Research and Analysis of Blockchain Data

IOP Conf. Series: Journal of Physics: Conf. Series **1237** (2019) 022084 doi:10.1088/1742-6596/1237/2/022084

Research and Analysis of Blockchain Data

Xiaojing Yang, Jinshan Liu, Xiaohe Li

College of Computer Science, Xi'an Shiyou University, Xi'an 710065, China

xjyang@xsyu.edu.cn

Abstract. Blockchain technology is characterized by anti-counterfeiting, non-tampering and easy to implement smart contracts, and is known as a new technology that will lead to social change. Therefore, the study of data in blockchain has important theoretical and practical significance. The author proposes a three-layer model from the perspective of data analysis; on the basis of this model, the data structure and data type of blockchain, as well as the data structure and operation principle of smart contract are studied. Finally, the seven problems and correlations of blockchain data analysis are summarized.

1. Introduction

Blockchain is a completely new distributed infrastructure and computing paradigm formed by reorganizing mature technologies such as hash function, Merkle tree, Proof of work (PoW) and combining cryptographic technologies such as public key encryption, digital signature and zero-knowledge proof. Blockchain technology is characterized by anti-counterfeiting, non-tampering and easy to implement smart contracts, and is known as a new technology that will lead to social change.

In order to realize trusted transactions in distributed environment, blockchain technology uses cryptographic technology to hide user information, while all transaction information is verified and stored by distributed network. Blockchain technology can be divided into public chain, alliance chain and private chain according to different application scenarios and network admission mechanism. Bitcoin, Ethereum etc., have no restrictions on the joining and exiting of nodes, are typical public chains. In the public blockchain, data can be easily accessed, which provides an unprecedented opportunity for data analysts to analyze various behaviors in the system by using transaction data.

At present, various public chains such as Bitcoin, Ethereum etc., have obtained a large number of users and accumulated a large amount of transaction data. Taking Bitcoin as an example, a joint study by ARK Investment Company and Coinbase points out that by the end of 2016, more than 10 million users worldwide hold Bitcoin, and Bitcoin trades $200 million a day. With the development of blockchain technology, various industries have introduced blockchain technology as the underlying technology, which will inevitably lead to the existence of a large amount of data in the form of blockchain data. Therefore, the study of blockchain data has important theoretical and practical significance. At the same time, since the blockchain technology is still in the initial stage, the analysis of blockchain data is also in the exploratory stage, and there is a lack of research on blockchain data.

Organizational structure of this paper: Section 1 briefly introduced the basic concepts of blockchain; Section 2 proposed a three-layer model of blockchain from the perspective of data analysis; Section 3 analyzed and studied the data structure and data type in blockchain; Section 4 analyzed the data structure and operational mechanism of smart contracts; Section 5 summarizes seven issues and interconnections in the field of blockchain data research. Section 6 conclusions.

2. Blockchain architecture

Blockchain is the underlying technology of Bitcoin. In 2008, a scholar named "Satoshi Nakamoto" proposed a digital currency called Bitcoin. In the absence of any authoritative intermediaries, people who do not trust can pay directly in Bitcoin. In October 2016, China Blockchain Technology and Application Development White Paper published by the Ministry of Industry and Information Technology described Blockchain Technology as a new application mode of computer technology, such as distributed data storage, point-to-point transmission, and consensus mechanism and encryption algorithm. In terms of the technical architecture of the blockchain, Yuan Yong et al. proposed a five-layer framework: data layer, network layer, consensus layer, incentive layer, contract layer and application layer. Literature proposes a three-tier framework from the perspective of privacy protection: network layer, transaction layer and application layer. From the perspective of facilitating data analysis, this paper proposed a three-tier framework: application layer, contract layer, transaction layer.

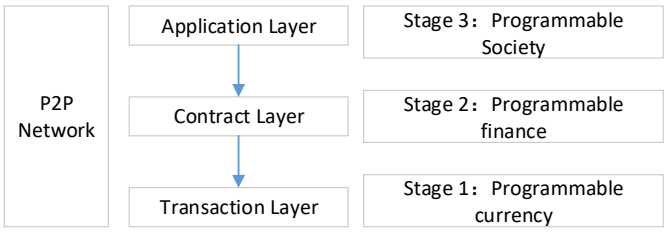

Fig 1: The blockchain framework

The three-tier framework not only represents three types of block chain data, but also represents three stages of block chain development. The bottom layer is the transaction layer, corresponding to the block chain development stage 1.0, can be programmed currency, represented by Bitcoin. Transaction is not only a means to change blockchain data, but also an important basic data in blockchain. Recording transactions to ensure the global uniqueness and irreversible modification of blockchain data is at the heart of the blockchain 1.0.

The middle layer is the contract layer, corresponding to the 2.0 phase of the blockchain development: programmable finance, Ethereum is representative. Smart contract was proposed by Szabo in 1994, a digital protocol that uses algorithms and procedures to compile contract terms, deploys on blockchain, and can be automatically executed according to rules. The smart contract itself is also an important data type for Blockchain 2.0, which we call contract data. Of course, the deployment and operation of smart contracts is not only inseparable from transactions, but also generates new transaction data.

The top layer is the application layer, corresponding to the 3.0 phase of blockchain development: programmable society, such as decentralized application, decentralized autonomous organization. The application layer can be built on top of the contract layer to implement many complex automation functions, or directly on the transaction layer (thus the contract layer is represented by dotted lines). At present, because the blockchain technology is still in the early stage of development, there is a lack of application data combined with actual scenarios.

The operating environment of the blockchain is distributed, and its three-tier framework is in a distributed environment. For example, the same smart contract is stored as a type of data on every node in a distributed environment. When a transaction triggering contract operation is received, each node runs the relevant contract based on the local blockchain data and stores the running results in the local data. Finally, through the consensus mechanism, the local data is consistent with the entire distributed network. In a distributed network environment of blockchain, there are usually a large number of nodes.

3. Blockchain data structure and data type

3.1 blockchain data structure

In order to achieve data tampering, the blockchain is constructed into a chain structure in units of blocks. The data structures on different blockchain platforms are slightly different, but basically the same. Taking Bitcoin as an example, each block is composed of a block header and a block body, and block body stores multiple transactions that occurred after the previous block; block header stores a pre-block hash (PrevBlockHase) , random number (Nonce), Merkle root (MerkleRoot), PoW etc. The blockchain guarantees irreversible modification of data based on two hash structures, Merkle tree and block list, as shown in Figure 2.

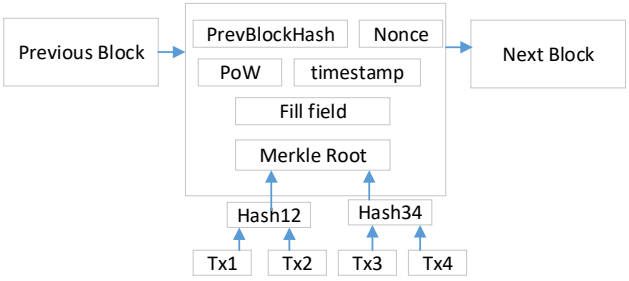

Fig 2: blockchain structure

(1) The Merkle tree. It was first proposed by Ralph Merkle and was originally used to generate digests of digital certificate directories. Bitcoin uses the simplest binary Merkle tree. Each node in the tree is a hash value, and stores a SHA256 hash value of a transaction data. The values of two leaf nodes are spliced together, and then the values of the parent node are obtained by hashing operation. Repeatedly calculate the hash value of the parent node until the root is generated, that is, transaction merkle root. With the Merkle root, tampering of any transaction data in the block is detected, ensuring the integrity of the transaction data. Without the involvement of other nodes in the tree, we can confirm whether a transaction occurs based on the SPV (Simplified Payment Verification) only according to the direct branch from the transaction node to the Merkle root path. For example, only the nodes Hash3, Hash12, and Merkle roots in Figure 2 can be used to verify that the transaction Tx4 is located in the block. In a block consisting of N transactions, at most 2log2 (N) hash operations can be performed to verify the existence of the transaction. Without the need for full data participation, customers can verify the existence of a transaction, making it ideal for building light clients or e-wallets. In Ethereum, the Merkle root is calculated using the Merkle Patricia tree because the transaction data in the block does not change much, but the status data often changes and the number is large. When building a new block, the MerklePatricia tree only needs to calculate the account status that has changed in the new block, and branches with no change in status can be directly referenced without recalculation.

(2) Block list. The hash value of the block header is obtained by performing a SHA256 hash operation on the metadata such as the Prev-BlockHash, the Nonce, and the Merkle root in the block header. As shown in Figure 2, PrevBlockHash stores hash value of the previous block. All blocks are linked together in the order of generation with PrevBlockHash as a hash pointer, forming a block list. The block header contains the transaction Merkle root, so you can verify whether the block header and transaction data have been tampered with by Hash operation. The block header also contains the previous block hash value PrevBlockHash, so it is also possible to verify whether all blocks from the previous block of the block to the creation block have been tampered with by block hashing. Relying on the PrevBlockHash, all blocks are interlocked, and any block is tampered with, causing a chain change of all the block hash pointers. When a block and all previous blocks are downloaded from an untrusted node, block hashing can be used to verify whether each block has been modified.

3.2 data type

(1) Transaction data model. A transaction in the blockchain is usually a transfer, and Figure 3 shows the data structure of the transaction in Bitcoin. Each transaction includes multiple transaction inputs and multiple transaction outputs, indicating that one transaction can merge the bitcoins in the previous plurality of accounts and transfer them to another plurality of accounts. Each transaction input is

mainly composed of the hash value of the last transaction (PrevTxHash), the output index (Index) and the input script (ScriptSig). The input of a transaction corresponds to the output of the previous transaction, and PrevTxHash saves the hash of the previous transaction output. Index indicates that this transaction input corresponds to which transaction output in the transaction history. The script (ScriptSig) contains the signature of the Bitcoin holder on the current transaction. Each transaction output includes the transfer amount (Value) and script script-PubKey containing the recipient's public key hash (script-PubKey). In the transaction-based model, all transactions depend on the hash pointer of the previous transaction to form a list of transactions as nodes. Each transaction can be traced back to the first transaction in the trading chain Coinbase (ie bitcoin from mining). Transaction-based model can effectively prevent counterfeiting, double-flower attacks against digital currency.

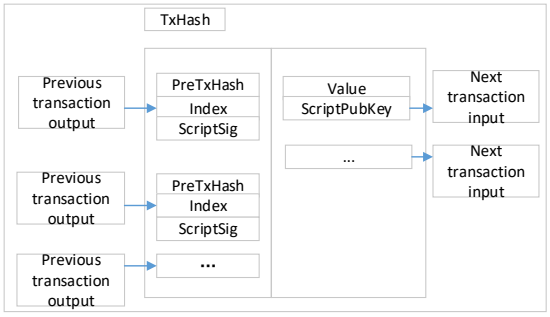

Fig 3: Data structure for bitcoin transactions

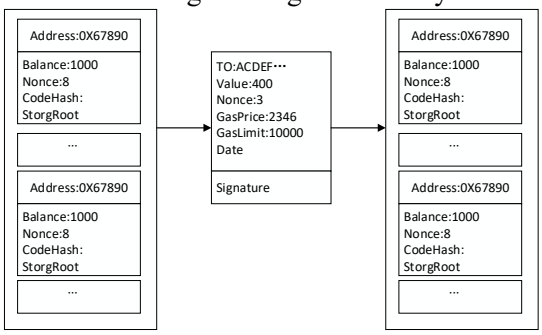

Fig 4: State transition of Ethereum accounts

(2) Account-based model. The transaction-based model makes it easy to verify transactions, but it cannot quickly query user balances. To support more types of industry applications, blockchain platforms such as Ethereum and HyperledgerFabric use an account-based model to easily query transaction balances or trading status data. Smart contracts are also better suited to build on an account-based model that is easier to handle complex business logic for state data. The accounts under Ethereum are divided into two types: external account (Externally Owned Account) and contract account (Contract Account). The external account is used to record the Ethereum balance, and the contract account is used to store the smart contract.

As shown in Figure 4. When a transaction occurs, it will trigger a change in account status. External accounts and contract accounts are represented by the same data structure in Ethereum, which contains four attributes: Balance, Nonce, CodeHash and StorageRoot. Balance records the Ethereum balance of the account. Nonce counts the number of transactions initiated by the account to prevent replay attacks. CodeHash is the hash value of the contract code when the account is applied to a smart contract. StorageRoot is the MerklePatricia root of contract status data. Ethereum's transactions include seven attributes: To, Value, Nonce, gasPrice, gasLimit, Data, and transaction signature. Attribute To saves the recipient's account address, Attribute Value is the amount of ethernic currency transferred. The attribute Nonce counts the number of transactions initiated by the transaction originator. GasPrice is the unit price of the Ether of Gas at the time of the transaction. gasLimit is the maximum amount of Gas allowed to execute this transaction. Data is the message data when the smart contract is invoked. Transaction signature is the initiator's ECDSA signature of the transaction.

4. Contract data

In 1995, cryptographer Szabo first proposed the concept of smart contract. He pointed out that "smart contracts stimulate contract execution by using protocols and user interfaces". Smart contract is a commercial contract written in a programming language that automatically enforces the terms of the contract when the predetermined conditions are met, achieving the goal of "code is the law." The existing smart contract works like the If-Then statement of other computer programs. Because blockchain data has important features such as non-tampering and complete credibility, building smart contracts based on blockchain data has natural advantages. The life cycle of smart contract includes three stages: contract generation, contract publication and contract execution.

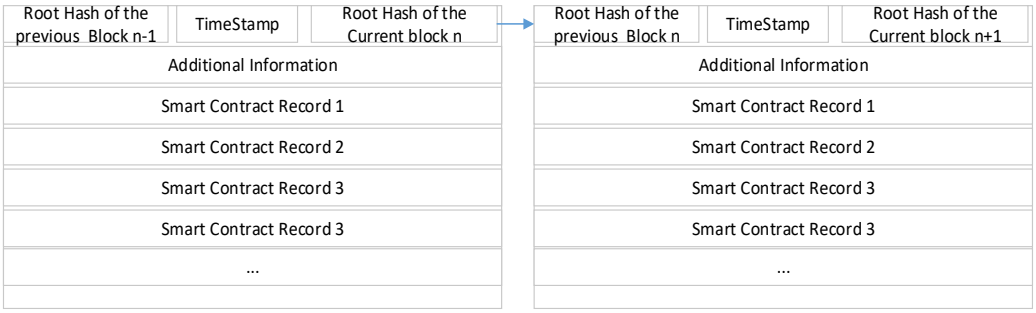

Fig 5: Blockchain diagram of smart contract

Contract generation mainly includes four aspects: multi-party negotiation, contract specification, contract verification, and contract code. The specific implementation process is as follows: the contract participants negotiate with each other; the parties clarify their respective rights and obligations, jointly determine the contract text, write the program, test the program, and finally obtain the standard contract code. Contract release is similar to transaction release. The signed contract is distributed to each node through P2P, and each node will temporarily store the received contract in memory and wait for consensus. Each node will package the temporary contract in the most recent period into a contract set, calculate the hash value of the set, and finally assemble the hash value of the contract set into a block and spread to other nodes of the whole network. The node that receives the block will compare the Hash value stored in the block with the Hash value of the contract set saved by itself. Through multiple rounds of transmission and comparison, all nodes will eventually reach a consensus on the newly issued contracts, and the agreed contracts will be distributed to all nodes in the network in block form. As shown in Figure 5. Each of the blocks contains the following information: Root Hash of the Current Block, Root Hash of the Previous Block, TimeStamp, contract data, and other descriptive information.

The execution of smart contracts is based on the "event-triggered" mechanism. The smart contract on the blockchain contains transaction processing and saving mechanisms and a complete state machine for accepting and processing various smart contracts. The smart contract periodically traverses the state machine and trigger conditions of each contract, pushing the contract that meets the trigger condition to the queue to be verified. The contract to be verified will spread to each node, just like a normal transaction. The node will first verify the signature to ensure the legitimacy of the contract. The validated contract will be successfully executed after reaching a consensus. Finally, all nodes will reach a consensus on the contract. The processing of the entire contract is automatically completed by the smart contract system built into the bottom of the blockchain, which is open and transparent and cannot be tampered with. Smart contract system can actively or passively accept, store, process and send data, and call smart contract, so as to control and manage digital objects in the chain. The smart contract technology platforms that have emerged, such as Ethereum and Hyperledger, have Turing's complete development scripting language, enabling the blockchain to support smarter contract applications for financial and social systems.

5. Problems to be solved in blockchain data analysis

Based on relevant literature analysis, the focus of blockchain data analysis can be summarized into seven aspects: entity identification, privacy risk analysis, network portrait, network visualization, transaction pattern recognition, market effect analysis, illegal behaviour detection and analysis. Figure 5 shows the relationship between these seven questions. Solid arrows indicate that the previous study supports the next one, or that the previous study is the basis for the next one. The dotted arrow indicates that the latter study is a specialization of the previous study from a certain perspective. Two-way arrows indicate different aspects of the same problem.

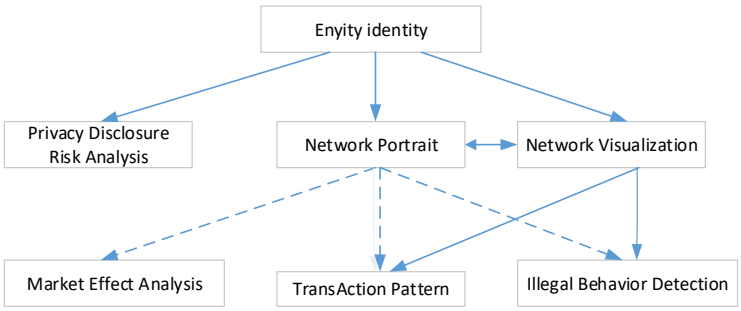

Fig.5 Research problems and their relationship

(1) Entity identification. In Bitcoin transactions, users are anonymous. And transaction involves multiple users, and one user may participate in multiple transactions at the same time. A natural question is: can you identify users from transaction records, or which addresses belong to the same user? Since it is not possible to confirm that a user is identified, it is generally considered in the literature that an entity is identified. An entity may be a user or an organization, and a user or an organization may also control multiple entities at the same time. Heuristic methods are often used to identify potential entities, which can be divided into two main types: common input method and change address method. The common input method means that all input addresses belong to the same entity in a transaction. The change address definition is a plurality of output addresses of a transaction, and the change address represents an entity.

(2) Privacy protection. Blockchain privacy protection can be divided into identity privacy protection and transaction privacy protection. Identity privacy protection requires that user's identity information, physical address, IP address are not related to user's public key, address and other public information on the blockchain. Any unauthorized node can neither obtain any information about the user's identity by relying on the data disclosed in the blockchain, nor can it track and analyse user transactions and identity information through measures such as network monitoring and traffic analysis. Transaction privacy protection requires that the data information of the transaction itself be anonymous to the unauthorized node. The transaction information in Bitcoin refers to the transaction amount, the sender's public key of the transaction, the address of the recipient, and the purchase content of the transaction. Any unauthorized node cannot obtain transaction-related knowledge through effective technical means. In some blockchain that require high privacy protection, it is also required to separate the relationship between transactions and transactions, that is, the unauthorized nodes cannot effectively infer whether the two transactions have continuity, whether they belong to the same user or the like.

(3) Network portrait. In the face of massive transaction data, researchers hope to analyse: How many users are involved in the transaction? What are the characteristics of these users? Does this huge payment network have the characteristics of a general complex network? Bitcoin as an "asset", how is it distributed among users? Do you meet the general laws of economics, etc.? And from the point of view of complex network, the characteristics of Bitcoin network are analysed.

(4) Network visualization. With the prevalence of blockchain technology, the transaction data stored in the blockchain increases rapidly. Therefore, in the face of a large and rapidly growing trading network, it is an important research direction to study its visualization tools. For example, the bitcoin trading network's visualization system, BitConeView, can be used to track bitcoin transactions in real time and intuitively. Especially, the system defines the concept of "purity" so as to find out the mixed currency transactions conveniently and monitor the potential money laundering banks in the Bitcoin network in real time.

(5) Market Effect Analysis. The prices of encrypted currencies such as Bitcoin are highly volatile. On the one hand, this extremely high volatility has attracted economists to discuss whether bitcoin is a currency or not from the perspective of finance. On the other hand, from the perspective of data analysis, researchers hope to explain the driving factors behind this extreme volatility. It is generally believed that the influencing factors include: miner behaviour; cryptographic currency system settings;

user participation; policy; number of potential users and market sentiment; competition or alternative relationship with other cryptocurrencies and traditional assets.

(6) Illegal behaviour detection. Unlike traditional bank payment systems, Bitcoin is an anonymous, non-centralized payment system. Anonymous features lead to many illegal activities in bitcoin transactions, such as money laundering, fraudulent gambling, and the sale of contraband. These illegal acts of exposure are only a small part, and a large number of illegal acts are not known. Identifying illegal behaviours based on transaction patterns and blockchain data analysis techniques is not only a need to promote the healthy development of blockchain technology. And can also provide reference for the supervision and legislation of the blockchain industry.

(7) Transaction pattern recognition. Unlike traditional payment systems such as Banks, bitcoin is an anonymous, decentralized payment system. What characterizes human payment behaviour in the context of anonymity is an interesting question. It is a valuable question whether specific patterns can be identified from blockchain transaction records to detect related illicit activity.

6. Conclusion

Blockchain technology is an emerging technology that is currently widely concerned by researchers. This paper proposes a three-layer model from the perspective of data analysis. Based on this model, this paper discusses the analysis of blockchain data structure, data types, and the data structure and operation principle of smart contracts. It also summarizes the seven problems and interrelationships of blockchain data analysis. The author hopes that this research can promote the healthy development of blockchain technology.

References

[1] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system [EB/OL]. [2017-07-19]. http://bitcoin.org/bitcoin.pdf.
[2] Szabo N. Smart contracts: Building blocks for digital markets [EB/OL]. [2018-01-30]. http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart_contracts_2.html.
[3] Churyumov A. Byteball: A decentralized system for storage and transfer of value [EB/OL]. [2018-02-08]. https:// byteball.org/Byteball.pdf.
[4] Buterin V. A next-generation smart contract and decentralized application platform [OL]. [2018-02-08]. https://cryptorating.eu/whitepapers/Ethereum/Ethereum_white_paper.pdf.
[5] Atzei N, Bartoletti M, Cimoli T. A survey of attacks on Ethereum smart contracts (SoK) [C] //Proc of the 6th Int Conf on Principles of Security and Trust. Berlin: Springer, 2017:164-186.
[6] Vasek M. The age of cryptocurrency [J]. Science, 2015, 348(6241): 1308-1309.
[7] Kosba A, Miller A, Shi E, et al. Hawk: The block chain model of cryptography and privacy-preserving smart contracts //Proceedings of the IEEE Symposium on Security and Privacy (S&P). SanJose, USA, 2016: 839-858.
[8] Wood G. Ethereum: A secure decentralized generalized transaction ledger [OL]. [2018-02-08]. http://gavwood. Com/Paper.pdf.
[9] Poon J, Dryja T. The Bitcoin lightning network: Scalable off-chain instant payments. White Paper, 2015.
[10] Zhu Liehuang et al. Survey on privacy preserving techniques for blockchain technology [J]. Journal of Computer Research and Development, 2017, 54(10): 2170-2186(inChinese) .
[11] Yuan Yong, Wang Feiyue. Blockchain: The state of the art and future trends [J]. Acta Automatica Sinica, 2016, 42 (4): 481-494 (inChinese) .
[12] Shao Qifeng, Jin Cheqing, Zhang Zhao, et al. Blockchain: architecture and research progress［J］. Chinese Journal of Computers, 2018, 41(5) : 969-988.