components and their corresponding data are quickly found. These giant components actually describe two completely opposite viewpoints, true and false. Formula (6) determines which viewpoint is true via the following procedure.

(1) Determining the strength of these data.
(2) Sorting them according to their strength in descending order.
(3) Determining the value of the critical probability c_0 according to Theorem 2.
(4) Finding the top $c_0 n$ big data.
(5) Determining the giant components generated by these data in the way of traceability according to the evolution process of big data.
(6) Defining two events A and B, (i.e., A means "Huang Haibo visited a sex worker is true", B means the viewpoint that "Huang Haibo visited a sex worker is fake" described by the data in the giant components).
(7) The data of "Huang Haibo visited a sex worker is fake" appeared later. According to the process from its appearance, forming closed loops, forming connected giant components, to the ubiquitous existence, the entire network big data evolution process was divided into discrete processes to determine the probability of the connected giant components that represent "Huang Haibo visited a sex worker is true" according to Formula (6).
(8) Judging the authenticity of the event "Huang Haibo visited a sex worker" according to this probability.
(9) Selecting data from step (4) that is consistent with the result of step (8), and judging that these data are real and valid.
(10) Selecting the corresponding data and conducting subsequent analysis according to the purpose of the study.

The conclusion obtained in step (8) is that Huang Haibo was wronged, consistent with the actual facts afterward. Once the Bayesian inference obtains the connected giant components corresponding to the valid big data, the corresponding data will be selected from the previous $c_0 n$ key data for analysis and various studies. Based on the different research purposes, we choose appropriate data. If the problem of "Huang Haibo was framed" is analyzed, its data volume is less than $c_0 n$. Even so, there are about 53,000 data available, and the sample size is entirely sufficient to reflect the properties of the event.

5 Conclusion

In essence, network big data represents the embodiment and dissemination process of participants' thoughts and opinions. For network big data analysis, it is necessary to obtain a sufficient amount of big data to analyze an event in a statistical sense. Nevertheless, big data with poor quality create wrong conclusions. Therefore, statistical analysis is helpful if and only if all of these big data are valid.

The network big data forms a dynamic complex random network during the formation and dissemination. The network big data generated by each agent will be transmitted according to the priority connection mechanism, forming a random complex network. Then, the revised random F-J model analyzes the evolution law of network big data, which satisfies the characteristics of random time-varying. Not only the parameters but also their structures are random, they can effectively describe the evolutionary dynamics of network big data. Also, scientists generally believe that the network is scale-free, i.e., various viewpoints are spread in a scale-free network, leading to the emergence of different types of big data in the network. Therefore, the spread of network big data is the formation and evolution process of connected giant components with different properties. Accordingly, we construct a corresponding simulation model based on percolation theory. The simulation results show that, due to the interactions among agents, the viewpoint constantly forms a connected giant component according to the principle of "birds of a feather flock together" in the evolution of network big data. If invalid data forms a connected giant component, invalid data rapidly spread to the whole network and annihilate valid big data, when the rank of the connected giant component reaches a certain degree. This effect causes the emergence of "false to true" results.

Based on the above analysis, suppose that network big data has formed several connected giant components with different viewpoints distributed in the social network platforms at a particular moment. After deleting some data, the network big data system continues to be connected, indicating that the remaining data can fully describe the network unstructured big data system corresponding to the event, and can reflect the nature of the event. If the network cannot remain connected, the remaining data cannot describe the nature of the event. There are two ways to delete: random and intentional attack. Random attack means that some data are randomly selected and retained according to a probability p, and the rest of the data is deleted. Intentional attack means that these unstructured big data are sorted from large to small according to their strength, and then the corresponding data are deleted according to probability c. The research results prove that network big data are always connected under random attack, but the system is disconnected with a critical probability $c_0 n$ under intentional attack. Finding valid big data with random attack is difficult, while finding them with intentional attack is easy. Hence, the key is to determine this critical probability. If some people in the network maliciously spread the invalid data, the critical probability will be minimal. This finding implies that only a little valid data in the network needs to be tampered with, which can turn the whole network big data into invalid data. By analyzing the evolution process of network big data, we find the critical point of its phase transition and explore the nature of the critical point. Then, we can find the vital big data through the rank distribution of the homogeneous data connected giant components on the critical point. In addition, the Bayesian inference compares the distribution with the distribution at a known time to analyze whether they belong to the same distribution, and judge the authenticity of key data and identify the valid network big data. Finally, this research represents an example of the above principle and process. The results show that the method of identifying valid network big data is feasible and has essential significance for the application of network big data.

When there are multiple related events with nonlinear interaction in big data at the same time, multiple events interact with each other, leading to more complex sub-selection of real big data, which will be the focus of our next research.

Data Availability

The data used to support the findings of this study are included within the article.

Funding Information

This work was supported by the National Natural Science Foundation of China [72174064, 12101021], and the Natural Science Foundation of Shandong Province [ZR2020MG004].

Conflict of Interest

We all declare that we have no conflict of interest in this paper.

References

[1] A. Sahaym, et al., "Mixed blessings: how top management team heterogeneity and governance structure influence the use of corporate venture capital by post-IPO firms", *Journal of Business Research*, vol. 69, no. 3, pp. 1208–1218, 2016.

[2] A. N. Tump, T. J. Pleskac, R. H. J. M. Kurvers, "Wise or mad crowds? The cognitive mechanisms underlying information cascades", *Science Advances*, vol. 6, no. 29, pp. 1–11, 2020.

[3] B. Bollobás, O. Riordan, "Robustness and vulnerability of scale-free random graphs", *Internet Mathematics*, vol. 1, no. 1, pp. 1–35, 2007.

[4] B. Bollobás, O. Riordan, "Sparse graphs: metrics and random models", *Random Structures & Algorithms*, vol. 39, no. 1, pp. 1–38, 2011.

[5] B. Bollobás, et al., *Handbook of Large-Scale Random Networks*. Springer, 2008.

[6] A. S. Boccaletti, et al., "Complex networks: Structure and dynamics", *Physics Reports*, vol. 424, no. 4–5, pp. 175–308, 2006.

[7] D. Hume, "Cause and Effect, Indianapolis, 1772", *An Enquiry Concerning Human Understanding*. Hackett Publishing Company, 1993.

[8] M. H. DeGroot, "Reaching a Consensus," *Journal of the American Statal Association*, vol. 69, no. 345, pp. 118–121, 1974.

[9] T. Dixon, "Emotion: The history of a keyword in crisis", *Emotion Review*, vol. 4, no. 4, pp. 338–344, 2012.

[10] D. S. Callway, et al., "Network robustness and fragility: Percolation on random graphs", *Physical Review Letters*, vol. 85, no. 25, pp. 5468–5471, 2000.

[11] N. E. Friedkin, et al., "A theory of the evolution of social power: Natural trajectories of interpersonal influence systems along issue sequences", *Sociological Science*, vol. 3, no. 20, pp. 444—472, 2016.

[12] N. E. Friedkin, et al., "Network science on belief system dynamics under logic constraints", *Science*, vol. 354, no. 6310, pp. 321–326, 2016.
[13] L. Guo, J. Ljung, G. J. Wang, "Necessary and sufficient conditions for stability of LMS", *IEEE Transactions on Automatic Control*, vol. 42, no. 6, pp. 761–770,1997.
[14] I. Kozic, "Role of symmetry in irrational choice", *Preprint*, arXiv:1806.02627, 2020. https://arxiv.org/abs/1806.02627.
[15] J. Q. Cao, "Application of overlapping community discovery algorithm in complex network big data", *Automatic Control and Computer Sciences*, vol. 55, pp. 8–15, 2021.
[16] K. Hu, T. Hu, Y. Tang, "Cascade defense via control of the fluxes in complex networks", *Journal of Statistical Physics*, vol. 141, pp. 555–565, 2010.
[17] D. Lazer, et al., "The parable of Google Flu: traps in big data analysis", *Science*, vol. 343, pp. 1203–1205, 2010.
[18] I. Leyva, "Entrainment competition in complex networks", *International Journal of Bifurcation and Chaos*, vol. 20, no. 3, pp. 827–833, 2010.
[19] H. X. Lv, X. J. Zheng, S. Chen, "Reveal the pattern of causality in processes of urbanization and economic growth. an evidence from China", *Scientific Programming*, vol. 2725113, pp. 1–17, 2022.
[20] S. A. Masato, "Functional advantages of Lévy walks emerging near a critical point", *PNAS*, vol. 117, no. 39, pp. 24336–24344, 2020.
[21] S. E. Parsegov, "Novel multidimensional models of opinion dynamics in social networks", *IEEE Transactions on Automatic Control*, vol. 62, no. 5, pp. 2270–2285, 2017.
[22] P. Ball, *Critical Mass – How One Thing Leads to Another*. New York: Farrar, Straus and Giroux, 2007.
[23] A. M. Smith, et al., "Competitive percolation strategies for network recovery", *Scientific Reports*, vol. 9, 11843, 2019.
[24] L. Tai, L. Li, J. Du, "Multimedia based intelligent network big data optimization model", *Multimedia Tools amd Applications*, vol. 78, no. 4, pp. 4579–4603, 2019.
[25] T. Jäger, A. Passeggi, "On Torus Homeomorphisms Semiconjugate to irrational Rotations", *Ergodic Theory and Dynamical Systems*, vol. 35, no. 7, pp. 2114–2137, 2015.
[26] V. Marx, "The big challenges of big data", *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

[27] X. J. Zheng, et al., "Random belief system dynamics in complex networks under time-varying logic constraints", *Physica A: Statistical Mechanics and its Applications*, vol. 566:125552, 2020.
[28] X. J. Zheng, J. J. Zheng, "Criticality of symmetry breaking of collective behavior driven by individual rules", *System Engineering Theory and Practice*, vol. 36, no. 2, pp. 413–426, 2016.

Biographies

Peng Wang has a Ph.D. in management, and graduated from Korea Woosong University in 2018. She is a lecturer for the School of Economic and Management in Weifang University of Science and Technology. She has been engaged in research work on unstructural data modeling in management system, estimation and controlling for 5 years. She has been involved in several corresponding scientific projects. She is good at modeling and analyzing management complex systems from a complex system perspective. In recent years, she has begun to identify risk in a management complex system by multi-agent modeling, and identification, estimation and control.

Huaxia Lv is a master of region economics, graduated from Henan University in 2007. She is a lecturer for the School of Economic and Management in

Weifang University of Science and Technology. She has engaged in research work on spatial patterns of regional economy and its evolution for almost 10 years, and has published approximate 10 papers in different journals. In recent years, she has begun to pay attention to the research field of individual behavior and collective behavior and has developed a strong interest, obtaining some achievements. She will continue her research work in this area. She is good at data analysis and application.

Xiaojing Zheng has a Ph.D. in management science and engineering, and graduated from Wuhan University in 2012. He is a Professor of the School of Economic and Management in Weifang University of Science and Technology. He has engaged in research work on complex adaptive systems of management and supply chain coordination and risk analysis for almost 15 years, and has published approximate 50 papers in different journals. He is good at constructing various models for management complex adaptive systems by multi-agent modeling. In recent years, he has focused on the emergence of collective behavior driven by irrational individual behavior, consisting of invariable distribution, self-similarity, criticality and percolation of the collective; the corresponding research results has been published in several important scientific journals, which have gained the recognition of other scientists.

Wenhui Ma is a master of supply chain management, and graduated from Harbin University of Commerce in 2009. She is a lecturer in the School of Economic and Management in Weifang University of Science and Technology. She has engaged in research work on supply chain coordination based on unstructural data for 5 years. She has been involved in several corresponding scientific projects. She is good at unstructural data modeling in a management system, supply chain modeling, coordination and controlling from supply chain coordination. In recent years, she has begun unstructural data modeling in a management system, supply chain modeling, coordination and controlling.

Weijin Wang is an M.D. of Economics, and graduated from Shandong University in 2007. He is an associate professor of the School of International Trade in Shandong College of Economics and Business. He has engaged in research work on international trade and agriculture economy for 13 years, and has published several papers in journals and several books. He has headed up and been involved in several corresponding scientific projects. He is good at economic systems in qualitative analysis from a philosophical perspective. In recent years, he has been researching the phenomenon and the evolution law of macroecnomics by introducing unstructural data, and has achieved some modest conclusions.