A Novel Data Traceability Model Based on Blockchain and Digital Watermarking in Edge Computing

Caiyun Liu, Xuehong Chen, Jun Li*, Shuaifeng Yang and Yan Sun
China Industrial Control Systems Cyber Emergency Response Team, Beijing 100040, China
liucaiyun@cics-cert.org.cn
Corresponding author email: cics_2020@163.com

Abstract. Most of the current data traceability models are centralized models, which cannot meet the traceability requirements of distributed edge nodes. Blockchain is a decentralized trust management mechanism based on the Internet. Its features that are difficult to tamper with and traceable provide a new solution to secure data traceability. However, many edge nodes have limited computing and storage resources, and cannot store large amounts of blockchain information. In order to solve the above problems, we propose a novel distributed traceability model based on digital watermarking and blockchain in edge computing. The model improves the security of the traditional traceability model by dividing the internal and external areas and electing the master nodes, reduces the storage capacity of resource-constrained nodes and solves the problem of poor scalability of the blockchain. The model can be applied to widely distributed edge node traceability scenarios, and can also trace the data in the case of watermark failure.

1. Introduction
With the development of emerging technologies such as artificial intelligence, the Internet of Things, and 5G, a large number of intelligent terminal devices have appeared and connected to the network. These devices are not only data consumers but also data producers. The number of devices connected to the network will reach 50 billion, and the total amount of global data will reach 175ZB by 2020[1]. If these data are uploaded to the cloud for processing, it will not only occupy a large amount of network bandwidth, but also increase the burden on the cloud data center, which causes the data transmission and information acquisition become worse and worse, resulting in problems such as service interruption and network delay. In addition, due to the relatively long transmission distance of terminal devices and centralized cloud computing data centers, the service quality of some applications that require mobile support, location awareness, and low latency is difficult to be guaranteed.

Therefore, edge computing, a new computing paradigm that is located at the edge of the network close to terminal devices, has gradually attracted the attention of the academia and industry[2]. It can provide compute and storage resources close to the user in open standards and ubiquitous manner in order to reduce latency. In the edge computing concept, any node with computing power can be used as an edge node to analyze data nearby and output analysis results, thereby reducing network bandwidth consumption and improving data real-time[3]. Industrial control systems have a wide range of data distribution and high real-time requirements. Therefore, in recent years, edge computing...
technology has been widely used in industrial control systems. The temperature and humidity environment observation data generated by massively deployed sensors, industrial field real-time data monitored by video monitors and a large amount of control and analysis data generated by the industrial field distributed control system (DCS) do not need to be uploaded to the data center, and certain storage, analysis, and applications can be realized in the edge nodes.

However, due to the widespread distribution of edge nodes and limited computing capability, their security protection ability is weaker than centralized data centers, resulting in data in these edge nodes facing security risks such as loss, tampering, and forgery. Therefore, data traceability technology is essential for data security analysis of edge nodes, but most traditional data traceability systems use a centralized storage model, which is susceptible to internal and external attacks and has the drawbacks of single points of failure. The edge nodes are highly dispersed, and after data security problems occur, it is difficult to track the location of data destruction through centralized data traceability technology.

Therefore, decentralized trusted data traceability technology has become a solution that needs to be studied urgently in the scattered edge nodes. Since 2014, the blockchain technology behind Bitcoin has received widespread attention and rapid development, and its anti-counterfeiting traceability feature is considered to be one of the most promising blockchain landing performance. The distributed chain structure of the blockchain can be well applied in the distributed structure of the edge nodes. Private chains and alliance chains can also meet the security scenarios of relatively closed industrial control systems. In addition, the latest digital watermarking technology has been able to trace data to a certain extent, and it takes up less storage space. Therefore, this paper designs a decentralized edge data traceability model based on digital watermarking and blockchain, to achieve traceability of edge node data and security of traceability information itself with minimal computing resource loss.

2. Related Works
The study of data traceability originated in the 1990s, also known as data log, data archive or data origin[4]. At present, there are many related researches on data tracing at home and abroad, mainly presenting four core views. One is to query the source and flow of historical data recorded by the log audit method[2], but this method is more difficult in actual production. A special person or machine audit is performed on each transmission end. At the same time, storing a large amount of log information increases the consumption of enterprise storage resources.

The second is to embed information or make annotations in the data. When data is leaked, the source of the data leak can be traced by embedding information or annotations. Specifically, it is divided into labeling method and watermark information embedding method. The annotation method traces the historical state of data by recording and processing related information such as background, author, time, provenance and other data information. Regarding labeling information, Buneman P et al.[5] first proposed a why-where-how type of provenance. In order to make the model more comprehensive, Sudha R et al.[6] then proposed a 7W model, which is who-when-where-how-which-what-why model for information annotation. This method is simple to implement, but consumes a large amount of storage resources. Watermark information embedding method is a new data source traceability method in recent years. Its core idea is to derive database fingerprint technology from database watermark information and distribute it to different users with different data. Because the distributed data is unique, it can be directly performed data traceability[7]. This method occupies small storage space and can effectively trace the attack path.

The third is the reverse query method, also called the inverse function method. This method inverts the query by inverse query or constructing an inverse function, or pushes it backwards according to the conversion process, and traces back the original data from the results. Because the computing is performed only when retrospective is needed, it is also called the lazy method. The difficulty of this method lies in the construction of an inverse function with good performance, high computational complexity, but relatively low storage consumption. The fourth is to implement data traceability based on blockchain technology. Blockchain technology is a key technology to solve the problem of trust. By sharing sharable data on the blockchain network in the form of a blockchain shared ledger,
multiple parties can participate in, maintain, and monitor the blockchain network ecology. Realizing the needs of shared data that cannot be tampered with and traceable can build a trust foundation at a low cost in the context of multi-stakeholder participation[8], so it is gradually applied to different data traceability fields. For example, Hao Z et al.[9] designed a food safety risk traceability method based on blockchain, Li C et al.[10] designs a digital point trading system based on blockchain technology, and Yu WJ et al.[11] designed a proposes a safe and feasible network transaction system model. In addition, Liang X et al.[12] proposed a decentralized and trusted cloud data provenance architecture using blockchain technology. It can provide tamper-proof records, enable the transparency of data accountability in the cloud, and help to enhance the privacy and availability of the provenance data.

Shafagh H et al.[13] proposed a distributed data storage and sharing method based on the blockchain, which is simple and effective for decentralized edge node information storage.

Therefore, based on the above research basis, this paper combines data watermarking information embedding technology with blockchain distributed storage technology for the first time to build a data source traceability framework suitable for decentralized edge nodes. This method can realize the traceability of edge data with less resource loss, thereby improving the ability of data security and traceability. It has important research significance for quickly locating and handling security threats and promoting production safety.

3. Distributed Data Traceability Model
Blockchain technology is a new type of distributed computing and storage paradigm. It uses distributed consensus algorithms to generate and update data, and uses peer-to-peer networks to transfer data between nodes. At the same time, it combines cryptographic technology to ensure that the data stored in the ledger cannot be tampered with, and uses automated scripts or smart contracts to implement upper-layer application logic. Its workflow includes three steps: block generation, consensus verification and ledger maintenance. During the ledger maintenance process, nodes will store the verified ledger data for a long time and provide functions such as back-testing. This belongs to the blockchain's own data traceability function. However, the blockchain also has problems such as low efficiency, large power consumption, and poor scalability. In industrial edge computing, the number of edge nodes is large, and the computing resources of the underlying edge devices are limited, which requires high real-time performance. Not all edge nodes have the computing and storage capabilities of blockchain. Considering that the current digital watermarking technology is gradually applied to data traceability due to its features such as strong concealment, security, detectability, robustness, and small space occupation[14], this paper uses digital watermarking and blockchain to build a safe and efficient data traceability model for edge nodes.

The proposed model is composed of blockchain network and internal network. Nodes in the blockchain network are elected by edge nodes and the elected node has the strongest computing power in its region. The internal network uses digital watermarking for data distribution and traceability. When the data to be traced crosses its area, the data is traced through the blockchain network composed of master nodes.

The steps for building a blockchain network are as follows:

The first step is to divide the area. We need to divide the edge nodes in the edge network into n areas according to their physical distance.

The second step is to conduct node elections. We need to select edge node managements for each region, which have the strongest computing power in the region they represent. Node managements are also called master nodes.

The third step is to perform node networking. In this step, the master node forms a peer-to-peer network according to Fig.1 to establish a communication link.
The fourth step is to deploy functional modules and security mechanisms. In this step, master nodes embed blockchain computing logic, and deploy algorithms or mechanisms such as data encryption, desensitization, signatures, and access control to ensure end-to-end data credibility, secure transmission, and authorization access.

The fifth step is data transmission. At this stage, data is transferred between master nodes according to the blockchain chain structure. Its flow structure diagram is shown in Fig. 2.

Each block in the blockchain contains the hash value of the previous block, so it can be traced from any block on the chain to the first block. Take Fig.1 as an example. If the data transmission path is $EP1 \rightarrow EP2 \rightarrow EP3 \rightarrow EP7$, and the operation experienced by each node is $EP1$ (DELETE)$\rightarrow EP2$ (INSERT)$\rightarrow EP3$ (MODIFY)$\rightarrow EP7$ (None), then each operation of the data in this block will be recorded in the transaction in each block. If we need to trace the original data from $EP7$, we can gradually infer the preamble from the current block until the source of the data.

After each master node is determined, the edge nodes in each area also need to generate their own watermark information. In order to reduce the amount of calculation and storage of edge nodes, each edge node uses its own fingerprint information and data manipulation information to generate watermark information, and embeds the watermark information into sacrificial digital bits. The operation of each node on data can be traced back. If there is data transmission outside this area, the master node should store the transmission record and the watermark information of the related node. Storage records include: watermark information, transmission object, transmission time. At the same time, the master node also needs to maintain the node list of this management area, and the hash value of the node’s ID is recorded in the list. When adding or deleting a node, you need to modify it in the list.
Assuming the existing data set states $DS_1$ and $DS_2$, $DS_1$ represents the data state of the starting node, and $DS_2$ represents the data state of the stopping node. $I_1$ represents the data transfer of edge nodes in the region; $I_2$ represents the data transfer of edge nodes and master nodes in the region; $I_3$ represents the data transfer between master nodes; $I_4$ represents the data transfer of edge nodes across regions, and $I_5$ represents the master node and other regions. The data flow at the edge node, $I_6$ indicates that the watermark information is invalid. $Q$ means tracing through the blockchain, $W$ means direct tracing through the watermark, $S$ means table lookup operation, and $O$ means exhaustive investigation. Taking the data set "flow path: state change" as the index of the flow method, taking the network in Fig.3 as an example, the data traceability path is as follows:

$$
\begin{align*}
I_{1}(E1 \rightarrow E2) &: (DS_1 \rightarrow DS_2) \\
&: E2 \xrightarrow{w} E1 \\
I_{2}(E5 \rightarrow EP_2) &: (DS_1 \rightarrow DS_2) \\
&: EP_2 \xrightarrow{s} E5 \\
I_{3}(EP_2 \rightarrow EP_3) &: (DS_1 \rightarrow DS_2) \\
&: EP_3 \xrightarrow{Q} EP_2 \\
I_{4}(E1 \rightarrow E5) &: (DS_1 \rightarrow DS_2) \\
&: E5 \xrightarrow{w} EP_2 \xrightarrow{Q} EP_1 \xrightarrow{S} E1 \\
I_{5}(EP_1 \rightarrow E5) &: (DS_1 \rightarrow DS_2) \\
&: E5 \xrightarrow{w} EP_2 \xrightarrow{Q} EP_1 \\
I_{6}(E2 \rightarrow E5) &: (DS_1 \rightarrow DS_2) \\
&: E5 \xrightarrow{O} EP_2 \xrightarrow{Q} EP_1 \xrightarrow{S} E2
\end{align*}
$$

Among them, the path from $EP_1$ to $E2$ in Case $I_6$ can be compared by the block time stamp and the data transmission time recorded by the master node, and the node with the smallest time difference can be temporarily determined as the data source.

This model not only considers the problem of limited storage resources of edge nodes, but also considers the problem of data tracing after watermark failure, and has stronger usability and robustness.

Compared with data watermark tracing, the combination of blockchain and watermark tracing can be more robust, and it can reduce the tracing path and tracing time when exhaustive tracing must be performed. For example, there are 100 nodes interacting with $E1$ in an hour. Among them, 30 nodes are from the area represented by $EP_1$, 20 nodes are from the area represented by $EP_2$, and 50 nodes are from the area represented by $EP_3$. In the data tracing process, even if the watermark information is damaged, the $E1$ node only needs to exhaust three nodes, and through these three master nodes to initially determine the data source.
Compared with the complete source tracing of the blockchain, the storage of the block information of the edge nodes can be reduced, and the trusted traceability can also be performed for the nodes that cannot have the computing power of the blockchain. At the same time, when adding or deleting edge nodes in the edge network, the edge nodes can be incorporated into the sub-region chain, which is equivalent to make up for the problem of poor scalability of the blockchain model. It effectively solves the problems of limited storage resources or limited computing resources of some nodes and poor scalability of the blockchain.

4. Performance Analysis

This section analyzes the advantages and disadvantages of the model proposed in this article by comparing and contrasting the model proposed in this article with the traditional data tracing model. Use "+" to indicate the increase of this model, "-" to indicate the decrease of this model, and "=" to indicate that they are equal.

### Table 1 Performance comparison of master node

| Traceability model | Robustness | Storage space consumption | Computational complexity | Safety |
|--------------------|------------|---------------------------|--------------------------|--------|
| Notation           | +          | +                         | +                        | +      |
| Reverse query      | +          | +                         | -                        | +      |
| Blockchain         | +          | +                         | =                        | =      |

### Table 2 Performance comparison of edge node

| Traceability model | Robustness | Storage space consumption | Computational complexity | Safety |
|--------------------|------------|---------------------------|--------------------------|--------|
| Notation           | +          | —                         | +                        | +      |
| Reverse query      | +          | —                         | +                        | +      |
| Blockchain         | —          | —                         | —                        | —      |

Exhaustive time analysis after watermark failure: Suppose there are $n$ edge nodes for communication, and $m(m < n)$ master nodes are elected. If the watermark fails and the proposed model is not used, the exhaustive complexity of the termination node is $n-1$. If the proposed model is used, only consume one table lookup and time comparison, one blockchain tracing and one Exhaustive operation with complexity $m-1$. Because the values of $m$ and $n$ are different, the performance of the model will change accordingly in practical applications. Suppose an exhaustion takes $1 ms$, a table lookup takes $0.1 * k ms$, where $k$ is the number of table records, and a blockchain traceback takes $qt$. Then the performance comparison between this model and the full watermark model is shown in Fig.4, Fig.5, Fig.6 and Fig.7. The horizontal axis represents $m$, and the vertical axis represents the time consumed.
In Fig.4, Fig.5, Fig.6 and Fig.7, when \( m = 0 \), it is equivalent to not selecting the master node, that is, the watermark model traceability is fully applied, and the \( t \) value at this time is at the peak. It can be seen that once the watermark information fails, the model proposed in this paper can trace the data source more quickly and has stronger robustness.

5. Conclusion

This paper proposes a new data tracing model based on blockchain and data watermarking technology for industrial edge nodes. This model selects master nodes for blockchain communication and uses data watermarking for data tracing between internal nodes. The model strengthens the security, robustness and usability of the data traceability model, and can quickly trace the source when the watermark fails, while making up for the problem of poor scalability of the blockchain model. The proposed model can be well applied to the data tracing of large edge networks, so as to improve the ability of data tracing.

Acknowledgments

This work was financially supported by The National Key Research and Development Program of China(2018YFB2100400).

References

[1] Vaquero L M, Roderomerino L. (2014) Finding your Way in the Fog: Towards a Comprehensive Definition of Fog Computing.AcmSingcomm Computer Communication Review, 44(5):27-32.

[2] Yousefpour A, Fung C, Nguyen T T, et al. (2019) All one needs to know about fog computing and related edge computing paradigms: A complete survey. Journal of Systems Architecture, 289-330.

[3] Kyoung-Don K , Chen L, Hyungdae Y, et al. (2017) Real-Time Information Derivation from Big Sensor Data via Edge Computing. Big Data & Cognitive Computing, 1(1):5-10.

[4] Wang C T. (2004) Research problems in data provenance. ieee data eng bull, 27:45-52.

[5] BunemanP , Khanna S , Wang C T. (2001) Why and Where: A Characterization of Data Provenance. proc icdt, 1973:316-330.

[6] Sudha R, Liu J. (2009) A New Perspective on Semantics of Data Provenance. In: international semantic web conference.Washington DC. pp. 35-40.

[7] Qasim A F ,Meziane F , Aspin R. (2018) Digital watermarking: Applicability for developing trust in medical imaging workflows state of the art review. Computer Science Review, 27(FEB.):45-60.

[8] Zheng Z ,Xie S , Dai H , et al. (2017) An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends . In:6th IEEE International Congress on Big Data. pp.557-564.

[9] Hao Z , Mao D , Zhang B , et al. (2020) A Novel Visual Analysis Method of Food Safety Risk Traceability Based on Blockchain. International Journal of Environmental Research and
Public Health, 17(7):2300.

[10] Li C, Dai B R, Zhao X F, et al. (2018) Design and Implementation of Digital Credit Trading System Based on Blockchain Technology. Modern Computer. 27.

[11] Yu W J, Wu Y. (2018) Research on Network Trading System Using Blockchain Technology. In: International Conference on Intelligent Informatics & Biomedical Sciences. IEEE Computer Society. Bangkok. pp. 93-97.

[12] Liang X, Sachin S, Deepak T, et al. (2017) ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability. In: 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). pp. 468-477.

[13] Shafagh H, Burkhalter L, Hithnawi A, et al. (2017) Towards Blockchain-based Auditable Storage and Sharing of IoT Data In: NSDI 2017 - 14th USENIX Symposium on Networked Systems Design and Implementation. United States. pp. 1-2.

[14] Zhou G, Wu K. (2015) Research on zero-watermarking model of relational databases based on improved C4.5 algorithm. Computer Applications & Software. 32(1):64-67.