Research on Anonymous Reconstruction Method of Multi-serial Communication Information Flow Under Big Data

Ying Li(✉), Feng Jin, Xiao-xia Xie, and Bing Li

Information and Communcition College National University of Defense Technology, Xi'an 710106, China
xuennxe@163.com

Abstract. The existing methods of dynamic reconfiguration of network information flow have some drawbacks, such as security, reliability and bad influence on the performance of the original network. Therefore, an anonymous reconfiguration method of multi-serial communication information flow under large data is proposed. Firstly, the original information flow is acquired in the communication network, and the cooperative filtering of multi-serial communication is carried out. After filtering, the notification information of relay nodes is obtained in the information flow, and the communication status of the information flow is extracted. The characteristic information of the information flow is reconstructed and anonymized. Finally, the anonymous reconstruction of multi-serial communication information flow is completed. By analyzing and comparing the experimental results, it can be seen that the method proposed in this paper is superior to the traditional method in terms of both the effect of anonymity and the efficiency of operation when reconstructing the anonymous information flow of multi-serial communication, it effectively solves the shortcomings of traditional methods, such as poor anonymous effect of information flow and slow speed of information flow reconstruction. It shows that the method has a high degree of anonymity and has a strong practicability.

Keywords: Large data · Information flow · Multi-serial communication · Reconstruction method · Anonymous

1 Introduction

Big data is a new processing mode, which has stronger decision-making power, insight and process optimization ability, and can effectively deal with massive, high growth rate and diversified information assets [1, 2]. With the continuous popularization of big data and Internet in personal and commercial communications, great changes have been brought to people's lives and work [3]. After a long time of use of computer internet and the development of big data industry, people have higher requirements for the network world in the use process [4]. Internet users hope to protect their privacy while enjoying the communication services provided by multi-serial network operators [5]. The emergence of serial transmission technology is an important condition to achieve the above requirements [6].

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 348, pp. 57–68, 2021.
https://doi.org/10.1007/978-3-030-67874-6_6

Traditional multi-serial-port parallel communication data transmission system can not acquire serial passwords independently, so it needs to select and open serial ports manually, and users need to know serial passwords beforehand, which greatly reduces the efficiency of the system. Reference [7] proposed a method of ACNS collision information reconstruction based on compressed sensing. The method developed ACNS terminal based on compressed sensing theory, and proposed the process of acceleration acquisition, compression and restoration in ACNS. Taking 20 km/ h as the critical speed triggered by ACNS, the collision acceleration data at this speed can be obtained through sled collision test; based on orthogonal matching pursuit algorithm, using discrete cosine transform matrix, the collision acceleration data of ACNS are obtained, The results show that the collision information can be reconstructed accurately, but the effect of anonymity is not good. Reference [8] proposes an automatic detection method for implicit information leakage in Business Process Execution Language (BPEL) based on information flow. This method constructs a BPEL representation meta model based on Petri net for transformation and analysis. Based on the concept of position noninterference of Petri net, Petri net reachability graph is used to estimate the interference of Petri net, so as to detect the components of hidden information leakage in Web services. This method can detect the hidden information accurately, but it takes a long time to reconstruct the information. Therefore, a multi-serial parallel communication data transmission system which can independently identify serial numbers has been developed, and has been widely used in real life and achieved good results.

XON/OFF is used to complete the data transmission control of multi-serial parallel communication based on software flow. When the input data of the software in the serial port receiver is higher than the threshold value, XOFF characters are transmitted to the serial port data sender. After the sender collects the XOFF spontaneously, the sending data is terminated. Otherwise, when the amount of data at the receiving end is below the threshold, XOF characters are sent to the serial data sender and data is sent. This is the working principle and mode of multi-export communication. In this way, the transmission efficiency of user information can be guaranteed, but in addition, the user also needs to protect the identity information and privacy, and also needs to gradually realize anonymity in the transmission of data. An important goal of anonymous communication is to hide the identity or communication relationship between the two sides, so that the eavesdropper can not directly know or infer the communication relationship between the two sides. In general, information flow re engineering is also called information flow re-engineering. It is a process of optimizing the combination of information flows in business processes according to the strategic objectives of enterprises and customer needs. It can be approximated as the early planning stage of business process re-engineering. In the network data transmission and communication, the user's data information is more anonymous by reorganizing the information flow, which makes the user's information more secure in the transmission process.

2 Design of Anonymous Reconstruction Method for Information Flow

The whole method of anonymity reconstruction of information flow is realized by two steps, namely, anonymity of information flow and reconstruction of information flow. The original information flow in the communication system is reconstructed and processed. Finally, the reconstructed information flow is encrypted. Finally, the function of anonymous reconfiguration of information flow is realized, and the processed information flow is output.

2.1 Raw Information Flow Acquisition

Firstly, the data flow analyzer takes the defined information flow rules as the basis, and takes the middle of the source code output by the static security checking tool as the input to get the information flow among the variables in the source code. Then the information flow is filtered to get the information flow among the hidden channel variables. On this basis, the information flow graph is constructed. Combined with the characteristics of the hidden channel information flow and the external security rules, the hidden channel in the system source code is detected by reverse iteration traversing the information flow graph. The information flow generation method is shown in Fig. 1.

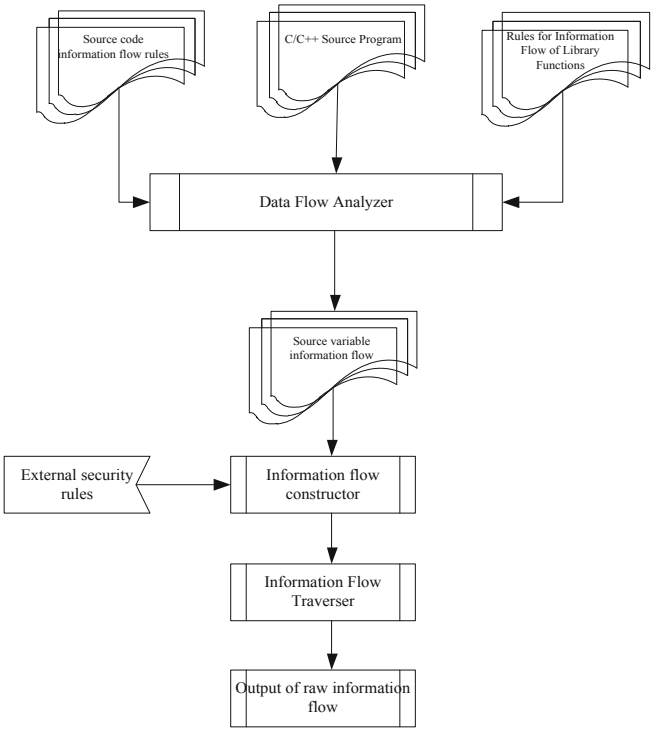

Fig. 1. Information flow generation acquisition flow chart.

The main function of the information flow generation part in the flow chart is to generate the information flow between source variables. Its main modules are as follows:

Source code information flow rule module. This module is based on the simple information flow rules given by Tsai and Gligor.

Library function information flow rule module. Functions that do not have function implementations in source code are called Library functions, including commonly known C/C ++ runtime libraries and third-party function libraries. According to the interface description of Library function, the module specifies the information flow between parameters and parameters, and between parameters and return values. When the data flow analyzer encounters a library function in the analysis process, the information flow generation rules of the matching function are extracted directly from the module, and the information flow is generated. As a supplement to the source code information flow rule module, this module ensures the integrity of the information flow.

Information flow graph constructor module. This module takes the information flow between source variables as input. Firstly, it filters out the information flow among the hidden channel variables, then organizes the information flow into an information flow graph. Finally, according to the implementation process of each module in the graph, the original information flow data can be obtained.

2.2 Collaborative Filtering of Multi-serial Communication

The data transmission of multi-serial communication includes serial data receiving module, parallel-serial conversion module and serial output selection module. The main work of serial data receiving module is level conversion and data transceiver. The original data stream is processed by multi-serial communication, which includes level conversion, UART IP characteristic parameter analysis and register control data baud rate design. The processing of data parallel-serial conversion module is mainly to analyze external channel signal, parallel-serial switching mode and timing simulation design. The processing of the serial port output selection module is mainly to analyze the serial port data of the target channel and to simulate the time series of ModelSim. It can be seen that the designed system realizes the function of multi-serial port data transmission, and the serial port baud rate can be adjusted. The multi-serial data is transmitted to the DSP processor through the UART IP of FPGA. The 8 UART IP implements the receiving of 8 kinds of serial data and the real-time adjustment of the baud rate of communication data. The data from 8 kinds of serial channels are fused into one channel and transmitted to the DSP processor serially. The expansion of multi-serial port of DSP is accomplished by FPGA, which simplifies the data transmission process of system communication and reduces the operation cost of the system. Serial parallel communication data is converted to data signal by level conversion circuit, and then transmitted to the pin of the FPGA. The designed serial data receiving module uses four MAX232 chips to complete the level conversion of 8 UART. The level conversion diagram of multi-serial communication circuit is shown in Fig. 2.

According to the conversion mode in the figure, the FPGA processor in multi-serial parallel communication data transmission is used to collect data and send it to the computer port. The display and control software is responsible for receiving data and

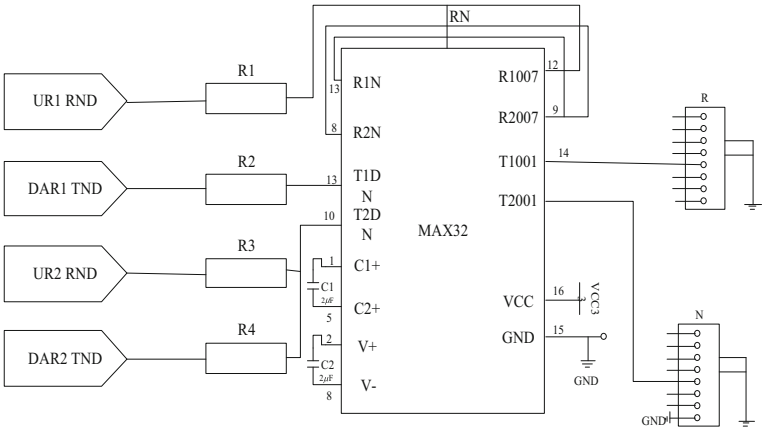

Fig. 2. Level conversion diagram of multi-serial communication circuit.

storing it. The processor transmits 8 kinds of serial channels via multiple UART IP, mainly transferring the communication data from PIO output to the parallel-serial conversion module. In this multi-serial communication architecture, collaborative filtering of the original input information flow is divided into two steps: computing the first arrival time of information, computing the potential value of users, and finally realizing the collaborative algorithm. Set up the first arrival time matrix M of each state in the information flow model. Its element m_{ij} takes i as the initial state and j as the expectation for the first time. Suppose $R = \lim_{t \to 0} Q(i,j,t)$, element r_{ij} in matrix R is the rate at which the semi-Markov process moves from state i to j. Taking constant $v \geq \max(r_i)$ and replacing the diagonal element of R with $1 - \frac{r_i}{v}$, discrete Markov chains equivalent to continuous-time Markov chains are extracted. Thus, discrete Markov chains equivalent to continuous-time Markov chains are extracted. The first arrival time matrix of discrete Markov chain is calculated and then converted to continuous time Markov chain first arrival time matrix.

If the discrete Markov chain transfer matrix and the first arrival time matrix are P_v and M_v respectively, the concrete expression formulas are shown in formulas 1 and 2.

$$P_v = I + \frac{1}{v} Q \tag{1}$$

$$M_v = \left[1 - Z_v + E(Z_v)_{dg}\right] D \tag{2}$$

In the formula, I represents the unit matrix of the original information flow; E is the matrix of all elements 1; D is a diagonal matrix with diagonal element $d_{ij} = \frac{1}{\pi(i)}$, and $\pi(i)$ represents the static distribution of discrete Markov chain state i; $(Z_v)_{dg}$ represents the matrix obtained by setting Z_v non-diagonal elements to 0, so the matrix M can be expressed as:

$$M = \frac{1}{v}(M_v)_{of} + \Lambda(M_v)_{dg} \tag{3}$$

In the formula, Λ represents a fixed constant; $(M_v)_{dg}$ represents a matrix of M_v non-diagonal primitive elements; $(M_v)_{of}$ represents a matrix with M_v diagonal elements all zeroed. The average first arrival time is calculated according to the stochastic process. Thus, the first arrival time of information flow in multi-serial communication circuits can be calculated. User's value is not only the choice of one resource, but also the influence on other users. That is, user's potential value. By calculating the potential value, the first time of information flow in multi-serial communication can be integrated, and the final collaborative filtering of original information flow can be realized. The specific process of information collaborative filtering is shown in Fig. 3.

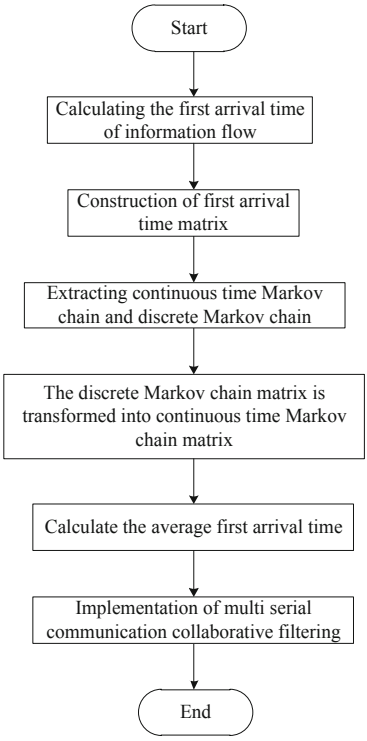

Fig. 3. Information collaborative filtering process.

2.3 Communication Information Known by Relay Nodes

Select the node T in the multi-serial communication network to receive information slices $\left(I^*_{Y_1}, I^*_{U_1}, I^*_{T_1}\right)$ and $\left(I^*_{Z_2}, I^*_{D_2}, I^*_{T_2}\right)$ from nodes S and S'. Node T grouped according to stream ID, and decoded data packets with the same stream ID number jointly.

Packets belonging to information slices T_1 and T_2 have the same ID number. Node T decodes the first piece of information received with the same stream ID number packet to get its own information. The number of information fragments contained in the data package is equal to the length of the established path, and the number of information fragments divided by the source information node is equal to the number of sub-nodes of the source information node. Node T decodes the information slice graph according to the path establishment stage, and regards the information slice received by the direct precursor node as an information slice of the data package sent to the corresponding direct successor node in the table, sorting out and sorting strictly according to the corresponding relationship in the chart. Node T takes the second piece of information received from node S as the first piece of information sent to node U. After the information slices are ordered by node T, the ordered information slices are coded by the same method as the source, multiplied by a random reversible matrix 8. The information slices received by node T are completely different from the information slices forwarded by node T, which is equivalent to further confusing and confusing the forwarded information. After T node coding, $\left(I_{U_1}^*, I_{Y_2}^*\right)$ becomes $\left(I_{U_1}^*, I_{Y_2}^*\right)'$, and its expression is:

$$\left(I_{U_1}^{*\prime}, I_{Y_2}^{*\prime}\right) = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \left(I_{U_1}^*, I_{Y_2}^*\right) = \left(I_{U_1}^*, I_{Y_2}^*\right)' \tag{4}$$

In the formula, $\begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$ is a 2 * 2 random reversible matrix, which is equivalent to a local coding coefficient matrix. The coefficient matrix obtained by the above multiplication must be reversible so as to ensure that the information received by the next relay node can be decoded. After processing, the information flow direction under relay forwarding is determined. When the direct communication mode between two nodes fails, different relay nodes are selected to complete relay forwarding of different levels of information by the relay node. On the basis of the above, extract the main features and reconstruct the state space of multi-serial communication information flow.

Let $H(x, y)$ denote two different data interference feature points of multi-serial communication information flow terminal, and the distance between the main feature vectors x and y of multi-serial communication information flow is expressed as data mining dimension. The data set is divided into $2n$ subsets. Based on sequential resampling method and principal feature matching, the expression of vector space set of feature points distribution for data to be mined is obtained. The phase space reconstruction of clustering features is achieved by defining a fuzzy clustering center and searching for multiple trajectories. When the disturbance of the state space of data storage is large, the fusion subset of state space features of data set is obtained by decomposing the semantic pheromone based on the fuzzy C-means clustering [9–11]. The reconstructed results of multi-serial communication information flow characteristics are obtained and output. According to the above method, feature extraction and state space reconstruction of large-scale multi-serial communication information flow are realized. Assuming that the density of the input data is a priori information random

variable, a large-scale data mining feature model in the cloud environment after state space reconstruction is obtained. Based on this model, data mining is carried out to improve the matching ability and balance of data mining.

2.4 Establishing an Anonymous Path to Output the Result of Anonymous Reconstruction

Source node *S* needs to establish anonymous paths before sending messages to destination node. Source node *S* needs to send the next hop IP address of each relay node to the corresponding node separately. Node S' is a reliable pseudo-source owned by source *S*, and destination *D* is randomly allocated to a certain location. Source *S* divides the IP addresses of all forwarding relay nodes except the successor nodes into two pieces. These two pieces of messages are sent to the corresponding relay node through two different paths. In order to prevent wiretappers from getting relevant information from a single message, the IP address information is multiplied by the random reversible matrix *A* before sending. At this time, the reversible matrix *A* is 2 * 2, which is equivalent to confusing encryption of the message. Source *S* sends the first half and the second half of node U's IP address to two direct forward nodes *T* and *W* along two different paths. Any successor node of its direct successor node can be known, because T receives only the first half of the IP address of *Y* and *Z* nodes, and does not know that s is the source information node and *D* node is the destination node. It only knows that node *S* is its predecessor node, and node *D* may be another forwarder. At this time, the information transmitted on the path is still the address information of the direct successor node of *T* node [12]. But information symbols have undergone tremendous changes. This ensures that even if an attacker intercepts all packets received and forwarded by all *r* nodes, it is difficult to visually find the relationship between data streams. In order to ensure the same size of data packets, the relay node should fill the data packets in the two information slices randomly before forwarding. How to effectively anonymize the reconstructed information flow is to achieve the best anonymity effect, the highest data availability and the least time and space cost. Generalization and suppression techniques are usually used to achieve anonymity of information flow, which makes information flow data more general and abstract. Bottom-up local re-coding anonymization steps are as follows:

Input: To be published table T

Step 1: PT is the result table for re-encoding identity attributes

Step 2: Check PT and add group labels to tuples that satisfy the anonymity requirement of identity preservation

Step 3: While (Number of tuples in PT with no grouping label > 0) and do (Quasi-identifying attribute groups are not generalized to the highest level)

Select a quasi-identifying property:

The selected attributes of the remaining tuples are generalized:

Adding group labels to tuples that satisfy anonymity requirements

Step 4: If (Number of tuples in PT with no grouping label > 0)

Remove tuples from Q1 groups that can be moved out of tuples and add them to the remaining tuples.

Adding group labels to tuples that satisfy anonymity requirements

Step 5: Return PT
Output: Publish table PT

3 Experimental Analysis

In order to test the application performance of anonymous reconfiguration method in multi-serial communication information flow under large data, simulation experiments are carried out. Firstly, the corresponding simulation scene needs to be built, and the simulation results of the fusion network built by the simulation scene are shown in the Fig. 4.

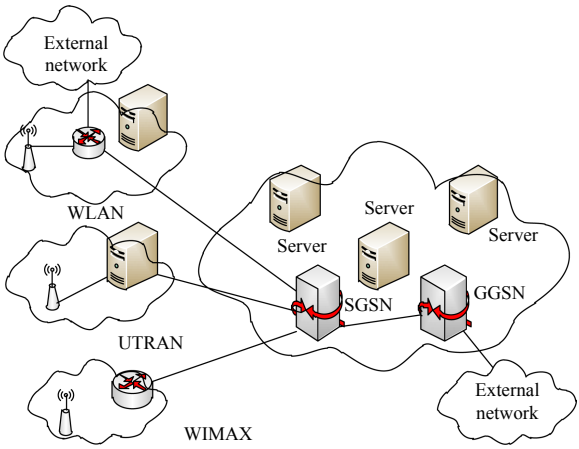

Fig. 4. Simulation experiment scene.

Because of the influence of various factors in the process of experiment, the experimental data will produce errors. In order to avoid the influence of errors on the experimental results, repeated measurement will be carried out to get the average value, so as to reduce the impact of errors.

Therefore, the designed anonymous reconstruction method is compared with the traditional reconstruction method for evaluation. In order to comprehensively evaluate and test the performance of the two methods, a set of artificial data with both uniform and Gaussian distributions is generated. Suppose that there are two attributes, and the range of each attribute is an integer in the interval [13]. The mean square deviation of the Gauss distribution is 1, and the default weight of each attribute is 1. The quality of anonymity and reconstruction error are used as experimental indicators for evaluation [14]. With the change of information flow radix, the quality of anonymity and reconstruction error of this method will also show different trends. The experimental results show that the comparison results are shown in Fig. 5.

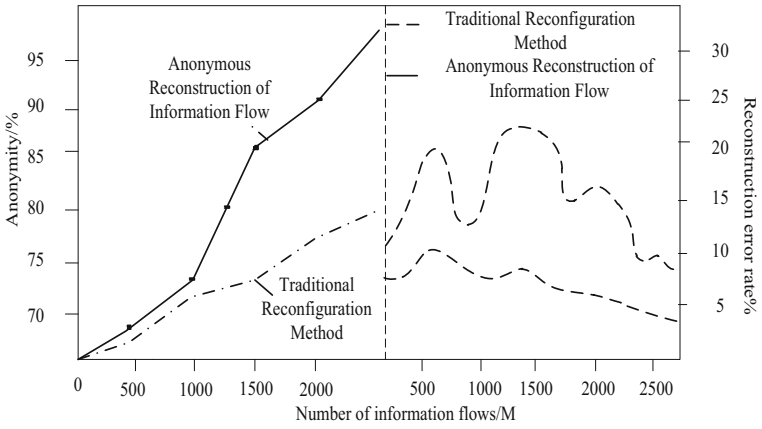

Fig. 5. Experimental comparison results

This experiment tests the reconstruction error rate and anonymity rate of the designed anonymous reconstruction method. The left side of the experimental comparison result graph represents the test result of anonymity rate. From the curve trend in the graph, it can be seen that when the amount of data in the information flow is zero, there is no anonymity rate for both methods. With the increase of information flow, the value of anonymity rate also increases. However, the traditional value of anonymity rate increases. The highest anonymity rate can only be maintained at about 80%, and the anonymity rate of the designed anonymous reconstruction method has exceeded 95% in the experiment, which can prove that the method has a high degree of anonymity. On the right side of Fig. 5, the error rate of reconstruction decreases with the increase of information flow, but the error rate of anonymous reconstruction method is more stable than that of traditional reconstruction method, which shows that the reconstruction accuracy of this method is higher.

Experimental data show that this method mainly realizes the function of anonymity and reconstruction of information flow. This method is superior to the traditional method in terms of both the effect of anonymity and the efficiency of operation.

In order to further verify the effectiveness of the design method, compare the anonymous reconfiguration time of multi serial communication information flow under different methods, and the results are shown in Fig. 6.

It can be seen from the analysis of Fig. 6 that the anonymous reconstruction time of the multi serial communication information flow in this design method is far lower than that of the traditional method, and its maximum time is only 3.9 s, while the reconstruction time of the traditional method is 7.9 s, which shows that the reconstruction efficiency of this design method is higher, and it can realize the real-time anonymous reconstruction of the communication information flow. This is because the method in this paper first obtains the original information flow in the communication network, and then cooperatively filters the multi serial communication to reduce the influence of redundant data on the speed of information reconstruction. After filtering, the

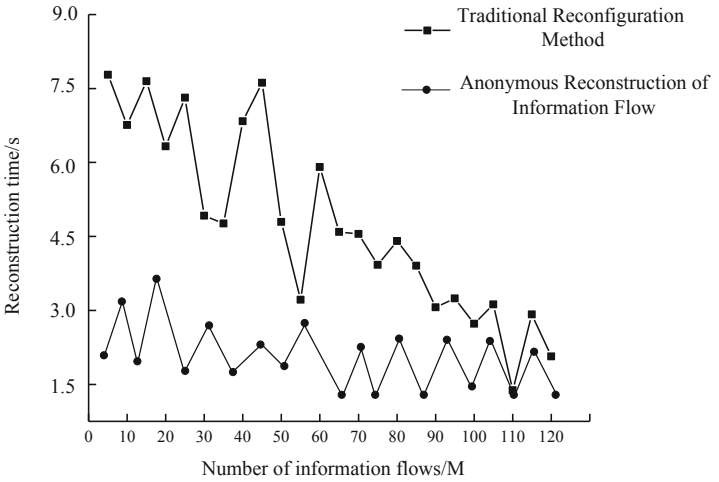

Fig. 6. Time comparison of information flow reconstruction.

notification information of relay node is obtained in the information flow, and the communication state of information flow is extracted, so as to improve the efficiency of information flow reconstruction.

4 Conclusion

Privacy protection has attracted more and more attention in various data applications. While enjoying various conveniences brought by information technology, people hope that their privacy information will be protected. Dynamic reconfiguration through anonymous reconfiguration of information flow in multi-serial communication can ensure important information security to reach the destination, and can also ensure the transmission performance of the overall network to a large extent through the diversion of information flow.

References

1. Su, Wei, Yu, Yongguang: Free information flow benefits truth seeking. J. Syst. Sci. Complexity **31**(4), 964–974 (2017). https://doi.org/10.1007/s11424-017-7078-4
2. Bijani, S., Robertson, D., Aspinall, D.: Secure information sharing in social agent interactions using information flow analysis. Eng. Appl. Artif. Intell. **70**(4), 52–66 (2018)
3. Bingwen, T.: Classified mining and optimizing technology for big data. Modern Electron. Technol. **40**(24), 34–36 (2017)
4. Ming, L., Ren, Z., Mei, H., et al.: An improved Bayesian network structure learning algorithm based on information flow. Syst. Eng. Electron. Technol. **40**(6), 25–28 (2018)
5. Wei, Q., Courtney, K.: Nursing information flow in long-term care facilities. Appl. Clinical Inform. **09**(02), 275–284 (2018)

6. Naghoosi, E., Huang, B.: Detecting the direction of information flow in instantaneous relations between variables. IEEE Trans. Control Syst. Technol. **28**(2), 542–549 (2020)
7. Lu, Y., Zhang, Y.C., Liu, Y., et al.: Reconstruction of crash data in acns based on compressive sensing. Comput. Appl. Software, **36**(09), 83–87 + 133 (2019)
8. Jiang, J.X., Huang, Z.Q., Wei-Wei, M.A., et al.: Using information flow analysis to detect implicit information leaks for web service composition. Front. Inf. Technol. Electron. Eng. **19**(04), 494–502 (2018)
9. Biondi, F., Kawamoto, Y., Legay, A., Traonouez, L.-M.: Hybrid statistical estimation of mutual information and its application to information flow. Formal Aspects of Comput. **31** (2), 165–206 (2018). https://doi.org/10.1007/s00165-018-0469-z
10. Luis, B., Andrés, T., Vicente, J., et al.: The information flow problem in multi-agent systems. Eng. Appl. Artif. Intell. **70**(4), 130–141 (2018)
11. Wahl, B., Feudel, U., Hlinka, J., et al.: Residual predictive information flow in the tight coupling limit: analytic insights from a minimalistic model. Entropy **21**(10), 1010 (2019)
12. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex and Intell. Syst. (2020). http://doi.org/10.1007/s40747-020-00161-4
13. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
14. Liu, S., Glowatz, M., Zappatore, M., Gao, H., Gao, B., Bucciero, A.: E-Learning, E-Education, and Online Training, pp. 1–374. Springer, USA (2020). http://doi.org/10.1007/978-3-319-49625-2