Hindawi
Scientific Programming
Volume 2022, Article ID 7219080, 13 pages
https://doi.org/10.1155/2022/7219080

Research Article

Countermeasure of Telecom Network Fraud Investigation Based on Big Data

Tianyu Wang[1] **and Bo Yang**[2]

[1]*College of Criminal Justice, China University of Political Science and Law, Beijing 100088, China*
[2]*School of Sociology and Political Science, China University of Political Science and Law, Beijing 100088, China*

Correspondence should be addressed to Tianyu Wang; cu204025@cupl.edu.cn

Received 8 January 2022; Revised 23 March 2022; Accepted 1 April 2022; Published 26 May 2022

Academic Editor: Ahmed Farouk

With the diversification of information data in the information age of big data and the integration of network technology and the development of different industries, criminals who carry out telecommunication fraud are also using the technical loopholes existing in the process of integration of big data with different industries as an opportunity to commit crimes. This paper studies the investigation process and countermeasures of telecom network fraud through big data technology. This paper first introduces the characteristics of big data, analyzes the challenge of personal information security under the background of big data, warns people to protect their personal information in the era of big data, puts forward the clustering algorithm based on big data, introduces the concrete steps based on big data clustering algorithm, and then puts forward the specific steps of big data clustering algorithm. The current situation of telecom network fraud is analyzed, and the telecommunication network fraud is clustered based on big data. The experimental results show that, based on the clustering analysis of telecommunication network fraud based on big data, it is found that through the information age of big data, as long as big data are used rationally, it can effectively suppress telecommunications fraud and reduce it by 80%.

1. Introduction

With the advent of the "Internet+" era, the integration of information network technology and traditional industries is deepening, as well as the effective use of information network space and information data resources is profoundly changing the way of human existence. The intersection and integration of network and traditional industries, on the one hand, promotes the overall promotion of society in various fields; this is beneficial for the development of society, and can promote common progress in all fields of society and develop competition. In addition, due to the infringement of computer technology and the diversity and dispersion in the development of big data analysis, the emerging telecommunication fraud crime cases have shown a growing trend; its modus operandi is diverse and complex, seriously threatening the people's sense of wealth and security. Therefore, it is necessary for us to study the basic legal characteristics and investigation measures of telecom fraud crimes under the background of the Internet and the times, and to carry out precision strikes by strengthening the joint, focusing on the police force, forming a joint force, and looking for weaknesses, starting from the change of reconnaissance concept and the promotion of reconnaissance skills. The increasing development of Internet globalization has made the problem of personal information leakage more and more prominent. Protecting personal information is an important basis for safeguarding property rights. On the one hand, we should strengthen the performance of the legal obligations of Internet information service providers to protect information security; on the other hand, we should try our best to protect personal information.

Telecom fraud is a new high-intelligence, contactless crime with a rising crime rate in recent years; its organizational structure is tight and gradually developed into a corporatized business management model, and the modus

operandi has reached intelligence and diversification by using high-tech means. Telecommunications fraud has the following characteristics: outstanding noncontact; wide scope of criminal violations; various tricks and quick renovations; the use of electronic high-tech achievements to carry out remote crimes; industrialization development; enterprise operation; and other characteristics. The research on the problem of network fraud and preventive measures can play an effective preventive effect, which is of great practical significance to ensure the development of the network information industry, to ensure the steady growth of the national economy, to maintain the security of personal property, to ensure the social order of the network, and to promote the stability and development of the network society.

According to the research progress at home and abroad, different researchers have also studied the countermeasures of telecom network fraud investigation: In order to solve the information security problem in the process of large-scale data aggregation on the Internet, Zou et al. proposed a privacy protection algorithm (PPA) based on large-scale network data aggregation in view of the deficiencies of existing standard large-scale network data aggregation. Experiments have shown that this algorithm increases time utilization and has excellent reversibility and security in increasing false positive rates, making it more useful [1]. Sliwczynski et al. show the results of time transfer using optical fiber. The results show that operators of telecommunications networks can use this stable fiber link as a reliable source of synchronous signals with better accuracy than those using the most advanced GNSS time receivers [2]. Bouhamida et al. is designed to power remote telecommunications networks (RTNs) with appropriate photovoltaic-based energy generation systems by evaluating their performance and monitoring the associated Smart Microgrid (SMG) to provide safe and energy-efficient energy RTN management [3]. Cerroni et al. focuses on telecommunications software, network virtualization, and software-defined networks. Software and virtualization are increasingly important and transformative in today's telecommunications world, bringing the level of abstraction, decomposition, distribution, scalability, and programmability in network infrastructure and services to unprecedented levels [4]. Chen et al. have developed a fraud analysis and detection system based on real-time message communication, which constitutes one of the most common human-computer interaction services in online social networks, and proposes an integrated platform consisting of various text mining techniques, such as natural language processing, matrix processing, and content analysis through potential semantic models. Then, build an Android-based application to alert you to suspicious log and fraud events. The application is designed to facilitate the emergence of new self-configured integrated computing communication platforms to uninstall and process big data streams from mobile/wireless devices with limited resources in real time [5]. Baccarelli et al. outline the key challenges of managing real-time energy savings from distributed resources available in mobile devices and Internet-connected data centers [6]. The purpose of Mauro et al. is to identify and describe the most prominent areas of research related to "big data" and to propose a comprehensive definition of the term. Mauro et al. analyze a large number of industry and academia papers related to big data and discovers the commonalities between the topics they deal with, and give a new concept for the term, including that big data is an information asset characterized by such high capacity, rate, and complexity that specialized processing techniques and analytical methods are needed to translate it into real value [7]. Zheng et al. have developed ways to combine big data analytics with Internet optimization technologies to improve the quality of the user experience. First, mobile network optimization architecture for big data drivers (BDDs) is provided. Then, it introduces the characteristics of big data collected not only from ordinary user devices but also from the mobile Internet, and discusses some techniques in the process of big data acquisition and analysis from the perspective of network optimization [8]. However, these scholars do not combine big data to analyze the prevention of telecommunication network fraud countermeasures.

The innovation points of this paper are mainly reflected in: (1) introduces the characteristics of big data, and challenges the security of personal information in the context of big data, (2) puts forward the algorithm based on clustering analysis in the context of big data, and analyzes telecom network fraud by using clustering algorithm.

2. Methods on the Investigation of Telecommunication Network Fraud Based on Big Data

2.1. Features of Big Data. Big data analysis involves a large amount of data processing, a lot of kinds, so it is necessary to use the software system within the time limit prescribed by law to process the data of the corresponding data set, and analyze the basis for providing decision-making reference, in order to truly highlight the useful value of big data analysis [9]. It is generally believed that big data mainly has the following four typical characteristics: scale, diversity, high speed, and value. Big data have some characteristics:

(1) The amount of data is large and complex. The amount of information is huge, but the general database system does not have the ability to collect and store information; the kind of big data is rich, the data come from all aspects of society; the sources of information are complex, present in different structures and in different media forms, far beyond the management and analysis that can be accomplished by conventional database systems, which requires a database system with powerful functions to discover the potential value of big data, in order to achieve the effect of big data technology to promote economic development [10]. Big data include structured, semi-structured, and unstructured data, and unstructured data are increasingly becoming a major part of the data.

(2) The processing of data is fast. The demand for computer technology in the era of big data analysis is also getting higher and higher, and because of the

diversity of data analysis, the data analysis path of Figure 1 is particularly important for the timely processing of data analysis; people not only obtain data but also require data information classification, data mining, and also need to analyze the preferences and behavior models of information subjects, so that data information can be quickly and continuously classified and processed. This enables the important reference requirements for real-timeness to be met [11].

(3) Use big data analytics to get valuable economic information. The core of big data analysis is mathematical modeling, the foundation is the actual business, and the result is an automated procedure. The core of big data analysis is not to save or simply manage a large amount of data but to classify these data in a specific way and then obtain some key information. For example, after analyzing the shopping lists of a large number of consumers in a large supermarket, it was found that beer products often appeared on the same shopping list with diapers, and the large supermarkets quickly came to the conclusion that consumers who bought beer products tended to choose diapers more, so when the goods were placed, the beer products and diapers were placed together, which not only made it easier for consumers to pick up goods but also increased the price of these two products [12]. In these cases, supermarkets use the analysis of financial data to obtain important information of real value to their operations, and big data analysis technology is also the same, but its information is larger, and data processing, data analysis methods are more complex.

2.2. Challenges to the Security of Personal Information in the Context of Big Data. The increasing awareness of the protection of personal information of the public is also increasing with the vigorous development of the Internet. This is because with the advent of the information age, it is more and more obvious that many criminals use network technology to infringe on citizens' information [13]. But the background of big data analysis also brings convenience to the transmission and acquisition of bad information and the transmission and acquisition of information that has not been approved by the information subject, and adds great risks to the information security and protection of citizens. The exploration of big data and the development of data analysis technology foundation are changing rapidly; different institutions and individuals are also beginning to compete to seize the information resources, and the background of big data analysis also makes information more and more easy to become a "commodity." Driven by commercial interests will inevitably produce all kinds of enterprises to infringe information for profit phenomenon, so there is a huge security risk of information [14]. Figure 2 shows a few ways for information to be compromised. What are the ways of leaking personal information? Currently, the main ways of leaking personal information are to use Internet search engines to search for personal information, compile it into a book, and sell it to those who need to buy it at a certain price. The development of the information age also has advantages and disadvantages. The era of information sharing can make the society develop more rapidly, but at the same time, information is easily stolen.

In the context of large data, human society is more and more dependent on information; because of the rapid development of science and technology, economic and social development tend to humanize, and the important value of information is more prominent; in the market economy, there is a need for a market, and information will become a "commodity" because in the information age, whoever masters the advanced information technology first will be called the leader of this era. Therefore, under the conditions of a market economy, information is equivalent to commodities. In recent years, CCTV's annual 315th party will report on the time of information leakage caused by different reasons, initially for banks, China Telecom, and other large companies to leak user information, and then slowly developed into mobile phone applications, free WIFI leakage information [15]. Under the background of big data, the collection of personal information is mostly carried out by software, but with the wide application of smartphones, tablet computers, etc., the danger of big data to the security of personal information is even more perennial. However, due to the difference in knowledge level between information holders and fraudsters, in order to reduce the cost of crime, more crimes will occur, making personal information also in trouble.

2.3. Big Data Clustering Algorithms. Because the complex network system formed by computer network information has a large number of nodes and a large scale, and the network system information aggregation method has high time-consuming requirements, the research on the overall discovery algorithm is not suitable for the cluster analysis of network system information. We will start from the perspective of local whole research and provide an information clustering algorithm based on local whole key nodes [16].

In complex networks, the identification of major network nodes not only has practical theoretical research meaning, but also has great practical research value. This theoretical method can also be applied in various fields such as bioinformatics, social, and systems science. Starting with local critical nodes, the entire community structure can be discovered quickly and efficiently [17].

Community evolution tracking is a major part of biological evolution analysis. How do societies where needles appear at different times connect them (i.e., find the successors and successors of the society at a given moment) and how do you determine whether a society is new, dead, or separated or incorporated? These issues are crucial to social networking sites [18]. For societies that appeared at different times, they all have their own meanings. There are two ways of traditional association evolution analysis: point overlap,

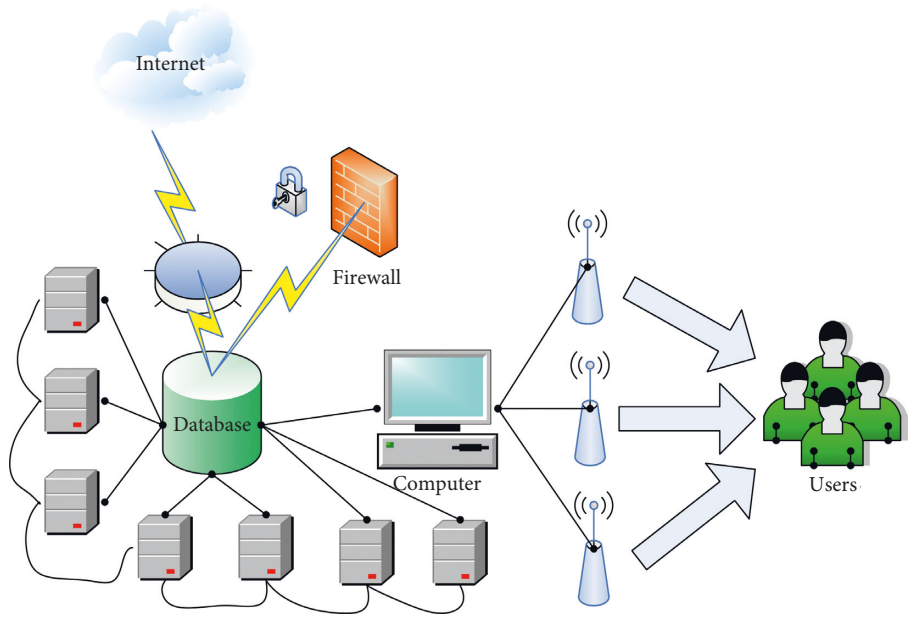

Figure 1: Data transmission path.

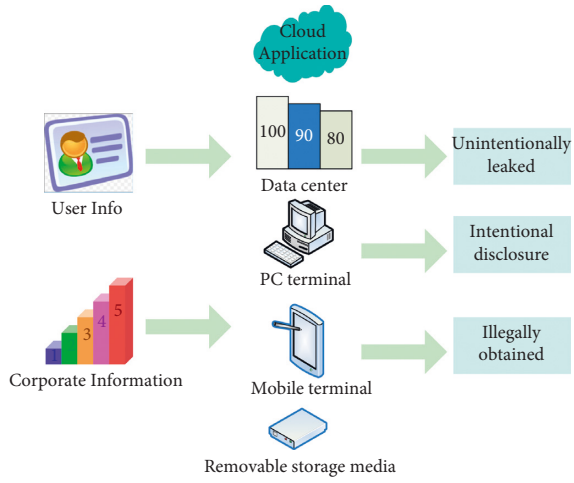

Figure 2: Information leakage path.

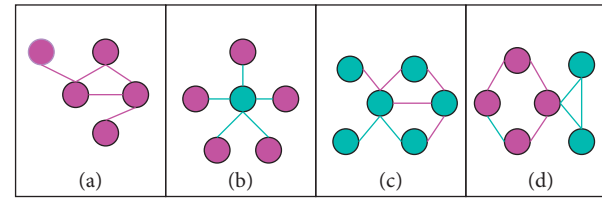

Figure 3: Correlation of structures and nodes.

refers to the degree of overlap between points at two moments before and after (generally using the Jaccard coefficient), when a threshold is exceeded, and there is a correlation between two associated points; the structure is similar (i.e. the topology of the edges), but both approaches have drawbacks. Figure 3 (a) and (b) although the node size is the same, their structure is completely different, (a) and (c) although the structure is exactly the same, its nodes have many differences, so it is difficult to say that their relationship is related [19]. Taken together, we feel that (a) and (d) are closer.

2.3.1. Improvements to the Fitness Function. Complex Internet community research methods based on adaptability, such as LFK algorithms, define adaptability functions as

$$j_Q = \frac{l_{\text{in}}^Q}{\left(l_{\text{in}}^Q + l_{\text{out}}^Q\right)^{\varphi}}. \tag{1}$$

At the same time, Q is a divided community that l_{in}^Q represents the sum of the number of interconnected edges between each node within social Q, i.e., twice the number of internal edges of the allocated social Q, wherein l_{out}^Q is the society Q between nodes connected to the number of edges outside the social Q, φ is a social control parameter, control adjusts the size of the social scale, but according to the experience, the φ value of 1.0–1.6 is more reasonable, the value selected in this article is 1.4. For any node φ D in the network, its adaptability to Community Q is defined as a change in Community Q join and without node D, the amount of change is

$$j_Q^D = j_{Q+D} - j_Q, \tag{2}$$

Among them, j_{Q+D} and j_Q, respectively, represent the adaptability of community $j_{D/Q}$ to node D, not the adaptability of node D to the original community Q. If $j_Q^D > 0$ means that the D node has increased adaptability after joining Community Q, and if so, the $j_Q^D < 0$ D node has joined Community Q the post-adaptation is reduced.

However, according to the adaptability $l_{\text{out}}^Q l_{\text{in}}^Q l_{\text{out}}^Q$ formula, determining whether a node can become a community requires calculations of the original community and the

community after joining the node, greatly increasing the computing time [20].

Make the following improvements to the adaptability formula.

Community Q by joining node D is available by (2):

$$j_{Q+D} = \frac{l_{\text{in}}^{Q+D}}{\left(l_{\text{in}}^{Q+D} + l_{\text{out}}^{Q+D}\right)^{\varphi}}. \tag{3}$$

When node D is joined, the edge of the original community Q connection node D becomes the inner edge, l_D^Q and the edge of the node connection node Q is also the inner edge l_{in}^D the edge of node D connected to the original community Q external node becomes the outer edge, l_{out}^D so you get

$$l_{\text{in}}^{Q+D} = l_{\text{in}}^Q + l_{\text{in}}^D + l_D^Q, \tag{4}$$

$$l_{\text{out}}^{Q+D} = l_{\text{out}}^Q + l_{\text{out}}^D - l_{\text{in}}^D. \tag{5}$$

You l_D^Q can see that the equivalent of is equal to l_{in}^D, and the equation (4) is

$$l_{\text{in}}^{Q+D} = l_{\text{in}}^Q + 2l_{\text{in}}^D. \tag{6}$$

The equivalent of the bring-in (5) and (6) bring-in (3) is

$$j_{Q+D} = \frac{l_{\text{in}}^{Q+D} + 2l_{\text{in}}^D}{\left(l_{\text{in}}^Q + l_{\text{in}}^D + l_{\text{out}}^Q + l_{\text{out}}^D\right)^{\varphi}}. \tag{7}$$

You only need to calculate the sum of the initial community Q once, $l_{\text{in}}^Q l_{\text{out}}^Q$ and each time you add a new node later, you only need to calculate and meet the requirements, $l_{\text{in}}^D l_{\text{out}}^D$ which greatly reduces the computational time consumption [21].

When nodes D, E, and F are added at the same time, you get

$$j_{Q+D+E+F} = \frac{l_{\text{in}}^Q + 2l_{\text{in}}^D + 2l_{\text{in}}^E + 2l_{\text{in}}^F + 2l_E^D + 2l_F^D + 2l_E^F}{\left(l_{\text{in}}^Q + l_{\text{in}}^D + l_E^D + l_D^F + l_{\text{in}}^E + l_E^D + l_E^F + l_{\text{in}}^F + l_D^F + l_E^F + l_{\text{out}}^Q + l_{\text{out}}^D + l_{\text{out}}^E + l_{\text{out}}^F - 2l_E^D - 2l_F^D - 2l_E^F\right)^{\varphi}}. \tag{8}$$

Among them, l_{in}^D is the edge connecting community Q to node D, l_{out}^D is the edge connecting node D to the edge, l_{in}^E is the edge connecting community Q and node E, l_{out}^E is the edge connecting node E to external nodes, l_{in}^F is the edge connecting community Q to node F, l_{out}^F is the edge connecting node F to external nodes, and the l_E^D is the edge connecting node D to node E. Is the edge connected to l_F^D F for D. Is an edge connected to l_E^F E by F. Think of D, E, and F as a community. Q_1 is available.

$$l_{\text{in}}^{Q_1} = l_E^D + l_D^E + l_F^D + l_D^F + l_E^F + l_F^E = 2l_E^D + 2l_F^E + 2l_E^F, \tag{9}$$

$$l_Q^{Q_1} = l_Q^D + l_Q^E + l_Q^F, \tag{10}$$

$$\begin{aligned} l_{\text{out}}^{Q_1} &= l_{\text{out}}^D + l_{\text{out}}^E + l_{\text{out}}^F - l_E^D - l_D^E - l_F^D - l_D^F - l_E^F - l_F^E \\ &= l_{\text{out}}^D + l_{\text{out}}^E + l_{\text{out}}^F - 2l_E^D - 2l_F^D - 2l_E^F, \end{aligned} \tag{11}$$

where *is the* $l_{\text{in}}^{Q_1}$ inner edge of Community Q_1, $l_Q^{Q_1}$ the edge connected to Q_1 for Community Q, $l_{\text{out}}^{Q_1}$ and Q_1 Edge connected to outside. Bring-in (9)–(11) bring-in (8) to

$$j_Q = \frac{l_{\text{in}}^Q + 2l_Q^{Q_1} + l_{\text{in}}^{Q_1}}{\left(l_{\text{in}}^Q + l_Q^{Q_1} + l_{\text{in}}^{Q_1} + l_{\text{out}}^Q + l_{\text{out}}^{Q_1}\right)^{\varphi}}. \tag{12}$$

Therefore, the adaptation of Community Q_1 to Community Q can be calculated by using type (12) [22].

2.3.2. Complexity Analysis. The space and time consumption of the algorithm is mainly in the community expansion section, taking the clustered data with size a and the number of categories v as examples. Space complexity is a measure of the amount of storage space temporarily occupied by an algorithm during its execution. Time complexity is a function that qualitatively describes the running time of the algorithm.

(1) Spatial Complexity. The expansion process of spatial structure of each society always follows the serial implementation method, so that the occupied storage space found between societies can be reused, so the total spatial structure complexity is b, of which b is the largest social reserve space structure, the total space cost of the allocation process is v, so the allocation of the last maximum social reserve space structure is $a - i$, where $a - i$ is the number of nodes repeated between societies, the total data storage space is a. So, the spatial complexity is

$$C(a) = b + v + 2a + i. \tag{13}$$

The spatial complexity available by formula (13) is linear $P(a)$.

(2) Time Complexity. Expansion within all communities scales out from within a key community, with data divided into $a - i_2$ and i_2 as nodes where communities overlap. Because a multithreaded strategy is used, assuming the number of threads is b, the time complexity is

$$R(a) = \frac{a + i_2}{b}. \tag{14}$$

In sparse networks, (14) can be equivalent to

$$R(a) = \frac{a}{b}. \tag{15}$$

Available by (15) Time Complexity is linear $P(a)$.

Therefore, the overall spatial complexity of the parallel algorithm is

$$C(a) = b_1 + v_1 + 4a + b + v + i. \tag{16}$$

The overall spatial complexity of the parallel algorithm available by (16) is linear $P(a)$.

Therefore, the overall spatial complexity can be

$$R(a) = \frac{a}{b} + i_1 + \frac{a + i_2}{b}. \tag{17}$$

In sparse networks (17) it can be equivalent to

$$R(a) = 2\frac{a}{b}. \tag{18}$$

Available by formula (18), the overall time complexity of the parallel algorithm is linear $P(a)$.

(3) Evaluation Function Based on Big Data. Nicosia et al. proposed functions to evaluate community structure in the evaluation of post-divided communities.

$$\begin{cases} Z_{ol} = \dfrac{1}{b}\displaystyle\sum_{f=1}^{a_f}\sum_{i,k}\left[J\left(\varphi_{i,f},\varphi_{k,f}\right)D_{ik} - \dfrac{\gamma^{\text{out}}_{i\longrightarrow,f}x^{\text{out}}_{i,f}\gamma^{\text{in}}_{k\leftarrow,f}x^{\text{in}}_{k,f}}{a}\right., \\ \gamma^{\text{out}}_{i\longrightarrow,f} = \dfrac{\sum_k J\left(\varphi_{i,f},\varphi_{k,f}\right)}{a}, \\ \gamma^{\text{in}}_{k\leftarrow,f} = \dfrac{\sum_k J\left(\varphi_{k,f},\varphi_{i,f}\right)}{a}, \end{cases}$$

$$J\left(\varphi_{i,f},\varphi_{k,f}\right) = \frac{\sum_k J\left(\varphi_{i,f},\varphi_{k,f}\right)}{\left(1+e^{-j\left(\varphi_{i,f}\right)}\right)\left(1+e^{-j\left(\varphi_{k,f}\right)}\right)},$$
$$j(m) = 2pm - p,\ p \in R, \tag{19}$$

Among them, D is the adjacency matrix of the network. When there are $x^{\text{out}}_{i,f}$ network nodes i, the progress of $x^{\text{in}}_{k,f}$ network nodes k is $\varphi_{i,f}$ and $\varphi_{i,f}$ represents i-to-f, that is, i is the membership factor of f, which is represented by $i\epsilon f$. $\sum_{f=1}^{af}\varphi_{i,f} = 1$ Nicosia et al. defined a more reasonable function through experimental testing, $J(\varphi_{i,f},\varphi_{k,f})$ to calculate, gave the empirical value of p.30. And, in research papers by Nicosia et al., it has been shown that higher values represent overlapping community structures with a higher degree of modularity.

(4) Specific Steps Based on Big Data Clustering Algorithms. The first step is to randomly select network node D.

The second step is to get the neighbor node for D and calculate the degree of D and D's neighbor node.

In the third step, select the node with the greatest degree, and if D is the node with the greatest degree, then D is the local critical node, otherwise repeat the second step with the node with the highest degree as the initial node, until the initial node is the node with the highest degree [23].

The fourth step is to obtain the local important community by obtaining the large cluster where the local important nodes are located through the parallel strategy discovered by the large cluster.

Step 5, repeat Step 1 until you have acquired the existing local critical community, and include all the data including the normal nodes and the local area critical section community consisting of a large cluster [24].

Step 6, select the largest local key community as the initial community Q1, and expand according to the adaptation formula. If it encounters all points in other local key communities Q2 plus all points in Q1, it observes how their adaptation changes. If the fitness becomes larger, add Q2 to Q1, if the fitness decreases, Q2 will not be added to Q1 until no neighbor nodes belong to Q1.

Step 7, select the larger locally important neighborhoods in the remaining locally important neighborhoods as the starting block, and then repeat step six until you have traversed all the area points [25]. The process is shown in Figure 4.

3. Experimental Results of Telecommunication Network Fraud Investigation Strategy Based on Big Data Analysis

At present, the research on the crime of telecommunication network fraud is based on the most achievements of public security investigation and judicial investigation, and the research focuses mainly on the detection bottleneck and measures of telecommunication network fraud crime, the current situation and the results obtained, the regulation system, the analysis of classic cases, and so on. This article is based on the era of big data Internet, investigating telecommunications network fraud, and gives the corresponding countermeasures; it is expected to make a certain contribution to network security.

3.1. Current Situation of Telecommunications Network Fraud. After research, it is concluded that fraud suspects most often pass through the five types of communication channels to transmit illegal information. As can be seen from Table 1, fraudsters transmit illegal information containing fraudulent content through the use of communication tools or services provided by Internet companies. The most commonly used means of information dissemination for fraud suspects is that 24.01% of fraud information is disseminated using QQ; use classified information websites for dissemination; 5.21% use WeChat for dissemination. QQ and WeChat are both instant messaging services provided by Tencent, and the proportion of services provided is 29.32%.

It can be found that the most commonly impersonated specific groups of suspected fraudsters are, in turn, human resources personnel, public prosecution law, financial practitioners, the closest family members, and shopping network customer service. Other groups that wantonly counterfeit often deceive victims because of the very low cost of identity counterfeiting on the Internet, which makes it impossible to easily identify authenticity. Although most

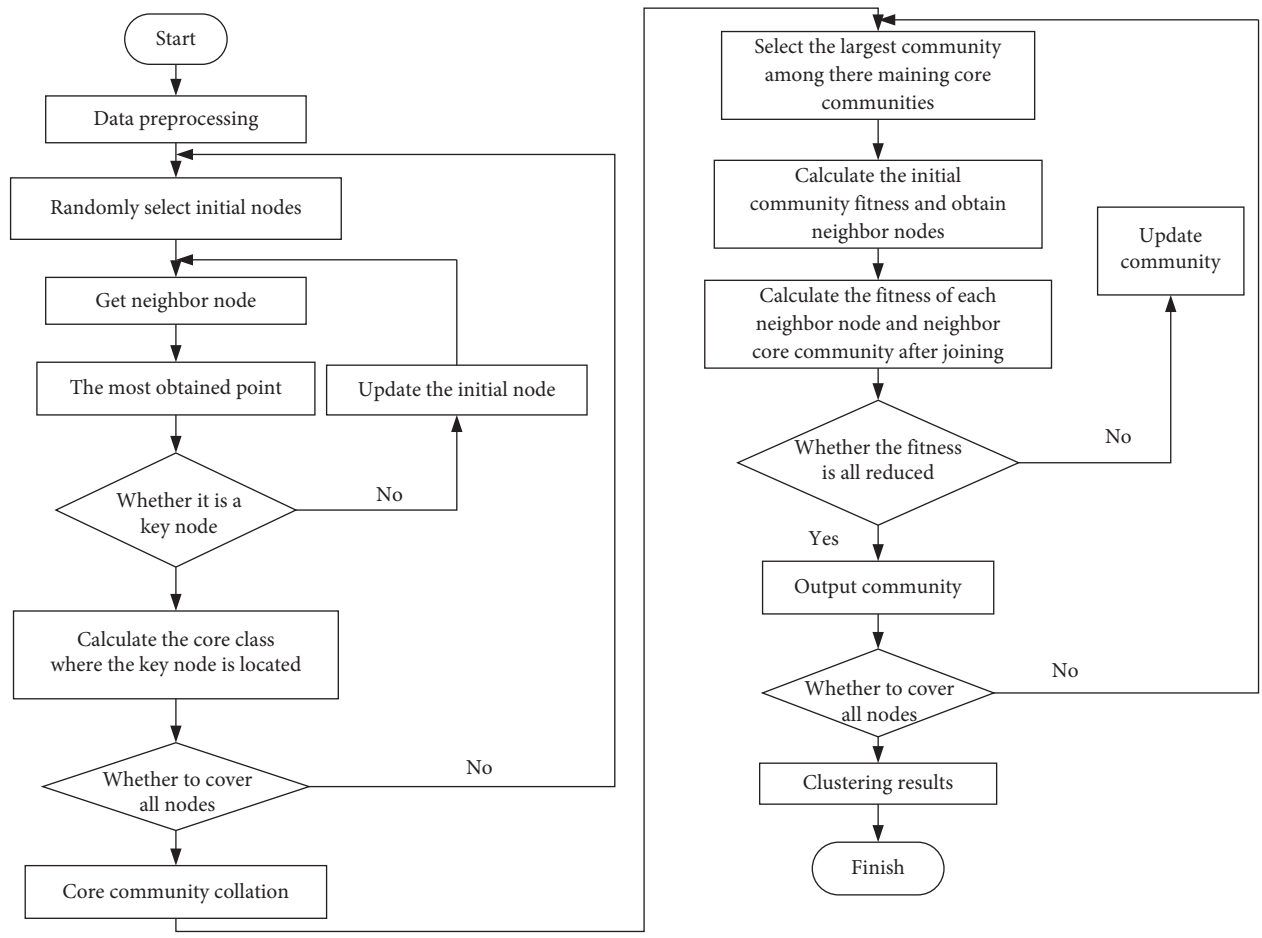

Figure 4: Flow chart of big data clustering algorithm for key communities.

Table 1: Statistics of communication channels and the most frequently impersonated people.

Channel	Count	Percentage	Impersonated group	Count	Percentage
Telephone	469	40.12	Impersonating human resources	219	17.99
QQ	291	24.01	Impersonation of public security law	171	14.01
Classified information website	147	12.49	Impersonating a financial practitioner	156	12.39
SMS	153	11.98	Impersonating someone close	139	11.98
WeChat	59	5.21	Pretend to be a customer service on a shopping site	112	8.89

people can identify the disguise of a fraud suspect, if the fraud suspect keeps sending messages in exchange for trust so as to pass false news to many people; this can successfully deceive some people.

By analyzing the data of the victims of the case filed by the public security organs, it can be found that most of the masses are more vulnerable to the impact of telecommunication network fraud, such as statistics on the sex and age of the victims, as shown in Figure 5.

It can be found from the figure that men aged 21 to 25 and women aged 26 to 30 are the most frequently victims of telecommunications network fraud.

At the same time, the research on banking and telecommunications network fraud also found that due to the manual inspection to complete the risk audit work, the workload and intensity of the bank office supervisors have increased. In recent years, the warning and write-off messages with high operational risk have gradually increased the trend. Since 2016, the main business data information generated is shown in Figure 6.

At present, the business operation risk problem extracted from the commercial bank's massive operation business still depends on manual, random, and sampling methods. In 2019, for example, of the 171.258 million underlying business data generated, only 362,000 were manually sampled and 2310 pieces of risk information were detected. It is not possible to strengthen risk prevention and management through data analysis, and there are many difficulties in the process of risk prevention. Since 2019, four types of operational risk write-off information have been randomly listed as shown in Table 2:

Since 2019, the number of write-offs per month has been irregular, such as the customer information governance check conducted in April 2019, when 107 write-off issues

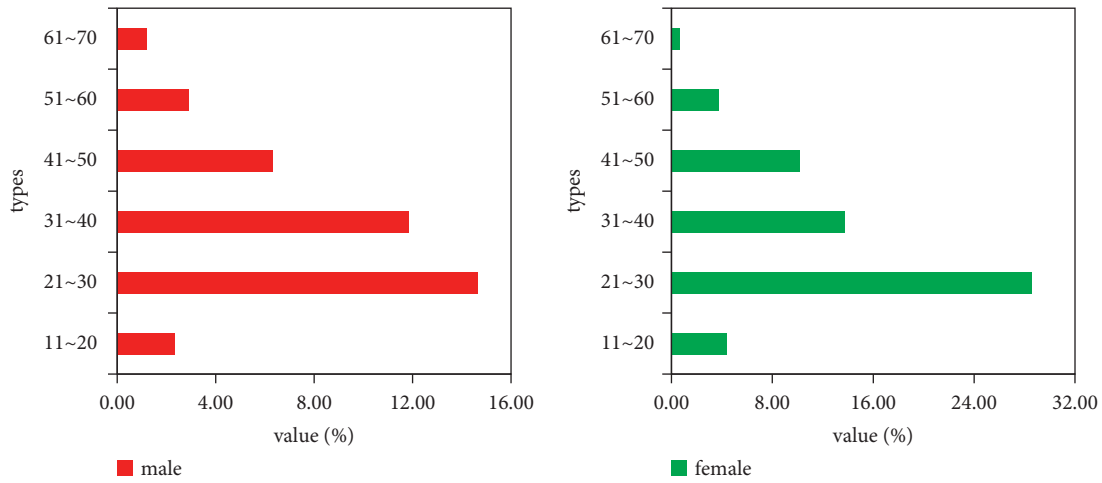

Figure 5: Age-sex ratio of victims.

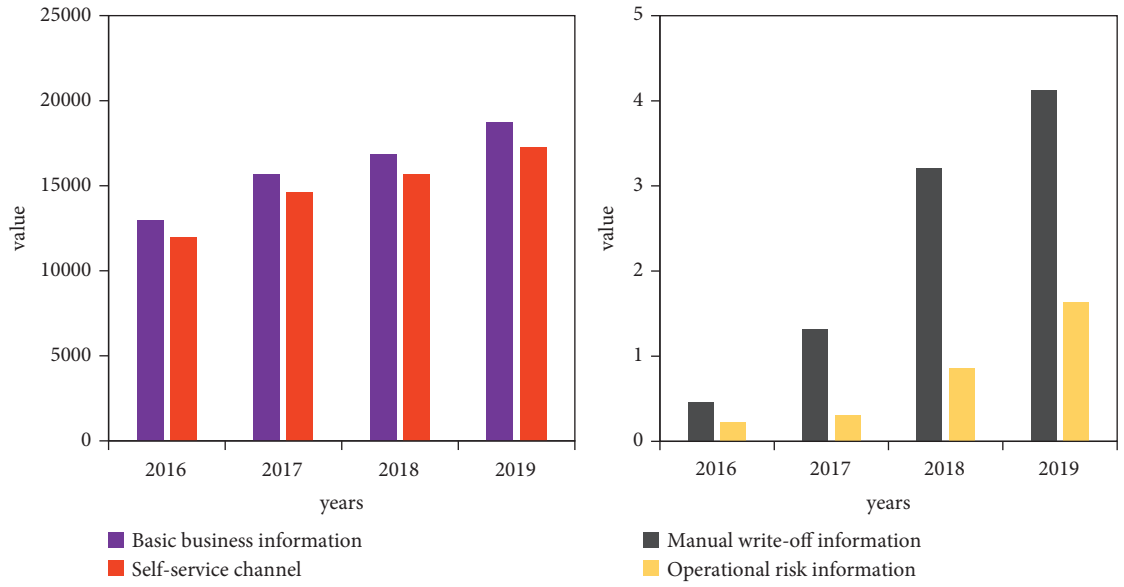

Figure 6: Banking business information in the past four years.

Table 2: Bank operational risk information table.

month	Write-off information	Operational risk information			
		Customer information protection	Real-name account system	Self-service equipment	Online merchants
January	2801	9	16	2	4
February	3298	16	9	5	2
March	3286	41	3	15	6
April	2995	111	10	10	8
May	3001	36	15	129	11
June	2869	9	47	18	5
July	3098	5	129	15	3
August	3168	21	31	10	8
September	3203	19	21	6	5

were identified. Not proportional to other monthly check write-off issues, 131 questions on the real-name system of accounts were verified during the account information rectification work carried out in July 2019; self-service equipment inspections were conducted throughout the city in 2017, with an average of no more than 10 write-offs per month, the number of write-offs reached 130 in May, while online merchants carried out special inspections in June

because of a single operating process There has been no increase in write-off issues.

In view of the characteristics of investment risk of banking business data, through clustering analysis technology, the company's operating data are divided into investment market operation risk, business operation risk, and credit risk, and then used to extract the control risk type data in a full sample of business data; the specific analysis process is as follows: Cluster analysis is an exploratory analysis method. Different from discriminant analysis, cluster analysis does not know the classification standards in advance, or even how many categories should be divided into, but will automatically classify according to the characteristics of the sample data.

(1) The initial cluster center is specified. The objects of the business class are used as clustering centers and are recorded as c1, c2, c3 in three categories.

(2) Clustering. For all sample ai of the data set, calculate the largest gap between it and the first three clusters. The main approach is to measure risk characteristics by the type of business that the data contains, and to attribute the risk types closest to the gap to the same class of values.

(3) Update the cluster center. When calculating the sample average, if it is not significantly reflected in the calculation results, re-cluster to arrive at the sample average.

(4) Judgment. If the center of the calculation to each class no longer changes, the iteration ends with a clustering result.

(5) Export the clustering diagram, as shown in Figure 7.

By using clustering analysis, all samples of commercial bank operating financial data can be clustered to obtain the correct risk assessment and analysis information, and provide data support for the next stage of research and analysis management.

3.2. Clustered Analysis of Telecommunications Network Fraud Based on Big Data. From the experimental results, we can see, as shown in Figure 8, that in the karate data set test, the Q values obtained by the LFK algorithm, the big data clustering algorithm based on local key nodes and the big data clustering algorithm based on local key communities are 0.661 and 0.692 and 0.689, the information clustering algorithm based on the important network nodes of the local region and the information clustering algorithm based on the important community of the local region are higher than the Q value obtained by the LFK algorithm. The time spent is 0.449 s, 0.319 s, and 0.371 s, respectively, and the difference between the first three can be found to be small in time frame. In the process of football data collection testing, the Q values obtained by the three algorithms are 0.631 and 0.701 and 0.698; however, it can also be found that the information clustering algorithm based on local key nodes is more modular than the clustering results obtained according to the information clustering algorithm of local important

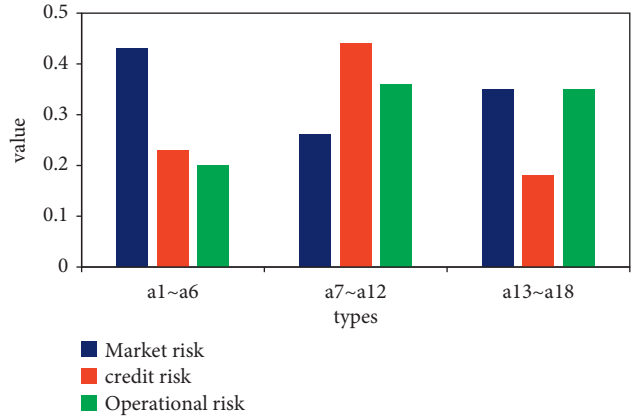

Figure 7: Cluster analysis diagram divided by risk type.

communities, and the average time limit used by the three is 8.364 s, 7.214 s, and 7.954 s, and the difference between the three average time is much worse than during the karate data set testing process, but still not obvious.

As can be seen from Figure 9, for smaller real network karate and football, the calculations we give are not very different from the original computational time, while the Q values are higher based on the information clustering algorithm of locally important nodes and the information clustering algorithm based on locally important communities. In the CA-HepPH data set test, the Q values obtained by the three algorithms were 0.561 and 0.701 and 0.709, using 639.185 s, 467.832 s, and 502.636 s. It can be found that the large data clustering algorithm using local whole important nodes is more modular than the large data clustering algorithm using local whole important community, and the reduction effect of time consumption is more significant. In the Enron data set, the Q values obtained by the three algorithms are 0.291, 0.509, and 0.569, using 863.205 s, 698.386 s, and 789.306 s, so we can find that the large data clustering algorithm that originally used local area important nodes was more modular than the clustering algorithm obtained by using the local whole important community big data clustering algorithm, and the average time consumption was more significant. In addition, it can be found in the conclusion that for smaller information aggregation, the use of local area important network nodes of big data clustering algorithm and the use of local important society of big data clustering algorithm is not very different, and because of the small data information gathering of important social area is relatively small, the time consumption found in important society is relatively small, occupied by a larger proportion of resources.

The Q values are not much different, as are the smaller data sets and the resulting clustering results are fixed, so the Q values are also relatively fixed. In larger data sets, the time-consuming gap between the two is much larger than in smaller data sets, but it is also within the acceptable range. This is because there are more communities divided in larger data sets and more key communities within the community, so there is more time spent on finding key communities. The Q values of the two are also different, as there are more critical communities per community in larger data sets, so

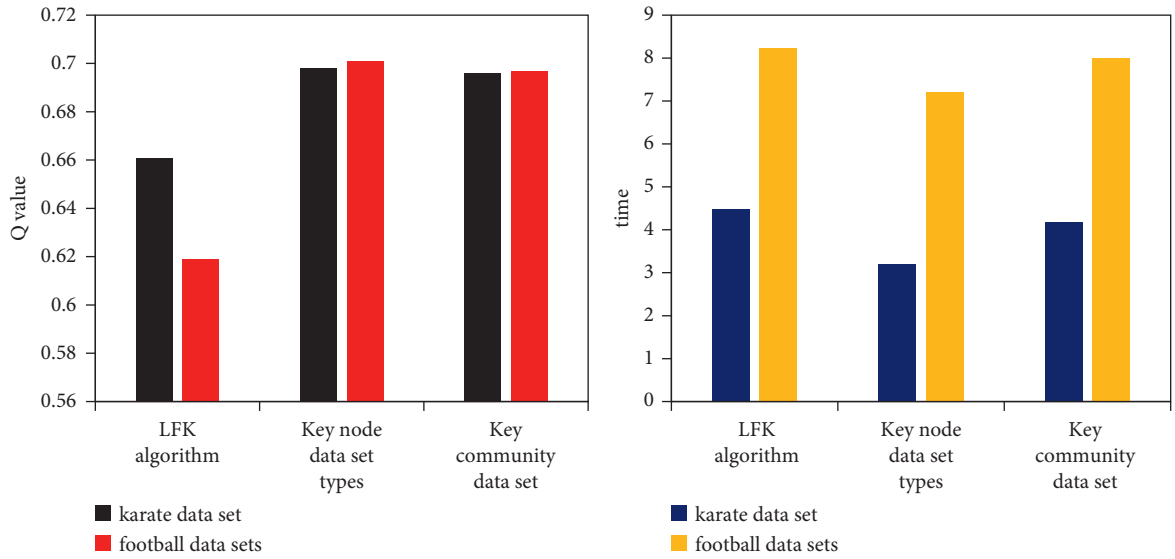

Figure 8: Test results of the three algorithms in the karate and football data sets.

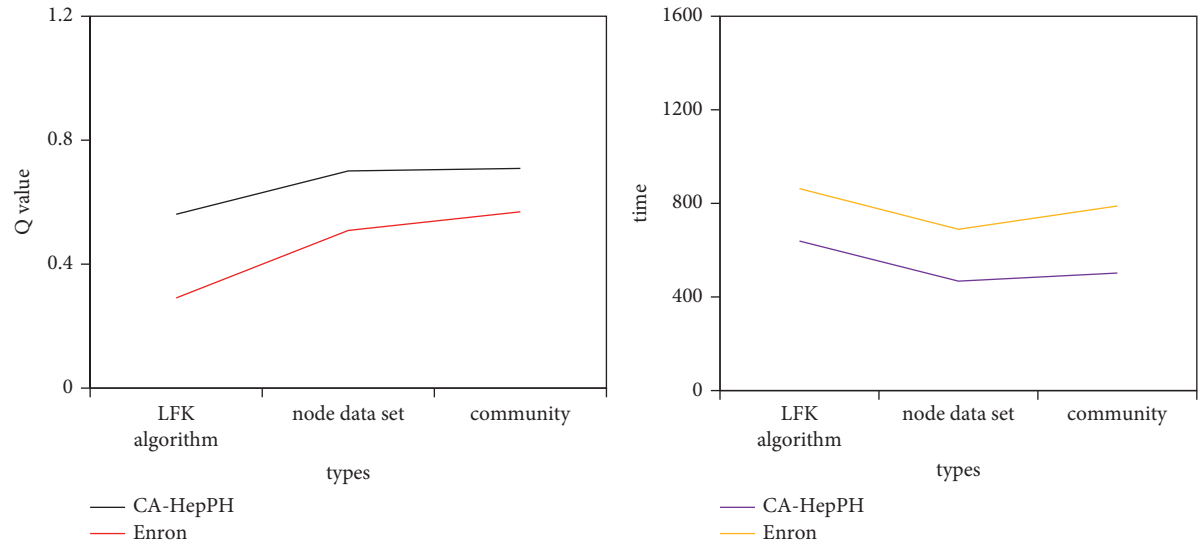

Figure 9: The test results of the three algorithms in the Enron and CA-HepPH data sets.

big data clustering algorithms based on local critical nodes may exclude key communities within some communities, resulting in a decrease in clustering results. Thus, it can be found that for large areas of real network data sets, through the information clustering algorithm based on local important nodes, and through the information clustering algorithm based on local important communities, the results of clustering analysis can enhance the clustering effect and effectively reduce time consumption. At the same time, the time consumrd to adopt the information clustering algorithm of the important nodes of the local whole region is relatively small, while the information clustering algorithm of the important community of the whole local area is higher.

As you can see from Table 3, in the karate data set test, the Q values of single-threaded, two-threaded parallel, and three-thread parallel are 0.692, 0.689 and 0.679, because the data set is small and the classification is fixed, single-thread, two-thread parallel, and three-thread parallel clustering have the same effect. The time spent of 0.398 s, 0.387 s, and 0.491 s shows that the time consumption of the three is not much different. In the football data set test, the Q values of single-threaded, two-threaded parallel, and three-threaded were 0.692 and 0.701, and 0.716, respectively, and you can also see that the clustering effect is not much different, the time used for the three is 8.002 s, 7.501 s, and 6.098 s, the time difference between the three is larger than in the karate data set test, but it is still not obvious.

As you can see from Table 4, in smaller real networks, the parallel time consumption of single-threaded, parallel with two threads, and three-threaded is not much different, and the Q values obtained by the three are not much different. In the CA-HepPH data set test, single-threaded, two-threaded parallel, and three-thread parallel Q values were 0.709, 0.753,

Table 3: Test results of parallel strategies in karate and football data sets.

	Karate data set			Football data set		
	Single thread	Two threads in parallel	Three threads in parallel	Single thread	Two threads in parallel	Three threads in parallel
Q value	0.692	0.689	0.679	0.692	0.701	0.716
My T	0.398	0.387	0.491	8.002	7.501	6.098

Table 4: Test results of parallel strategies in CA-HepPH and Enron data sets.

	CA-HepPH data set			Enron data set		
	Single thread	Two threads in parallel	Three threads in parallel	Single thread	Two threads in parallel	Three threads in parallel
Q value	0.709	0.753	0.712	0.569	0.562	0.571
My T	235.561	125.673	101.623	987.382	839.212	789.322

and 0.712, respectively, using a time of 235.561 s, 125.673 s, and 101.623 s; from then on, you can see that the clustering quality of the three is not much different, but the time consumption of multithreaded parallel operations is significantly reduced. In enron data sets, single-threaded, two-thread parallel, and three-thread parallel operations result in Q values of 0.571, 0.559, and 0.567, respectively, using 987.382 s, 8.39.212 s and 789.322 s; the clustering quality of the three is not much different, but the reduction in multithreaded time consumption is more obvious and proportional to the number of threads.

So, Figure 10 is a parallel strategy run efficiency graph, as can be seen from the experiment. For smaller data information integration, the time-consuming difference between multiple threads and single threads is not significant. This is mainly because the data set is small, so the time consumption for clustering is less. Although multithreaded can reduce the loss of information clustering and also improve the time loss of hardware, multithreaded and single-threaded time loss and smaller data sets are not very different.

For larger data sets, however, the reduction in time consumption for multiple threads is even more significant, as the amount of time spent on clustering will account for the bulk of the total time consumed in larger data sets, and the overall time consumption is more significant than the negligible hardware loss caused by the following multithreads. For smaller and larger data sets, however, the Q values derived by double multithreads are not much different from those obtained by single threads, indicating that single multithreads do not degrade the quality of clustered results. Thus, it can be found that in multithreaded technology for large real network data set, it can reduce the time consumption without reducing the quality of clustering, while reducing the time consumption is proportional to the number of threads. Based on the clustering of telecommunication network fraud based on big data, it is found that through the information age of big data, the rational use of big data can effectively curb telecommunications fraud and reduce the occurrence of similar fraud incidents by 80%.

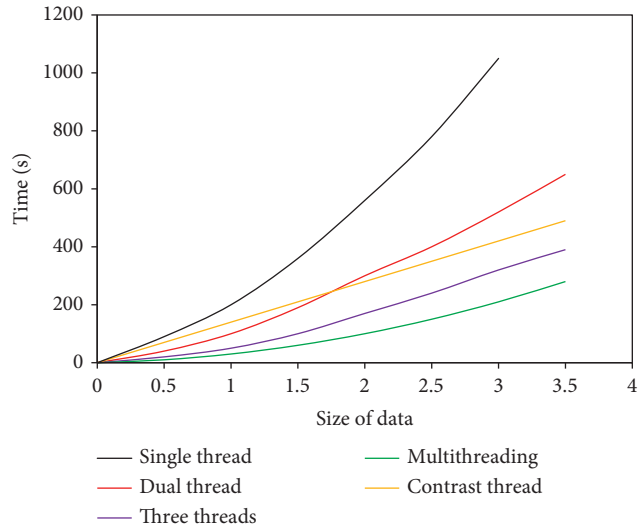

Figure 10: Operational efficiency graph of parallel strategy.

4. Discussion

With the rapid development of Internet society, the problem of network fraud is becoming more and more serious, and it has a great impact on individuals and families, which has seriously damaged the whole network society. At present, the general preventive measures of online fraud are not perfect, and the problem of online fraud is still serious, so it is particularly important to improve the prevention measures of online fraud. Under the background of big data analysis, the important value of information has also begun to be gradually highlighted, which has a positive impact on economic development, scientific and technological progress, and the construction of a harmonious society. If the civil law protection of personal information is still strictly protected, it cannot meet the requirements of the development of the times, nor can the social value of personal information be well found and used. Therefore, the government must adopt unified legislation to recognize people's legitimate use of personal information resources and policy guidance to avoid the risk of personal information rights being infringed, and thus prompt the government to use big data technology to use information resources legally, and thus create a convenient life and a more humane working environment. At the same time, we also need to pay attention to the improvement of the legal assistance channels for victims whose personal information rights and interests

have been violated, so as to balance the rational use of information and ensure the proper balance between the interests of both sides. In the case of information leakage, government departments should also take corresponding measures to reduce the leakage of public private information and protect the public from all aspects.

5. Conclusions

In order to achieve the rational use of information, it is necessary to build on the broad trust of human beings in the system, and implementing special legislation on information protection will help human beings realize that they enjoy the right to information and understand their own information, so that they can understand the rational use of information, and can adopt a more open mind to accept and use information brought about by the convenience. It is true that the development of law has the characteristics of lag, especially in today's network, and information light speed developed; no matter how sound the establishment of personal information protection law, it is also possible to completely deviate from the intended direction of development. In this way, information needs to adapt its connotation to the adjustment of the times according to the social development requirements of each era, in order to realize the good relationship between the rational use of information and civil law guarantee, and provide people with a good social atmosphere with a sense of security, so as to achieve a better future of information.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2018YFC0831002).

References

[1] Y. Zou, W. He, L. Zhang, J. Ni, and Q. Chen, "Research on privacy protection of large-scale network data aggregation process," *International Journal of Wireless Information Networks*, vol. 26, no. 3, pp. 193–200, 2019.

[2] L. Sliwczynski, P. Krehlik, J. Kolodziej et al., "Fiber-optic time transfer for UTC-traceable synchronization for telecom networks," *IEEE Communications Standards Magazine*, vol. 1, no. 1, pp. 66–73, 2017.

[3] H. A. Bouhamida, S. Ghouali, M. Feham, B. Merabet, and S. Motahhir, "PV energy generation and IoT power consumption for telecom networks in remote areas," *Technology and Economics of Smart Grids and Sustainable Energy*, vol. 6, no. 1, pp. 1–11, 2021.

[4] W. Cerroni, A. Galis, K. Shiomoto, and M. F. Zhani, "Telecom software, network virtualization, and software defined networks," *IEEE Communications Magazine*, vol. 57, no. 10, pp. 40-41, 2019.

[5] L.-C. Chen, C.-L. Hsu, N.-W. Lo, K.-H. Yeh, and P.-H. Lin, "Fraud analysis and detection for real-time messaging communications on social networks," *IEICE Transactions on Information and Systems*, vol. 100, no. 10, pp. 2267–2274, 2017.

[6] E. Baccarelli, N. Cordeschi, A. Mei, M. Panella, M. Shojafar, and J. Stefa, "Energy-efficient dynamic traffic offloading and reconfiguration of networked data centers for big data stream mobile computing: review, challenges, and a case study," *Computers & Chemical Engineering*, vol. 91, no. 2, pp. 182–194, 2016.

[7] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of big data based on its essential features," *Library Review*, vol. 65, no. 3, pp. 122–135, 2016.

[8] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5G," *IEEE Network*, vol. 30, no. 1, pp. 44–51, 2016.

[9] M. Zaharia, R. S. Xin, P. Wendell et al., "Apache Spark: a unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[10] E. Zeydan, E. Bastug, M. Bennis et al., "Big data caching for networking: moving from cloud to edge," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 36–42, 2016.

[11] N.-H. Bao, M. Tornatore, C. U. Martel, and B. Mukherjee, "Fairness-aware degradation based multipath re-provisioning strategy for post-disaster telecom mesh networks," *Journal of Optical Communications and Networking*, vol. 8, no. 6, pp. 441–450, 2016.

[12] J. Wang, Y. Guo, X. Wen, Z. Wang, Z. Li, and M. Tang, "Improving graph-based label propagation algorithm with group partition for fraud detection," *Applied Intelligence*, vol. 50, no. 10, pp. 3291–3300, 2020.

[13] N. K. Bhattacharya, "Telecom services industry -technological evolution and future opportunities," *Gigabit Newsletter*, vol. 25, no. 10, p. 71, 2018.

[14] Y. Zhang, "GroRec: a group-centric intelligent recommender system integrating social, mobile and big data technologies," *IEEE Transactions on Services Computing*, vol. 9, no. 5, pp. 786–795, 2016.

[15] S. J. Leistedt and P. Linkowski, "Fraud, individuals, and networks: a biopsychosocial model of scientific frauds," *Science & Justice*, vol. 56, no. 2, pp. 109–112, 2016.

[16] I. Ahmad, "Revenue assurance & fraud management is a constant journey of discovery and action," *Gigabit Newsletter*, vol. 24, no. 1, pp. 66-67, 2017.

[17] J. Melvin, "Energy infrastructure investments must extend to utilities' telecom networks," *Platts Megawatt Daily*, vol. 23, no. 50, pp. 6-7, 2018.

[18] A. Gent, "Fighting fraud on mobile networks," *Computer Fraud & Security*, vol. 2017, no. 2, pp. 10–13, 2017.

[19] J. Longworth, "VPN: From an obscure network to a widespread solution," *Computer Fraud & Security*, vol. 2018, no. 4, pp. 14-15, 2018.

[20] S. Galzarano, R. Giannantonio, A. Liotta, and G. Fortino, "A task-oriented framework for networked wearable computing," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 621–638, 2016.

[21] D. Specht, "The data revolution: big data, open data, data infrastructures and their consequences," *Media, Culture & Society*, vol. 37, no. 7, pp. 1110-1111, 2016.

[22] J. Wang, W. Liu, S. Kumar, and S.-F. Chang, "Learning to hash for indexing big data-a survey," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 34–57, 2016.

[23] E. D. Siew, R. K. Basu, H. Wunsch et al., "Optimizing administrative datasets to examine acute kidney injury in the era of big data: workgroup statement from the 15th ADQI consensus conference," *Canadian Journal of Kidney Health and Disease*, vol. 3, no. 1, pp. 1–12, 2016.

[24] H. Stevens, "Big data, little data, no data: scholarship in the networked world," *Journal of the Association for Information Science & Technology*, vol. 67, no. 3, pp. 751–753, 2016.

[25] Y. Zhang, M. Qiu, C. W. Tsai, M. M. Hassan, and A. Alamri, "Health-CPS: healthcare cyber-physical system assisted by cloud and big data," *IEEE Systems Journal*, vol. 11, no. 1, pp. 88–95, 2017.