TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 17/18 pp176–206
DOI: 10.26599/TST.2021.9010080
Volume 28, Number 1, February 2023

Blockchain Data Analysis from the Perspective of Complex Networks: Overview

Wanshui Song, Wenyin Zhang*, Jiuru Wang, Linbo Zhai*, Pengkun Jiang, Shanyun Huang, and Bei Li

Abstract: Cryptocurrency based on blockchain technology has gradually become a choice for people to invest in, and several users have participated in the accumulation of massive transaction data. Complete transaction records in blockchains and the openness of data provide researchers with opportunities to mine and analyze data in blockchains. Network modeling and analysis of cryptocurrency transaction records are common methods in blockchain data analysis. The analysis of attribute graphs can provide insights into various economic indicators, illegal activities, and general Internet security, among others. Accordingly, this article aims to summarize and analyze the literature on cryptocurrency transaction data from the perspective of complex networks. To provide systematic guidance for researchers, we put forward a blockchain data analysis framework based on the introduction of the relevant background and reviewed the work from five aspects: blockchain data model, data acquisition on blockchains, existing analysis tools, available insights, and common analysis methods. For each aspect, we introduce the research problems, summarize the methods, and discuss the results and findings. Finally, we present future research points and several open questions in the study of cryptocurrency transaction networks.

Key words: blockchain; cryptocurrency; transaction record; complex network; data analysis; data mining

1 Introduction

Blockchain technology[1] is used to create a network environment with collective maintenance and peer-to-peer trust, which can be combined with artificial intelligence, cloud computing, and other technologies to create an intelligent business entity. As the underlying technology of Bitcoin and other cryptocurrencies[2], its unique decentralization, distrust, traceability, and other characteristics build a "trustworthy world computer", which has aroused extensive research attention from scholars.

At present, blockchain technology is still in the exploratory stage, so there are relatively few successful cases of blockchain technology applications in various industries. Transaction data generated by relatively mature public chains, such as those represented by Bitcoin, Ethereum, and EOS.IO, are still analyzed as data sources. Cryptocurrencies based on blockchain technology have gradually become the choice of investment with the development of blockchain technology. By the first quarter of 2021, the number of active cryptocurrencies soared from more than 1000 in 2018 to more than 7000. Mainstream coin arrays are represented by Bitcoin Cash (BCH), Ether (ETH), and Litecoin (LTC), etc., among others. Polkadot ecological plates are represented by Polkadot (DOT), Kusama (KSM), and Reef Finance (REEF). DeFi plates are represented by Uniswap (UNI) and SushiSwap (SUSHI), and many other currencies have emerged. The cryptocurrency total market capitalization has exceeded

• Wanshui Song, Linbo Zhai, and Shanyun Huang are with the School of Information Science and Engineering, Shandong Normal University, Jinan 250307, China. E-mail: wanshuisong@163.com; zhai@mail.sdu.edu.cn; shanyunhuang1997@163.com.
• Wenyin Zhang, Jiuru Wang, Pengkun Jiang, and Bei Li are with the School of Information Science and Engineering, Linyi University, Linyi 276005, China. E-mail: zhangwenyin@lyu.edu.cn; wangjiuru@lyu.edu.cn; jpk0220lzy@163.com; 17853099931@139.com.
* To whom correspondence should be addressed.
Manuscript received: 2021-08-27; revised: 2021-10-21; accepted: 2021-10-22

© The author(s) 2023. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

$2 trillion, of which BTC accounts for more than 50%.

From cryptocurrency to material supply chain management, an increasing number of application scenarios have promoted the application of blockchain technology in public and various industries and simultaneously increased the amount and complexity of data stored on blockchains. The data in blockchains are increasing, and the data types are becoming increasingly abundant. However, only data obtained through data analysis are valuable. The data analysis of massive blockchain data can not only bring great commercial value but also help us to understand data on blockchains.

In the face of the economic ecology of cryptocurrency with a large amount of money, mining and analyzing the transaction data of cryptocurrency stored in blockchains can not only be used to study transaction behaviors in a complex economic environment but can also be used to understand the distribution of the amount of money that pertains to users or addresses and to perform financial activities, such as price prediction through the analysis of transaction behaviors. They can also help combat illegal financial activities that use cryptocurrency as a medium of payment, such as drug trafficking and money laundering, thereby enabling law enforcement agencies to better regulate cryptocurrency markets and establish a healthy blockchain ecosystem[3].

At present, the research on cryptocurrency has a history of nearly ten years, and there is much research literature on blockchains from different perspectives, such as privacy protection[4], consensus algorithms[5], and Internet of things[6]. These papers summarize and analyze the key technologies, basic concepts, research hotspots, and applications in the field of blockchains but do not summarize and analyze the related work of blockchain data analysis.

de Haro-Olm et al.[7] investigated the relationship between privacy and anonymity in different applications of blockchain technology. Akcora et al.[8] introduced a method in their review to represent the modeling of blockchain data, summarized some auxiliary analysis tools, and introduced the insights gained through the analysis of transaction data. Chen and Zheng[9] summarized the data types in a blockchain and provided a summary of seven research problems on the data analysis of the blockchain.

With the increasingly large cryptocurrency transaction networks, an ever-increasing number of blockchain data analysts have chosen to apply social network analysis methods to cryptocurrency data analysis. Complex networks[10] have been widely proven to be a powerful tool for modeling and characterizing various complex systems. Using this method, the relationship between addresses and user entities is constructed as a network to provide insights into various economic indicators, illegal activities, and general Internet security.

Different from the existing review articles, this research comprehensively summarizes the blockchain data analysis work related to graph analysis and graph mining of the data on blockchains from the perspective of complex networks. As shown in Fig. 1, by introducing

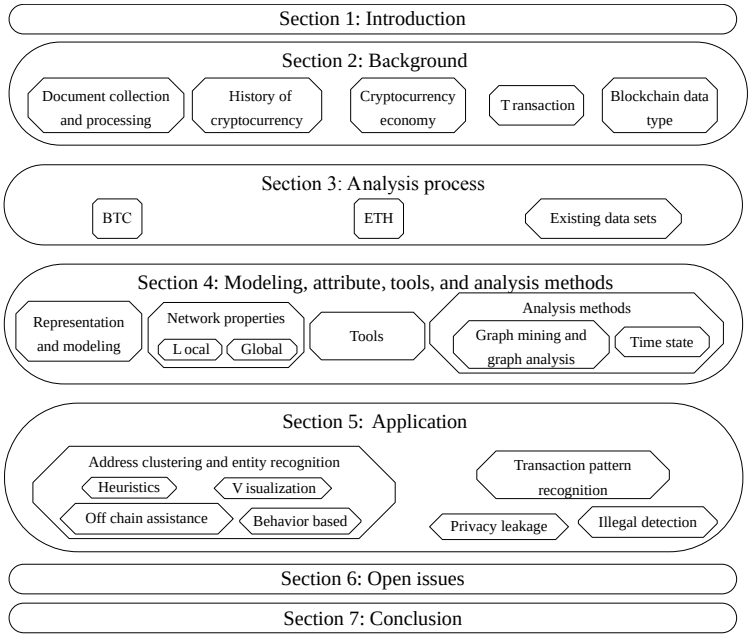


Fig. 1 Structure of this article.

the relevant background of cryptocurrency, we propose the framework of the blockchain data analysis.

This study also summarizes and analyzes the five aspects of the blockchain data model, acquisition and analysis of the data on blockchains, existing analysis tools, available insights, and common analysis methods. It also focuses on the knowledge discovery existing in the transaction records of cryptocurrency. The contents of future research in this field and several open problems are given.

The main contributions of this paper are as follows:

(1) This paper summarizes and discusses the research work on graph analysis and graph mining for the knowledge discovery of cryptocurrency transaction records from the perspective of a complex network.

(2) The detailed process of obtaining and analyzing blockchain data is given, and the basic framework of blockchain data analysis is proposed.

(3) This paper summarizes the important achievements and enlightening contents in the existing literature, puts forward several possible research and application directions for the analysis of blockchain data, and provides opinions on several important analysis methods and tools. It also provides a good research direction for blockchain data researchers.

2 Background

This chapter first presents how the references in this paper were collected and processed. Then, it describes the development history of cryptocurrency, cryptocurrency economy, cryptocurrency transaction, and blockchain data model based on addresses and accounts to better introduce the blockchain data analysis work and lay a foundation for the further analysis of transaction data.

2.1 Document collection and processing

We screened the references and related analysis tools mentioned in this article from the paper query websites and online resources. In this section, we will describe the literature search, processing methods, and collection sources.

2.1.1 Determine search keywords

After determining the theme of the research, the first step in our collection is to determine reasonable search keywords, that is, the keywords of the content of this research, to determine the most relevant articles.

We chose six initial search terms: “blockchain”, “cryptocurrency”, “transaction records”, “complex network”, “data analysis”, and “data mining”. We used the Boolean operator “and”. We first adopted the method of simultaneously searching multiple keywords to collect documents that are closest to the writing topic as much as possible. At the same time, the term “blockchain” must be included to reduce the scope of the literature search. Then, we used a snowball approach, which uses references from a document or its citations to identify additional documents. The application of this method requires that the documents found in the early stage must be very suitable for the thesis writing theme. Accordingly, after the snowball method is used, other research-related articles can be found and used as information sources to better explain the research points of the thesis.

2.1.2 Search academic articles

We mainly retrieved related articles from the following search engines or databases according to the above-mentioned search methods. These documents include published and arXiv papers. The tools introduced in the article are summarized in the literature or from the Internet.

Computer science bibliography (https://dblp.org/);
Web of Science (https://www.webofscience.com);
Google Scholar (http://scholar.google.com);
IEEE (http://ieeexplore.ieee.org);
Science Direct (http://www.sciencedirect.com);
Scopus (http://www.scopus.com);
Springer (https://www.springer.com/cn);
Semantic Scholar (https://www.semanticscholar.org/).

After searching the literature, we deleted the duplicate literature and read the paper title, abstract, keywords, and article title frame one by one to further determine whether the paper will be included in our reference range. The work of reading papers is a part of daily learning, which spanned one and a half years. The screening sorted out papers that had been read according to their title frame. Only a few documents are new or have not been read before.

2.1.3 Data collection results and classification scheme

To sort out as many documents related to the subject of this paper as possible and explain some professional terms or knowledge points, in this study, a total of 97 papers, 7 online analysis tool websites, and 1 entry website were selected.

When defining our classification scheme, we considered the research directions contained in the title of the paper and other problems in the process of data

analysis, such as data acquisition and processing. After several summaries, we summarized the references into five parts. That is, we summarized and analyzed the blockchain data analysis work from the perspective of complex networks from five parts: blockchain data model, data acquisition on the chain, existing analysis tools, available insights, and common analysis methods.

2.2 History of cryptocurrencies

The development history of cryptocurrency can be divided into four stages: namely, the early stage of digital currency germination, the Blockchain 1.0 stage represented by Bitcoin, the Blockchain 2.0 stage represented by Ethereum, and the stable currency stage represented by Tether (USDT).

In the early embryonic stage of digital currency, e-Gold, a digital currency with the concept of digital currency as the anchor of gold, appeared earlier and started its operation in 1996, reflecting the early decentralized payment concept. Subsequent digital currencies, such as Digital Monetary Trust (DMT) and VEN, as the products of early exploration, ended in failure.

The official emergence of cryptocurrency represented by Bitcoin marked the arrival of the Blockchain 1.0 stage. The period from 2009 to 2014 was in the early stage of the development of cryptocurrency. Bitcoin, with a chained structure as its data structure, is mainly used for storage and value transmission. However, this kind of currency is based on the proof-of-work algorithm and has the disadvantages of high confirmation time, high cost, and low transaction throughput, which makes it unable to be used in scenarios with high transaction rates. Bitcoin has had a huge impact on the current blockchain system, and its scarcity and hype have made its holders refer to it as “digital gold”.

Since 2014, in the Blockchain 2.0 stage represented by Ethereum, many cryptocurrencies with platform properties have emerged. With the Ethereum Virtual Machine, the implementation of smart contracts based on the Ethereum platform has become very convenient. Compared to Bitcoin, Ethereum has faster data processing speeds.

With the increasing popularity of blockchain technology, the scale of the cryptocurrency transaction market is becoming increasingly large. The stable cryptocurrency represented by USDT has begun to be used for the transactions of other cryptocurrencies against the US dollar, playing the general functions of legal currency, such as value measure, circulation means, and payment means.

As of the first quarter of 2021, there were more than 7000 active cryptocurrencies with a total market capitalization of more than $2 trillion, with BTC as the largest, and accounting for more than 50% of the market capitalization.

2.3 Cryptocurrency economy

The promotion of blockchain technology has made the cryptocurrency market prosper. As a virtual asset, the rise of cryptocurrency is affected by several linkage factors, among which the trend of the US dollar is the most influential. Similarly, as an investment option, the rise and fall of cryptocurrency prices are usually negatively correlated with the trend of the US dollar. Other factors include precious metals[11], stock markets[12], and cryptocurrencies with complementary relationships. The study also found that price changes are also affected by news and policies.

In the primary market, token-issuing companies either sell their cryptocurrencies (project tokens) directly to investors through initial coin offerings or private placements or airdrop tokens into the accounts of exchange users to promote their projects and attract investors. However, initial coin offerings may be overvalued or undervalued, as in the case of the MEME (pineapple coin) project, which airdropped 350 tokens to each of the early telegram participants, worth $1 million today.

If we call the primary market “factory direct sales”, then we can call the secondary market “mall shopping”. The secondary market is the place where all kinds of cryptocurrencies are publicly traded. Since the first public transaction of Bitcoin on an exchange in 2010, more than 7000 cryptocurrencies have been circulating on exchanges represented by Huobi, Binance, and OKEx. People can choose fiat, coin, contract, and other forms of transaction in the exchange.

Cryptocurrency is a kind of digital currency. In addition to being circulated in the market as a commodity, cryptocurrency is also a payment means similar to currency. Examples include paying miners for mining and buying goods in places that support cryptocurrency payments. Criminals also often take advantage of the anonymity of cryptocurrencies to engage in illegal activities, such as money laundering and hacking.

Cryptocurrency, similar to stocks and funds,

has gradually become the investment choice of people, gaining the participation of several users and accumulating a large amount of transaction data from exchanges. However, exchanges cannot make such data public, and analysts cannot access them through exchanges. Complete transaction records recorded on Bitcoin, Ethereum, and other public chains and the characteristics of data disclosure provide an opportunity for researchers to perform data mining and knowledge discovery on-chain data.

2.4 Transactions

Transaction data are the finest granularity of blockchain data, and the data uplink must be packaged by the miner node to the main network in the form of transactions. In the cryptocurrency market, a transaction is a process of transferring cryptocurrency from one account address to another. If this process is finally confirmed, then the transaction will be recorded in the "block". Basic fields, such as sender, receiver, amount, time, and remarks, can be obtained from transactions.

Only transactions recorded in the chain can be called effective transactions. For example, when Alice transfers money to Bob, the transaction produced goes through five processes: transaction creation, broadcasting, mining, packaging, and confirmation.

After the transaction is created, the transaction information will be broadcast to other nodes to verify the legitimacy of the transaction, and the verified transactions will be placed in the transaction pool for packaging. Then, these transactions will be constructed into a Merkle tree, and the Merkle root will be calculated. Version, prevBlockHash, target, timestamp, and nonce are combined to perform the hash operation until the calculated value is $\leqslant$ target, which means that mining has been completed. Miners package the transaction into blocks and publish it to the blockchain network and broadcast it to other nodes for verification. If the verification is passed, then it can be published to the blockchain.

The transaction confirmation process is a process from an effective transaction to an instrumental transaction, that is, from a prepayment status to a successful status. Such a process becomes an instrumental transaction after being confirmed by six blocks.

In the secondary market, each user can charge and withdraw money through his own address or the address provided by the exchange. Taking USDT as an example, there are three main types of USDT charging and withdrawing coins; Omni-USDT based on a Bitcoin network with high security, expensive handling fees, and slow transaction speed. Good security, high handling fees, and relatively fast transaction speed are suitable for USDT ERC-20 based on the Ethereum network in daily transactions. USDT TRC-20 based on TRON has a short release time and low security, but it has a zero transfer fee and submillisecond arrival. There is no difference among the three types of USDT in the exchange, but they are not interoperable among different chains. Moreover, it is difficult to retrieve them once they make mistakes during transfers. As mentioned above, in the current stable currency stage represented by USDT, USDT plays the role of a universal equivalent in the cryptocurrency economic circle, which promotes the circulation of cryptocurrency among different users and provides a large amount of analyzable data for data analysts.

2.5 Blockchain data model

At present, the dataset of data analysis mainly comes from the public chain, that is, the address-based blockchain represented by Bitcoin, Litecoin, and Monroe and the account-based blockchain represented by Ethereum and EOS. In the next sections, we will briefly introduce the blockchain transaction data model based on addresses and accounts.

2.5.1 Address-based model

In a Bitcoin blockchain, transactions are performed through addresses, not accounts. UTXO is a transaction model of Bitcoin, and it is also a core concept of Bitcoin transaction generation and verification. In a Bitcoin wallet node, we can see our balance, but there is no concept of balance in address-based blockchains, such as Bitcoin. The balance we can see is actually the product of the wallet, which is obtained by calculating all unused transactions in the user UTXO.

According to Ref. [8] in a UTXO-based blockchain, three rules of transactions pertain to the source, mapping, and balance, which can be seen through transaction data. Based on the sources of transactions, the UTXO transaction model makes transactions constitute a chain structure, and all verified and effective Bitcoin transactions can be traced back to the output of one or more transactions.

The forward can be traced back to the coinbase transaction, and the backward can find the transaction output that is not currently spent, that is, the available balance displayed in the wallet. As shown in Fig. 2, there

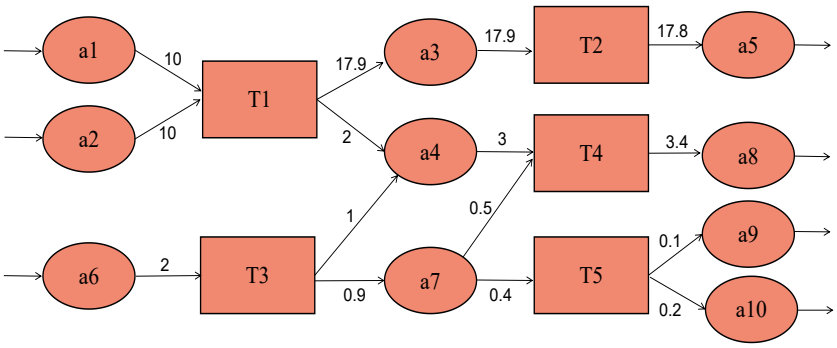


Fig. 2 Transaction model based on UTXO.

are four transaction modes: a single input and single output, single input and multiple outputs, multiple inputs and single output, multiple inputs and multiple outputs. In Fig. 2, coins from addresses a1 and a2 are combined and spent in T1 and further outputted to addresses a3 and a4. Then, address a3 participates in T2, and address a5 receives the input from transaction T2. That is, the number of all transactions comes from one or several previous transactions of the UTXO, and the coin flow in the UTXO-based blockchain changes accordingly.

A transaction may be accompanied by multiple sources. Although the transaction is traceable, it is unclear where the coins in the address come from because each transaction corresponds to N inputs and M outputs. For example, the coins in address a4 may come from a1, a2, or both; therefore, the number of coins in the address before and after the transaction cannot be mapped.

The total amount of coins inputted in any transaction is greater than the total amount of coins outputted. Before a transaction starts, it is necessary to check whether the balance is sufficient. If it is insufficient, then the transaction fails, and the balance is insufficient. However, the total input of transactions in the whole chain is equal to the total output. This is because when a transaction occurs, regardless of how many coins need to be paid to the other party, all coins should be transferred out, and surplus coins should be returned to the source address through the address called change address or in a direct manner. This process actually produces two unspent transactions. That is, coins received from the previous transaction must all be consumed in the next transaction. Any amount that is not sent to the output address or change address is regarded as the transaction fee, which is charged by the miner who created the block. The process of creating addresses is free and easy. To enhance anonymity, it is generally recommended to create new addresses for the change. Therefore, in the multi-output transaction mode, the first and only address in the chain may be the change address. More details about the change address will be described in Section 5.1.

The transaction chain in the unconsumed transaction output mode does not form a network and can be called forward-branching trees[8]. In recent years, this kind of blockchain system, represented by Bitcoin, has also been used for information hiding by embedding information into certain fields of transaction data or transmitting information through special transaction amounts. In Ref. [13], special transactions in the signature data were embedded by modifying the signature algorithm and transaction filtering mechanism in the blockchain, and the receiver hid the communication relationship of data transmission by detecting the signature data to filter the special transactions and realize hidden data transmission.

2.5.2 Based on the account

As the representative of Blockchain 2.0, Ethereum has adopted an account-based model. Bitcoin is based on the address, the system does not record how much money is in the account, and the balance displayed in the wallet is calculated, which is sometimes inconvenient in actual use. Thus, it is necessary to explain the source of the currency when transferring it. In fact, explaining the source is necessary only for saving money, all the money in the account needs to be spent when transferring, and change needs to use another address. Ethereum, which supports smart contracts, needs both parties in the contract to be clear and basically unchanged, especially for the contract account to be highly stable, so Ethereum adopts an account-based model.

As shown in Fig. 3, Ethereum has two types of accounts. One is an externally owned account controlled by a key that has an account balance and no code and can trigger a transaction (transfer money or execute a smart contract). The other is a contract account with an account

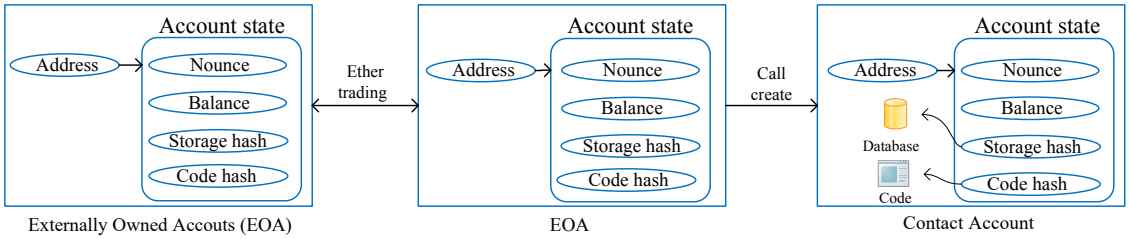


Fig. 3 Transaction model based on UTXO.

balance and code, which can be triggered to execute the smart contract code and run automatically after the creation of the smart contract and is controlled by the smart contract code. Each contract account is associated with an executable code and maintains status information. A transaction in Ethereum is a signed packet from one account to another that contains only one input and one output, unlike the scenario in Bitcoin.

3 Data Analysis Process from the Perspective of Complex Networks

Inspired by the data partitioning and sampling methods and techniques in the big data processing and analysis methods used in the literature[14], as shown in Fig. 4, we divided the work related to the research and analysis of blockchain transaction records into four processes from the perspective of networks, namely, data collection, network modeling, data analysis, and data presentation. This section introduces the mainstream public chain data acquisition methods as the primary task of data analysis.

The first step of data analysis is data acquisition and preprocessing, and subsequent work is equally important and necessary. Different from existing big data analysis methods, blockchain data analysis faces many challenges. Blockchain data are stored in a client in a heterogeneous and complex data structure, which cannot be directly analyzed. Usually, original data need to be parsed and stored in a comma-separated value table or database for analysis. As shown in Fig. 5, data sources in the existing work are mainly obtained from crawling the address and label existing on the web page, synchronized node data, using the Application Programming Interface

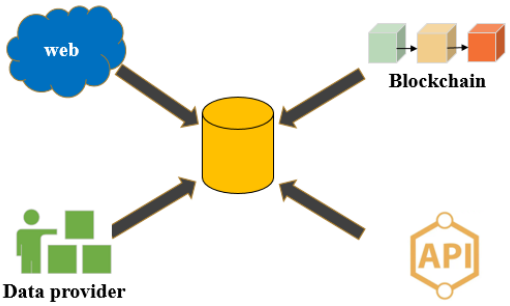


Fig. 5 Main data sources.

(API) provided by the blockchain browser or exchange, and using a well-processed dataset.

3.1 Data acquisition of Bitcoin

As the number one market capitalization, Bitcoin is the most successful application of blockchain technology, and its popularity has led more people to buy and sell Bitcoin as value investment options, producing considerable transaction data. Fortunately, compared with the traditional financial transaction information that is not available, people can download data on the Bitcoin blockchain for analysis to understand the economic ecology of the huge amount of cryptocurrency. Mining and analyzing cryptocurrency transaction data stored in the blockchain provide the means to study the transaction behavior in the complex economic environment. Currency transaction data are further studied in transaction behaviors in complex economic environments. Studying Bitcoin transaction data can also predict price movements and detect illegal behavior.

However, systematically exploring Bitcoin data is not easy because it involves a large number of heterogeneous data, which are generated and stored in different ways.

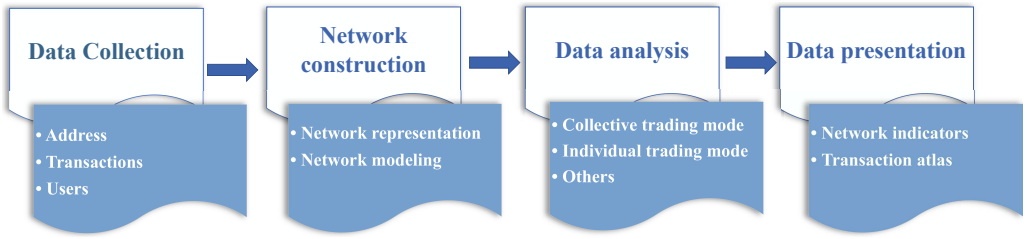


Fig. 4 Blockchain data analysis process.

Some existing work attempts to obtain data through different methods to support their research, which is mainly divided into the following types:

(a) Synchronization based on full-node Bitcoin data: This refers to downloading the installation official client (Bitcoin Core for Bitcoin) to access the area blockchain network and synchronizing data, producing historical transaction data. However, most raw data of a blockchain are stored in a special format, which requires further parsing. We can build a parser to extract transactions from the original data according to its data structure.

(b) Blockchain browser API. Some clients provide a JavaScript Object Notation Remote Procedure Call interface for users to obtain historical transaction data. For example, transaction data can be retrieved using the address, transaction, or block in the blockchain browser.

(c) In addition, some institutions or organizations provide well-processed datasets, such as the Bitcoin OTC and Bitcoin Alpha datasets for predicting illicit Bitcoin transactions; Google BigQuery that supports importing Bitcoin and Ethereum data; ChainAnalysis, a digital currency analytics company; and WalletExplorer that provides tag services.

(d) Collecting out-of-chain data based on a web crawler. Many addresses used for rewards will also be published on web pages or related forums, which can be obtained by writing crawler codes. Generally, this part of the data belongs to entities.

Different data acquisition methods have different advantages and disadvantages. Method a can obtain complete transaction data, which can be used to analyze the development trend of the entire transaction network but requires large memory to maintain increasing data, and data parsing is difficult. Method b can query basic information, such as blocks and transactions. This method is simple and requires the least investment, but its performance is limited by the API provider, and the data type is limited. This method is suitable for small-scale experiments and verification procedures. In Method c, after goods processing, data attributes are fixed, which is suitable for specific analysis works. However, it usually only contains data within a certain time and hence lacks continuity. The data provider will charge a part of the usage fee, expecting the affiliated institutions or organizations to regularly update the data.

Different from the other three methods, Method d is used as an auxiliary means to collect out-of-chain data, often used in address clustering or entity analysis classes.

In existing work, researchers typically use two or more data acquisition methods to make up for single-class methods and provide comprehensive and accurate data analysis results. Table 1 lists some ways in which a Bitcoin data source is sourced.

3.2 Data acquisition of Ethereum

Compared to Bitcoin data, Ethereum has more data types, such as block data, transaction information, smart contracts, execution tracking, and log data, such as accounts. Log data are the key to parsing ERC20 tokens and other smart contract data.

Considering different requirements, the following methods have been presented to obtain Ethereum data.

(a) Blockchain-based browser API. Similar to the Bitcoin blockchain browser, Etherscan and other Ethereum browsers provide an API to query blocks, transactions, and other basic information.

(b) Institutions or organizations providing well-processed datasets. Infura is a dedicated provider of Ethereum data, providing a rich and powerful Ethereum interface and querying a very comprehensive range of data types. GoogleBigQuery supports importing Bitcoin and Ethereum data, enabling researchers to analyze data online. Other notable companies include ChainAnalysis, a digital currency analytics company, and EtherscamDB, an open-source database that tracks all current Ethereum scams.

(c) Full-node data synchronization based on Ethereum. By running the Ethereum node locally, all Ethereum data can be obtained, and the original data can be parsed to obtain the required data format.

(d) Collecting out-of-chain data. This process refers to writing crawler codes to crawl the account address existing on the website of the blockchain browser.

Methods a and b are easy to use and have the least investment, but their performance is limited by API providers, and their data types are limited. These methods are suitable for small-scale experiments and verification procedures. Method c has the highest cost (node maintenance cost and parser cost), but it has the strongest scalability and can realize various types of data services as required. This method is also suitable for systems with high requirements on the data type and

Table 1 Common data sources in BTC data analysis.

Method	[15]	[16]	[17]	[18]	[19]	[20]	[21]	[22]	[23]	[24]	[25]	[26]	[27]
a			√	√	√	√	√		√			√	√
b	√	√						√					
c										√		√	
d				√	√	√	√	√	√		√		√

performance. Meanwhile, Method d is usually combined with other data collection methods as an auxiliary means. Table 2 lists the data sources of Ethereum in some works.

3.3 Existing datasets

To speed up the work progress, some well-processed datasets can be selected for experimental or effect comparison. Common and emerging datasets are shown in Table 3. Di Francesco Maesa et al.[24] and Poongodi et al.[40], compared the results of their methods with those in a website to prove that the method's effect is more comprehensive and accurate than that in the website. Google BigQuery[41] imports data from Ethereum and Bitcoin for online analysis. XBlock-EOS[42] and XBlock-ETH[43] come from the same laboratory, which is an open-source dataset framework for analyzing EOS.IO and Ethereum blockchains. There are well-processed EOS.IO and Ethereum datasets that can be used for further exploring EOS.IO and Ethereum blockchain systems.

Using well-processed datasets helps researchers to explore data at a certain time. In the dynamic analysis of transaction networks, Refs. [31, 45] used snapshots and progressive datasets to construct transaction networks of different scales to analyze network changes in the short term, drew conclusions, and extended them to the whole network.

4 Network Modeling, Attribute Analysis, Visualization Tools, and Analysis Methods

Transaction data should be presented as a network before graph analysis and graph mining. In this section, we first introduce the methods of representing transaction data as a network, then summarize the commonly used network metrics and analysis and visualization tools, and finally summarize the commonly used methods for blockchain data analysis.

Table 2 Common data sources in Ethereum data analysis.

Method	[28]	[29]	[30]	[31]	[32]	[33]	[34]	[35]	[36]	[37]	[38]	[39]	[40]
a		√			√	√	√			√			√
b	√		√	√					√				
c						√		√	√		√	√	
d											√		

4.1 Network representation and modeling

A transaction network can be expressed as $G = (V, E)$, where V represents a non-empty set of user vertices for transaction activities and $v_i, v_j \in V$. If there is a transaction relationship between v_i and v_j, then there is a connecting edge between them, which is denoted as e_{ij}, and E represents a transaction link set. In an actual transaction situation, transaction networks should are called a directed time-weighted graph $G = (V, E, W, T)$. Each pair of transactions is labeled as $e_{ij}(v_i, v_j, w, t)$. In a directed time-weighted graph, W represents the weight set of connecting edges between vertices, that is, the number of coins transferred between users, and can also represent the number of transactions of users in the considered time interval, showing the thickness of connecting lines in the transaction graph, and T represents the transaction time.

To build transaction data into a network, the transaction addresses of the sender and receiver are usually regarded as nodes, and the interaction between nodes is called an edge. Transaction data network construction based on the UTXO blockchain can be divided into three types: address networks, transaction networks, and user networks. After extending Fig. 1, in Fig. 6a, the address is taken as a node and the coin flow as a directed edge. In Fig. 6b, transactions are taken as nodes and directed edges as the flow direction of coins. In Fig. 6c, users or entities with one or more addresses are taken as nodes and the capital flow between them as a directed edge. Based on the characteristics of the transaction, the interaction between nodes can be divided into one-way and two-way interactions. Under the network science framework, the complex network

Table 3 Common and emerging datasets.

Name	Target	Ref.	Content or application
Google BigQuery	BTC&ETH	[41]	With Google BigQuery, Chrome users can read all the data stored in the Bitcoin and Ethereum blockchains and update them regularly.
XBlock-EOS	EOS	[42]	An open-source dataset framework for analyzing EOS.IO, including many types of EOS.IO datasets, is used to further explore the EOS.IO blockchain system.
XBlock-ETH	ETH	[43]	An open-source dataset framework used for analyzing Ethereum, which includes many types of Ethereum datasets.
WalletExplorer	BTC	[20, 44]	It provides address clustering grouping and wallet tag function, which is often used to compare experimental results.

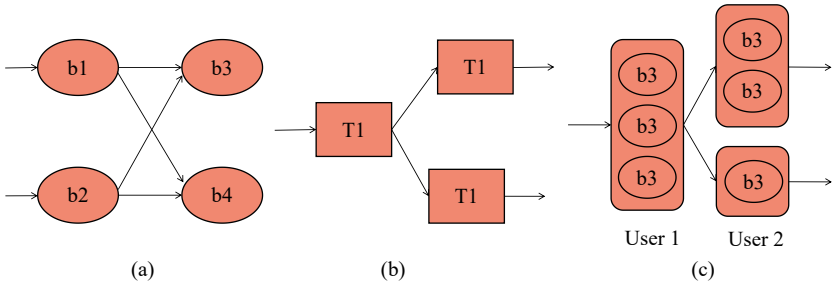


Fig. 6 Bitcoin transaction network modeling. Ellipses represent addresses, rectangular nodes represent transactions, and rounded rectangles represent user nodes.

composed of these interrelated nodes is modeled, and transaction networks are measured by the measurement method of network science.

At present, there is little research on graph analysis in blockchains, and existing studies[15, 24, 30] mainly focus on the static snapshot of graphs, ignoring the dynamic changes in transaction networks. In other studies[16, 31, 39, 45], the graph $G\{G1, G2, G3, \ldots, GT\}$ has been used to represent graph changes from time 1 to time T, which represents the cumulative modeling of transaction networks with different scales and complexities in the continuous time range.

4.2 Network properties

Complex network theory has been widely proven to be a powerful tool for modeling and characterizing various complex systems. This part summarizes the network attributes and related work in the research of cryptocurrency transaction networks according to the local and global network indices.

4.2.1 Local network properties

Local network attributes are used to express the local characteristics of a single vertex. These indicators mainly include the nodes, edges, degree distribution, node importance, and classification coefficient.

(a) Nodes and edges. The number of nodes and edges in a network can directly reflect the size and multiple identity attributes of a network. In Bitcoin, nodes usually represent addresses or entities, whereas edges represent the value transferred between two nodes. Using nodes and edges as the indices, Liang et al.[46] used the number of nodes and edges in the network to express the size of a network.

In an account-based blockchain network, such as Ethereum, an edge represents a variety of meanings. In addition to expressing the value transferred between two nodes, it can also express the number and time of transactions between two nodes, which also represents a contract invocation in a smart contract. In Ref. [31], the edges indicates the amount of the transaction account.

(b) Degree distribution. In a data structure, an undirected network can be represented by $N \times N$ symmetric adjacency matrix, where vertices are connected as 1 and disconnected as 0. By contrast, a directed network $N \times N$ can be expressed by an asymmetric adjacency matrix, in which the connection between vertices is marked as 1, and the rows of the matrix represent out-degree edges and the columns are in-degree edges. In an undirected graph, the degree of vertex i is the non-zero number in the column (row) where vertex i is located. In a directed graph, the degree of vertex i is the nonzero number in its row, and the degree of vertex i is the non-zero number in its column.

By calculating the degree of all nodes in the graph, let the degree of randomly selected nodes be k, and let N_k be the number of nodes with degree k. The degree distribution $P(k)$ of undirected networks can be calculated by the following formula:

$$P(k) = \frac{N_k}{N} \tag{1}$$

For a directed network, degree consists of out-degree and in-degree. The out-degree of node i refers to the number of edges pointing from node i to other nodes. The in-degree of node i refers to the number of edges pointing from other nodes to node i. Node j is the neighbor node of node i, and a_{ij} is a pair of adjacent nodes, which can be expressed by the elements of the adjacency matrix:

$$k_i^{\text{out}} = \sum_{j=1}^{N} a_{ij} \tag{2}$$

$$k_i^{\text{in}} = \sum_{j=1}^{N} a_{ji} \tag{3}$$

Assuming that the total nodes in the graph are N, because the degree of each node is at least 1 and at most $N-1$, then the degree distribution has the following relationship:

$$\sum_{k=1}^{N-1} P(k) = 1 \tag{4}$$

That is, the sum of degree distributions is 1.

The degree distribution index has been used by several studies to calculate the number of incoming and outgoing transactions, and it provides researchers with a macroscopic view of a transaction network. Guo et al.[39] used the Poisson model to observe the degree of nodes and found that most users tend to transfer ETH to a certain number of "close friends". They discovered that the degree distribution of transaction networks constructed monthly has an evident heavy-tail distribution. Hence, the degree distribution of most addresses is low, whereas the degree distribution of a small number of addresses is relatively high. Moreover, when the number of users reaches a certain number, the degree distribution approximately conforms to the power-law. In Ref. [31], the degree distribution of the cumulative network constructed by Ferretti and D'Angelo also complied with the heavy-tail phenomenon, which accords with the power-law distribution, and the power-law distribution was found to have time invariance with the expansion of the network scale.

(c) Node importance. In an undirected graph, four main indicators are used to evaluate the importance of nodes, which are mainly used to judge the nodes that play a pivotal role and to understand the influence of a node in the network. The most direct application of (i) degree distribution is to measure the importance of the node, that is, "degree centrality", which aims to describe the influence of the node. (ii) Intermediate centrality is an index that describes the importance of a node by the number of the shortest paths through a node. (iii) Near centrality refers to the average distance between node i and all nodes in the network. (iv) In the PageRank algorithm, according to the link relationship between nodes, the more times a node is linked, the more important it is.

Using the above method, Lee et al.[30] identified the most central vertices of the largest strongly connected components in the network they constructed, calculated the ranking of these vertices, and concluded that height nodes are very important in a network. In Ref. [47], the node importance calculation method was introduced and applied in detail, and the exchange, wallet service, and gambling industries were considered the most important subjects in Bitcoin ecology.

(d) Classification coefficient. The classification coefficient is an important method used to calculate the tendency of nodes in a network, which refers to the tendency that vertices in a graph attach to other vertices similar to them. Usually, it is described by the degree between nodes and is calculated by linking the Pearson correlation coefficient between node pairs. This value is between -1 and 1. If the classification coefficient is negative, then nodes of higher degrees tend to connect to nodes of lower degrees. For example, in an exchange, nodes with higher degrees tend to trade with nodes with lower degrees. Although there are many cryptocurrency exchanges at present that show a certain trend of centralization, the "rich club" phenomenon often does not appear in transaction networks; that is, there are few transaction links between large households.

4.2.2 Global network properties

Global network attributes are used to describe the entire network level. Such indexes mainly include the clustering coefficient, average shortest path length, diameter, connected graph, and community structure.

(a) Clustering coefficient. The clustering coefficient shows the trend that nodes in a graph gather together. Different from random networks, in real networks, a large number of nodes present an aggregation situation, and the clustering coefficient can be used to measure the transfer in the network. That is, if node i is connected to j and k, then the probability that j and k are also connected exists; that is, two connected nodes have the same neighbors. The clustering coefficient C_i of node i of the network mode d_i is defined as

$$C_i = \frac{e_i}{(d_i(d_i-1))/2} = \frac{2e_i}{d_i(d_i-1)} \tag{5}$$

where e_i is the actual number of edges between the d_i neighboring nodes of node i, that is, the number of triangles centered on node i, which can be recorded as $|\Delta|$. For an undirected graph, $|\Delta|$ represents the number of triangles with three nodes and three sides in graph G; $|\Lambda|$ indicates the number of triples of three nodes and two sides centered on node i, that is, the number of triples $d_i(d_i-1)/2$ in Eq. (7). Therefore, the clustering coefficient C_i of node i can be expressed as

$$C_i = \frac{|\Delta|}{|\Lambda|} \tag{6}$$

The average value of the local clustering coefficients of all vertices in the graph can be calculated to obtain the network average clustering coefficient C, that is

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{7}$$

Motamed and Bahrak[48] constructed and analyzed monthly transaction charts and cumulative monthly transaction charts and found that the clustering coefficient of Ethereum is high because account-based systems reuse addresses frequently and a person is likely to use a single address, thus making the graph dense and likely to be clustered. Baumann et al.[49] observed that a Bitcoin user network has a high average aggregation coefficient and typical "small-world" characteristics. Chen et al.[50] calculated that the clustering coefficient of top trader nodes is 0.192, which is much higher than that of the token transfer graph, which indicates that these top traders tend to gather.

(b) Average shortest path length and diameter. The path $P_{i\longrightarrow j}$ between nodes is defined as a sequence connected between the two. If there is a path $P_{i\longrightarrow j}$, then node i is said to have a path to j, that is

$$L = \frac{1}{\frac{1}{2}N(N-1)} \sum_{i \geqslant j} d_{ij} \tag{8}$$

where N is the number of nodes and d_i is the number of sides of the shortest path connecting the two nodes. For a path, the number of longest edges traveled by $P_{i\longrightarrow j}$, nodes i to j is called the network diameter. Ferretti and D'Angelo[31] showed the graduality of time and scale when constructing a network. They found that the wider the diameter of the network, the larger the nodes appearing in the network, which means that some nodes in the network are very active. We can also see the history of the development of large nodes with the development of time. Similarly, by calculating the average shortest path length and diameter of the network, Guo et al.[39] found that although the network diameter is rapidly increasing, the distance between users is very small, and the relatively low average shortest path indicates that the circulation period is very short, indicating that the Ethernet coin has good liquidity.

The average shortest path and network diameter can also be used to identify some phenomena. Di Francesco Maesa et al.[24] found that with the evolution of time, the network diameter diameter did not increase much but decreased, which proved the small-world hypothesis. In their other work on Bitcoin user graph analysis[51], they found that the average distance between nodes is short, but the graph shows a high diameter value. Through the study of outliers, they found that the outliers is the result of a specific abnormal transaction chain, which they classified as a Pseudo-Spam (PS) transaction.

(c) Connected graph. In an undirected graph, a connected graph is a subgraph in which any node can be reached from other nodes. In a directed graph, if you can reach another node j from node i, then the graph is connected. If there is a path from i to j and a path from j to i at the same time, then the connected graph is called a strongly connected graph. By replacing all directed edges of a directed graph with undirected edges, the resulting graph is called the base graph of the original graph. If the base graph of a digraph is a connected graph, then the digraph is weakly connected. The statistics and analysis of connected graphs can help us to understand the structure of the network. As a subgraph of the whole network, performing network analysis on a representative subgraph is more efficient than analyzing the whole network.

Guo et al.[39] found that a directed network presents a bow structure, and most nodes in a network are included in the strongly connected graph. Lee et al.[30] found in their research that the number of weakly connected graphs is smaller than that of strongly connected graphs and the size of maximum weakly connected graphs is similar to that of the original network. This is mainly because most vertex pairs have fewer bidirectional arcs, and the largest weakly connected components span important parts of the entire network (approximately 98% and 99% of vertices and arcs, respectively). In the study of contract networks, it is found that the strongly connected graph is the least of all other networks, which indicates that the connectivity in intelligent contracts is relatively strong, and a two-way arc exists between many pairs of intelligent contracts.

(d) Community structure. The community structure provides clues about what types of individuals or entities may be included in the network[52]. Only nodes with certain relationships can be included in a community because being close within the regiment and sparse between groups indicates that it is doomed to be closely related to people who can be close within the regiment. The network can be divided into several communities using a community detection algorithm. The k-means algorithm and Louvain algorithm are two different methods. k-means algorithm classification is accurate, whereas community division can only divide related addresses or accounts into the same cluster and cannot determine the type of cluster. However, the value of k using the k-means algorithm is too subjective. Although elbow technology can be used to optimize it, only the point that makes the rate of change of income lower at the elbow is selected. Therefore, the community

division method is better, and structural things should be dealt with in a structured way. If there is a network structure, it can be divided by communities. However, the community division method can only divide the nodes on a network into different communities according to the edge relationship. The community division method is very simple, and the imported data format is generally composed of three columns: source, target, and weight. Data are imported with DataFrame. After data importing, the imported data are constructed into a network with networks, and community division can be performed.

There is a lot of research on community division, which can be used to explore the relationship between addresses or between users and can also be used to identify entity identities in batches. Kumar et al.[53] collected Bitcoin data from 2016 to 2020 and observed that the triangles, transitivity, and clustering coefficient in the Bitcoin transaction network are higher than those in nonsocial networks, which indicates that entities in the network form a dense community trend. Zheng et al.[21] used heuristics to cluster addresses and the Louvain algorithm to determine the relationships between Bitcoin users. This work is an improvement on another clustering-related work in their team. In Ref. [54], three heuristic addresses of output transactions were combined, with multiple inputs and outputs; the address to clusters was changed; and the results were compared with those in WalletExplorer[55]. Zheng et al.[20] recently proposed an incremental clustering algorithm based on historical clustering that employed three heuristic algorithms, which can realize address clustering of large datasets. They built an address database related to historical transactions and implemented real-time updates. Comprehensive clustering results can be obtained by combining historical clustering and incremental clustering. Compared with the previous two works, this work implements the theoretical model and enriches the algorithm, and experiments were performed to verify each part of the model.

4.3 Analysis and visualization tools

Blockchain analysis tools provide researchers with a convenient way to effectively complete tasks, such as address tracking, building transaction networks, and knowledge discovery. Examples of these types of tools include BitConduite[17], a visual analysis tool used for the exploratory analysis of financial activities in Bitcoin networks, and BitIodine[18], an analysis tool that can identify and visualize those belonging to the same user address.

Based on different functions, we divided the existing tools into three categories: extract, transform, and load (ETL) tools, visual operation tools, data providers, and websites that provide full services in the cryptocurrency data analysis and illegal behavior detection. It covers commonly used tools for analyzing the relationship between addresses, analyzing the relationship between users, tracking currency flows, exploring user behaviors, and analyzing market effects and tools for providing blockchain data analysis websites and services.

The anonymity of a blockchain makes the transaction graph complex, so more reference information is needed to connect the wallet address with real users and perform different analyses. Block data are written to files on the disk, which makes data query time-consuming and, in turn, stimulates the development of blockchain data analysis tools.

The existing tools are mainly used to analyze user relations, user behaviors, wallets/markets, and transaction fees. BitConduite[17] is a visual tool that can identify Bitcoin network entities based on public addresses, focusing on identifying and describing individuals or groups of entities for the long-term exploratory analysis of meaningful activities.

BitIodine[18] can be used to build complex applications for forensic analysis against the Bitcoin blockchain. The framework parses the blockchain, clusters the addresses that may belong to the same users, classifies these users and marks them, operates through the front-end interface, and can also cluster and label the public addresses obtained from the Internet. However, it depends on Neo4j, which is not a universal graphic database designed for blockchain data, and its only additional properties make it inefficient for common blockchain analysis tasks.

BlockETL[56] is a Java development package used for data extraction, transformation, and loading in Bitcoin blockchain data analysis. It can be used to directly read the original block file and load the original block and transaction data into the SQL database. Similarly, in Ref. [16], a system was designed to make the process of blockchain data extraction and clustering scalable and to provide an SQL database that preserves transactions and address differences. The system meets the needs of clustering addresses in entities and stores the extracted data in the traditional database, which makes it possible to query the database for data analysis. More details can be found on the website[57].

BlockSci[58] is an open-source, extensible blockchain analysis system that supports a variety of analysis tasks. The tool integrates an in-memory analysis database, which makes it hundreds of times faster than its competitors and makes it more versatile, supporting multiple blockchains other than Bitcoin.

Blockchain.info[59] is a wallet that can also be used to provide Bitcoin blockchain data query services.

Chainalysis[60], which specializes in digital forensics and data analysis on the blockchains, has built a database to link cryptocurrency transactions to real-world activities to discover the purpose of anonymous transactions in the real world. Over the past five years, a large amount of money has been received from the US government to analyze blockchain network data and assess the risks associated with Bitcoin transactions. Currently, law enforcement agencies use their transaction analysis software in cybercrime investigations involving Bitcoin.

BlockAPI[61] is a blockchain data analysis framework that supports MySQL and other databases. This tool can combine out-of-chain information with in-chain data for a comprehensive analysis. The tools build data into a graph and store it in the database and then use the query language to analyze the constructed graph.

A common theme of existing blockchain analysis tools is to provide data to meet a range of analytical objectives to be delivered through different functions. In most cases, a complete analysis requires combining data on the chain with external data obtained from wikis or forums. Although existing tools have performed much development work in exploring and widely obtaining information about blockchain coding, most of the work focuses too much on Bitcoin and a small number of features. In the future, through the use of general real-time analysis tools, work can be more efficient and effective in addressing a wider range of blockchain data.

Analysis tools provide researchers with data analysis results with different characteristics to meet a series of analysis goals. However, the analysis operation needs to aggregate and associate the information from different sources, existing tools still use relatively single sources of data, and relatively single types of cryptocurrency are also analyzed. There is a lack of general blockchain data analysis tools, and the capability of combining data on the chain with data under the chain for a complete analysis should be addressed. In addition, these general tools should meet the ability of real-time transaction analysis at a minimum cost, which requires the availability of tools that obtain and process data from the chain and out of the chain regularly or even in real time. Table 4 shows a list of nonexhaustive tools classified according to different categories.

4.4 Common analytical methods

4.4.1 Graph mining and graph analysis

The rapid development of blockchain technology has aroused widespread concern, and data stored on chains are of great research value, especially the analysis of a large number of transaction data recorded in

Table 4 Analysis and visualization tools.

Category	Cryptocurrency	Function	Tools
ETL	BTC	Analysis of financial activities in a Bitcoin network	BitConduite[17]
	BTC	Address clustering, tagging, and visualization	BitIodine[18]
	BTC	Data extraction, conversion, and loading for Bitcoin blockchain data analysis	BlockETL[56]
	Address-based	Traversing address-based blockchain system	BlockSci[58]
	BTC, ETH	Construct a view of the blockchain and save it in a database;analyze the view using the query language of the database	BlockAPI[61]
Transaction network visual analytic tools	BTC	Visualization of Bitcoin transaction data flow	BitConeView[62]
	BTC	Explore the evolution of the Bitcoin market	BitExTract[63]
	BTC	Visualization and analysis of the Bitcoin network	BiVA[64]
	BTC	Uncover privacy issues with using Bitcoin in Tor hidden services	BlockTag[65]
Data providers and websites	BTC, ETH	Obtain all data stored in Bitcoin and Ethereum and update it regularly	Google BigQuery[41]
	ETH, EOS	Provide well-processed Ethereum and EOS.IO blockchain datasets	XBlock-ETH/EOS[42, 43]
	Various cryptocurrencies	Provide data query services including but not limited to the Bitcoin blockchain	Blockchain.info[59]
	Various cryptocurrencies	Provide data tracking service on the chain	Chainalysis[60]
	ETH	Use the API provided by Etherscan to access the Ethereum blockchain	EtherScan[66]
	BTC	Bitcoin network block query, grouping addresses, and labeling wallets	WalletExplorer[55]

blockchains. The existing research mainly constructs transaction data into address graphs or user graphs to analyze transaction data. Existing blockchain transaction graph analysis work can be divided into data mining tasks for graphic structure data and describing graphic attributes through some metrics. Regarding data mining for the graph structure of transaction data, some studies aim to use various network features to engage in deanonymization[25], illegal behavior detection[67], and cryptocurrency price prediction[40]. This kind of work usually requires analyzing and clustering the original data first, training the model through supervised learning or unsupervised learning or a combination of the two methods, and then carrying out the subsequent prediction work. For instance, in Ref. [68], after the process of obtaining and retrieving data is described, how to use two unsupervised machine learning algorithms to analyze the acquired common chain data was discussed, and some outliers and intelligent contracts were found. Then, they compared the effectiveness of this method by comparing the two machine learning methods and cross-linking them with public websites. As a new direction in the field of machine learning, deep learning can effectively learn the internal rules and representation levels of sample data. The same method can be used to train blockchain data to obtain data features.

Graph analysis refers to describing the attributes of a graph using some metrics to understand how the blockchain system and its transaction graph are formed and developed. That is, with the theoretical framework of network science, we measured the network characteristics of network graphs, particularly strongly connected subgraphs. These subgraphs often contain most nodes and can be used as a single data unit. They can also be used to aggregate connected nodes through the Louvain algorithm for morphological or mass verification. In an account-based blockchain, all transactions are one-to-one, which makes traditional graphical analysis tools easy. Of note, account-based blockchain systems, such as Ethereum and EOS.IO, need to extract internal transactions from ordinary transactions so that all relationships between addresses can be graphically modeled.

4.4.2 Time characteristic

A time series is a series of data points indexed in chronological order[69]. Based on the time characteristics of the data, the analysis of transaction data can be divided into static analysis or dynamic analysis. Static analysis refers to selecting the data before a fixed time point or a certain time to construct a graph, measuring the graph attributes, or mining the constructed graph structure from the perspective of a complex network.

Dynamic analysis can be divided into relatively real-time data update analysis and data analysis using segmented acquisition blocks to build a network. Relatively real-time data updating and analysis mean that data are regularly acquired from the chain, analyzed into readable data and stored in the database, and then retrieved and analyzed according to the requirements. For example, in Ref. [16], the ETL process was applied, which obtained 150 blocks of original data from online sources every time, stored the original data in a target database after the conversion stage, calculated the address cluster, and then transferred it to the SQL database for subsequent retrieval and analysis.

Data analysis using segmented acquisition blocks to build a network means selecting data groups with progressive periods to conduct experiments and observing the changing trend of transaction charts with time. Most research has been conducted in the form of time snapshots or cumulative networks, such as cumulative weekly networks or monthly networks. Kondor et al.[19] studied transaction data from the birth of Bitcoin to May 2013 and divided this stage into an initial stage and a transaction stage, showing the development process of Bitcoin ecology. Tasca et al.[23] extracted the network diagram of payment relationships, analyzed the transaction behavior in each business category, and used it to explore how the Bitcoin economy was formed. The author divides the bitcoin transaction from its birth to 2015 into three stages. The stages start with the early mining stage, then the second stage of growth, most of which are "evil" enterprises (i.e., gambling and black market), and finally the third stage, with exchanges as the leading factor. In Ref. [39], two datasets were selected, which represent early data and the latest data. The experimental results show that the degree and degree distribution of a transaction graph have the same pattern and the power-law is an invariable pattern of connectivity. Similarly, another study[31] obtained different networks with different sizes and complexities by changing the number of transaction blocks used to extract records and obtained network topology metrics with different results. It explored the dynamic structure of the Ethereum blockchain network from the perspective of complex networks and concluded that the wider the network diameter, the more large nodes in the network, which means that some nodes in the network are very active.

5 Application of Blockchain Data Analysis Based on a Complex Network

Complex network methods are widely used in address clustering and entity recognition, transaction pattern recognition, privacy disclosure risk analysis, and illegal behavior detection and analysis. Moreover, they belong to a subdivided research field included in the blockchain data analysis. In general, we can compare the same or different static or dynamic cryptocurrency transaction networks within a certain time range, apply the measurement method of a complex network to obtain the static or dynamic changing transaction network, perform a graph analysis, and obtain knowledge discovery to better understand the cryptocurrency trading and market effect represented by the network. Moreover, it can also perform graph mining on the cryptocurrency trading network to identify transaction patterns and detect illegal acts. Researchers can select some methods for in-depth research, but there is a certain order between directions.

5.1 Address clustering and entity recognition

Blockchain has two distinct characteristics: openness and pseudoanonymity. Openness allows researchers to download data from the chain for analysis, whereas pseudoanonymity not only protects users' privacy but also enhances the difficulty of data analysts. Therefore, as the premise of privacy leakage risk analysis and illegal behavior detection, the first job of blockchain data analysis is to cluster addresses and identify entities. Entities can be users or organizations, and users and organizations can also own one or more entities. In clustering transaction data, each transaction data point corresponds to one or more addresses, and these addresses are connected one by one. Gathering them together and putting them in a container is an entity. The addresses in this container can be as few as one, and the top is not capped. The process of identifying the container name is called entity identification. Reclustering close users (entities), finding the relationship between users, analyzing what this group of people is doing, and improving the accuracy of identity recognition have a certain enlightening significance for future research on user portraits, transaction pattern recognition, and illegal behavior detection.

The existing address clustering and entity recognition can be roughly divided into four types: heuristic method, behavior-based method, out-of-chain auxiliary information method, and visual analysis method.

5.1.1 Heuristic methods

Heuristic methods need to be divided into UTXO-based blockchain and account-based blockchain.

(a) UTXO-based blockchain. (Multi-input multi-output). Spagnuolo et al.[18] noted that if there are multiple input addresses in a transaction, it can be inferred that they are owned by one user without using multiple signals because both private keys need to exist to sign the transaction as valid. When there are two addresses at the output end of the transaction and one of them is the address first discovered in the blockchain, if the unused Bitcoin delivered to the address is delivered to the user again at the input end, then the address is derived from one address. As shown in Fig. 2, a1 and a2 are both input terminals of transaction T1, so a1 and a2 belong to the same entity. If a4 and a7 are both input terminals in transaction T4, then a4 and a7 are controlled by the same entity. Because there are many limited conditions in the common input method, it is easy to cause two kinds of errors, i.e., underestimation error and overestimation error[70]. Therefore, we cannot find all or as many addresses as possible by this heuristic method. For example, in a coinbase transaction, this transaction mode has only output but no input, so if only multi-input multi-output is considered, this kind of address cannot be found.

In recent years, the combination of the multi-input multi-output and change (shadow) address method has been adopted in the literature. When there are two addresses on the output side of the transaction and one of them is the address first discovered on the blockchain, if the unused Bitcoins delivered to that address are delivered to the user again on the input side, then it can be inferred that the address is derived from an address, and the newly derived address is often the change address. In the Bitcoin network, the output of a transaction is used as the input for another transaction, and if the input is greater than the new transaction output, then the client generates a new Bitcoin address and sends the difference back to that address[44]. Meiklejohn et al.[71] used this information and the above two methods to identify the major players and entities on the Bitcoin blockchain network.

However, Bitcoin exchanges using mixed-currency services do not satisfy this method. In a mixed-currency transaction, n input addresses come from different users[4]. At present, there are many versions of mixed-currency services, and the proportion of mixed currency is gradually increasing. Moreover, the more sensitive

the transaction is, the higher the probability of using the mixed-currency service is, and the worse the accuracy of this method is. In addition, shadow addresses are an unofficial concept; that is, there is no concept of change addresses in the format of Bitcoin transactions, which are defined spontaneously by users in the process of use. For example, the new or often fixed address in the output address is regarded as a change output, or the output amount is observed. That is, according to transaction habits, the decimal place is more or less a change address. However, the amount of the change address is not necessarily relatively small, nor is it necessarily integer. Therefore, the conditions for distinguishing the change address are not reliable and lack theoretical analysis.

To improve the above heuristic method, in Ref. [44], the output transaction heuristic based on the above conditions was added, the accuracy and comprehensibility of the method were verified using self-controlled addresses several times for comparative experimental analysis, and the influence of the number of iterations on the clustering effect was analyzed. The results show that the greater the iteration times of the clustering program, the more comprehensive the data that will be obtained. By contrast, the time will also increase in a linear trend. As further work, the author-related team personnel in the literature[20, 21] will provide a heuristic method combined with the Louvain algorithm and put forward clustering based on the history of the incremental clustering algorithm. This algorithm can achieve large datasets of the address of the cluster, build a historical transaction record-related address database, and realize real-time updating. It can relate relationships between Bitcoin addresses and between users. A recent paper[72] also proposed an incremental clustering method for block wallet address data using the algorithm of union-find disjoint sets. By introducing the lookup address index table to extract the relationship between clustering entities, a new Bitcoin entity relationship can be obtained, and then the entity type can be inferred. This method takes into account the result fusion problem after the emergence of new data and improves the efficiency of the algorithm. At the same time, the entity is identified and labeled, and the visual analysis of entity transaction behavior is realized using visual tools.

(b) Account-based blockchain. Most of the existing heuristic methods are based on the UTXO model of the Bitcoin blockchain, which has N inputs and M outputs. However, in account-based blockchain systems, such as Ethereum and EOS.IO, there is usually only one input and one output, so the method based on multi-input and multi-output address heuristics is not suitable for this case. Moreover, different from the transaction chain under the UTXO output mode, the transaction network is formed by all parties of the blockchain system node based on the account model. Possible problems, such as fraud, wealth distribution, and Ponzi schemes in intelligent contracts, can be obtained by mining and analyzing the hidden information in the transaction networks. Therefore, address clustering on the blockchain based on the account model is urgent to identify entities.

As in an early article on the heuristic method of clustering algorithms for the Ethereum account model, Victor[73] compared three heuristic methods of deposit address reuse, airdrop multiparty participation, and self-authorization and concluded that most addresses can be obtained via deposit address reuse. At the same time, another study[28] applied the clustering algorithm in machine learning to the analysis of Ethereum. By filtering data collected in Ethereum and clustering account nodes contained in it, the node embedding algorithm was used to quantify the external account nodes and intelligent contract account nodes in Ethereum. Based on the feature vector, the external account nodes and intelligent contract account nodes were clustered into groups, and the results were visualized. Then, based on the observation, a new method of malicious user detection based on clustering results and known nodes is proposed.

5.1.2 Behavior-based methods

In transaction networks, we can observe the user's behavior habits to discover the user's characteristics, such as transaction time, amount, and frequency. Monaco[74] identified and verified account holders by observing transaction time intervals. The function of this method is to portray the identity of a user in the blockchain, which does not belong to the same category as the method mentioned above. The aforementioned method classifies anonymous blockchain transactions and then decomposes them into datasets or clusters belonging to different users. This method is essentially the practice of traditional identity image technology in the field of blockchain. Its implementation premise is to find the "long-term transaction data" of specific users. In the blockchain system, users are suggested to adopt a one-time address strategy to improve anonymity so that

their transaction data will be scattered to many addresses and "long-term transaction data" of specific users will not be easily obtained. Nonetheless, only a few studies on this method have been performed.

Ron and Shamir[70] analyzed data based on transaction entities by describing the statistical attributes of a Bitcoin transaction graph. The results of transaction characteristics show a large number of dormant Bitcoins that have not participated in any expenditure and a large number of transactions coming from the same transactions in 2010.

Although the address may be controlled by a certain entity, it is difficult to avoid error aggregation in the process of clustering, and it is difficult to perform quantitative inspections. Therefore, many researchers use pair or multiple combinations of the above methods to perform address clustering. At present, the data stored in a blockchain are still dominated by transaction data. In the Blockchain 3.0 stage, the system will generate massive data from different sources to meet the needs of traceability of deposits and certificates, and the distributed system is faced with difficult monitoring. Effective data analysis and prediction of system behavior are the keys to optimizing the design of blockchain mechanisms. In recent years, the application of machine learning to blockchain data analysis can make the two methods work together efficiently. Machine learning can simplify the data verification process and identify malicious attacks and dishonest transactions in the blockchain. In Ref. [75], the application of machine learning in blockchains is summarized, which has great application prospects in illegal behavior detection and cryptocurrency price prediction, among others. For example, using machine learning methods to collect addresses leaked or voluntarily displayed by users from the network (to accept donations or rewards) or addresses leaked by exchanges, these data are used as a training set to find the best separation space in the given data and classify them. The mapping relationship obtained in the first step is used to constantly determine new node mapping relationships, and the new node is classified into the identified type according to the characteristics of the analyzed new node.

5.1.3 Out-of-chain auxiliary information

Out-of-chain data refer to blockchain-related data that are not stored on the blockchain. Similar to out-of-chain data collection, it can be used as an auxiliary means for network analysis, such as deanonymization. These data include but are not limited to addresses posted on web pages or related forums for reward, IP addresses of nodes, data leaked by exchanges, and websites that provide tagging services. Crawler codes can be written to obtain these data, and such data can generally find the entity to which it belongs.

Zhu et al.[76] introduced a system that analyzes Bitcoin and tracks the transaction flow of Bitcoin by comparing blockchain data and network traffic data. Blockchain data store all transactions related to Bitcoin addresses, whereas network traffic data reflect transactions related to IP addresses. Then, the system tries to anonymize users' Bitcoin identity by associating the Bitcoin address with the IP address. Finally, through experiments, it is proven that the deanonymity method is effective.

Bres et al.[77] identified and applied several identifiers derived from address reuse to analyze Ethereum and de-anonymize. However, the authors only applied the on chain data to de-anonymize Ethereum users. Later, if combined with out-of-chain data, more identifiers may be obtained to analyze Ethereum user data.

In Ref. [78], another way was proposed, i.e., using transaction records in a wallet address leaked by Mt. Gox to match the transactions recorded in the blockchain. When the ID of the wallet of the user and the address inside the wallet are known, the Bitcoin revenue and expenditure transaction network generated by these addresses is under control. That is, the deanonymization of the Bitcoin revenue and expenditure transaction network is realized. This method can be called a mapping-based approach. Biryukov et al.[79] analyzed the collected data, classified different transaction relay patterns, and designed a heuristic that uses summary statistics in transaction data to derive and evaluate Bitcoin address-to-IP mapping to assume the transaction owner. This method belongs to the Bitcoin address traceability of the network layer and has high requirements for hardware resources and a high threshold for research.

5.1.4 Visual analysis

Some system tools for transaction visualization can also be used to deeply study Bitcoin data and can perform a systematic, long-term, and dynamic visual analysis and explore Bitcoin transaction activities. Such tools include BitConduite[17], BitIodine[18], BitConeView[62], BitExTract[63], and other tools mentioned in this article. Other researchers, such as Kairam et al.[80], developed a system called Refinery under their proposed

design standard for a visualization system that supports associated browsing.

Thus, the purpose of entity analysis and address clustering is to explore the relationship between UTXO-based blockchain addresses, such as Bitcoin, to construct a user map through the address transaction graph for further user privacy disclosure risk analysis and network description, and to achieve network visualization and fully mine the hidden value in the blockchain data. Essentially, the selection of a clustering algorithm needs to consider three principles: the first is accuracy, that is, the address obtained by the analysis really belongs to the control of the same user group, and the second is the comprehensiveness of the result. Hence, to find multiple addresses controlled by the same user group as much as possible, it is necessary to cover all the addresses in the clustering analysis. The third is real-time, long-term, and stable automatic division. Here, the clustering process should be real time, dynamic, and stable as far as possible and should be clustered automatically, which requires training with known results and comparison of experimental results.

5.2 Transaction pattern recognition

Different from the traditional payment model, people's payment behavior becomes direct under the decentralized system operation mode. However, identifying the characteristics of human payment behavior caused by anonymity is an interesting subject. In addition, anonymity leads to many illegal behaviors based on Bitcoin, such as money laundering and fraud. Hence, whether we can identify special patterns from the transaction records of blockchain and find related illegal behaviors is a valuable question. The key to solving these problems is to identify and analyze the transaction patterns in Bitcoin.

Many methods are used to identify transaction patterns, and common methods can be divided into visualization methods and dynamic analysis methods. High-fidelity visualizations, such as those demonstrated by McGinn et al.[81], can detect high-frequency transaction patterns, including automated money laundering operations, and the evolution of multiple different algorithmic denial-of-service attacks on the Bitcoin network. Turner et al.[82] created an analytic framework—the Ransomware–Bitcoin Intelligence–Forensic Continuum framework—to search for transaction patterns in the blockchain records from actual ransomware attacks. This work is similar to Ref. [44], which uses known data for testing. Zola et al.[83] revealed changes in entity behavior through the time analysis of blockchain data and compared whether some transaction patterns are repeated in different batches of Bitcoin transaction data. Based on deep learning techniques that can be used for pattern recognition, Guo and Zhang[84] proposed sparse deep NMF models to analyze complex data for accurate classification and efficient feature interpretation.

The study on cryptocurrency transaction networks can understand the evolution of the cryptocurrency ecosystem, grasp the economic rules of the cryptocurrency market, explore the network characteristics of cryptocurrency transaction networks, and reveal some special transaction structures. This section summarizes and analyzes the knowledge discovery of transaction patterns in the graph analysis and graph mining of cryptocurrency transaction networks in the existing work from the perspective of complex networks.

5.2.1 Small world

Considering the average shortest path length and clustering coefficient described above, we can judge whether the network has a small-world phenomenon. In the small-world phenomenon, the network often has a small average shortest path length and high average clustering coefficient. As a comparison standard, it is necessary to generate random graphs with the same number of nodes and edges.

Particularly, the small-world phenomenon can only occur when the network scale reaches a certain level. By constructing a network of increasing size, Ferretti and D'Angelo[31] found that the clustering coefficient was 0 when fewer transactions contained in blocks were selected to construct a network, and there was no small-world phenomenon in the network at this time. When the number of blocks gradually increases, the clustering coefficient of the obtained network becomes larger, and correspondingly, the average shortest path length of the network becomes smaller, and its value conforms to the conclusion of "six degrees of separation" in small-world networks[85].

Based on Ref. [31], we also find that with the passage of time and the expansion of the network scale, the average shortest path length changes from small to large, and then from large to small. This is because, with the popularity of cryptocurrency and national support, blockchain technology has attracted a large number of

new users to join the blockchain network and created several new nodes. These nodes have less participation, and no close contact has been made. In the later stage, the reduction of the average shortest distance of the network may be caused by the newly joined nodes participating in exchange transactions or services, such as wallets[86].

Watt used a method of randomly reconnecting edges to explore the middle zone between the regular network and random network[87]. When the probability of selecting connected edges is $P<0.01$, there are many long connected edges in the regular network, which reduces the distance between the nodes at both ends of the long edges and the distance between their neighbors. In this case, random reconnection has little influence on the clustering coefficient of the network but has a great influence on the average shortest path length of the network, which makes the network transform into a small-world network. This kind of "long edge" can be called a "shortcut". The existence of a small-world network accelerates the ability of information transmission between nodes and has the effect of improving efficiency and reducing cost. Hence, changing a small number of nodes in the network can improve the performance.

For more work on verifying the small-world phenomenon, please refer to Refs. [20, 55, 88, 89].

5.2.2 Assortativity

Assortativity coefficient is an important method used to calculate the tendency of nodes in a network. Homogeneity is used to examine whether nodes with similar values tend to connect with one another, that is, to show whether nodes tend to communicate with other nodes that are similar to them. In some cases, nodes are neutral when communicating with other nodes.

Guo et al.[39] calculated the negative classification coefficient of a relevant dataset, indicating that although there are many cryptocurrency exchanges showing a certain trend of centralization, the transaction networks often do not have the "rich club" phenomenon.

Liang et al.[46] studied three types of cryptocurrencies, i.e., BTC, ETH, and Namecoin, in their work, and they found that Bitcoin's and Ethereum's early transaction networks were mismatched, but it was difficult for Namecoin to judge. In the latest cryptocurrency transaction networks, for Bitcoin and Ethereum networks, the function is monotonously decreasing using the K-nearest neighbor algorithm, which shows that nodes with high degrees can easily connect with low nodes, so the network is mismatched. For Namecoin networks, the curve is almost constant, which means that the degree of connected nodes does not depend on each other, so Namecoin networks are matched.

Zhao et al.[88] calculated that the classification coefficient of the coin exchange chart and contract authorization chart is negative, indicating that in coin exchange activities and contract authorization activities, height nodes tend to be connected to small nodes. Motamed and Bahrak[48] calculated the degree matching of the transaction graph to observe whether the account traded with the counterparty. By observing the transaction charts of different periods, it is concluded that for all emerging currencies, their homogeneity is negative.

5.2.3 Connectivity of transaction networks

From the perspective of the network structure, we first pay attention to whether the nodes in a network are connected, that is, the connectivity of the network. In undirected networks, the description of many network topology properties also depends on the connectivity of the network. For all the networks based on node user relations in the cryptocurrency network world, although we cannot give complete structural data, we can still intuitively judge that such a huge network should not be connected because if there is a node user who has not made a transaction, then the whole network is disconnected. From this point of view, connectivity is a very fragile property. Hence, large-scale networks, such as cryptocurrency transaction networks, are unconnected, but they will form connected slices or connected components with a large number of node users.

In the connected graph introduced earlier, the number of weakly connected graphs in the transaction networks is smaller than that of strongly connected graphs, and the size of the maximum weakly connected graph is similar to that of the original network, spanning an important part of the whole network (approximately 98% and 99% vertices and arcs, respectively), forming a blockbuster, and the blockbuster in the network is always unique. Because if there is a connection between two nodes in two sub-blockbusters, a large blockbuster containing more nodes will be formed.

Many studies have analyzed connected graphs rather than the entire network. As a subgraph of the whole network, the connected blockbuster (maximum weak

connection graph) contains almost all the nodes in the whole network. Thus, it would be more efficient to perform network analysis on a representative subgraph than to analyze the entire network.

There is a huge connected component in the directed network, which contains the vast majority of nodes of the whole network. In a directed graph, the structure of components becomes complex. Guo et al.[39] found that the component structure follows a bow structure.

(1) Strongly connected kernel (SCC): it is also known as a strongly connected blockbuster. As mentioned earlier, the strongly connected graph SCC in the network is the largest subgraph, so every pair of nodes in SCC is strongly connected, with paths $P_{i\longrightarrow j}$ and $P_{j\longrightarrow i}$ graphs.

(2) In component: This includes nodes that can reach the SCC through the tube but cannot be reversed.

(3) Out component: This includes nodes that can be reached from the SCC through a tube but cannot be reversed.

(4) Tendrils: These include nodes that cannot be reached from the SCC, cannot enter the SCC from the outside or are disconnected from large connectivity components.

There is also a component called the tube, which can be thought of as a bridge from the IN component to the OUT component.

5.2.4 Power-law distribution

In recent years, degree distribution, as an important topological feature of the network, has played an important role in network science research[52]. The power-law distribution is the distribution of node value in the network. The power-law distribution of many real networks can be well expressed by double logarithm $P(k) \sim k^{-y}$, where $y > 0$ is a power index, usually between 2 and 3. We can verify the existence of the proportional constant C and the power exponent y, such that there is approximately $P(k) \sim Ck^{-y}$, and the logarithm on both sides is $\ln p(k) = \ln C - y \ln k$. According to the given data, if we see an approximately straight line in the double logarithmic coordinate system, then the processed data approximately conforms to the power-law, and the corresponding power index can be obtained from the slope of the straight line. To reduce the error, the least square straight line fitting method is usually used. Moreover, only when the value is large, the distribution is approximate to the power-law form, and then the tail accords with the power-law distribution.

Ferretti and D'Angelo[31] listed linear scales to show the degree distribution. To better understand the degree distribution of complex networks and reflect the indicators of network authenticity, Guo et al.[39] conducted a Likelihood Ratio (LR) test; LR shows that the logarithmic power-law model is very suitable for observations because the number of nodes with large degrees is relatively small and will not significantly affect the fitting[39]. In a logarithmic graph, the power-law distribution model provides a reasonable fitting for observation in a statistical sense.

With the increase in the amount of data, the shape of a logarithmic graph accords with the power-law distribution model, although the coordinate axis index is changing. However, the out-degree, in-degree, and degree distribution of the transaction dataset still conform to the power-law distribution. To obtain an abundant node representation, Lin et al.[32] used a graph representation algorithm based on a random walk to represent the node characteristics of large networks in a low-dimensional space for graph analysis and mining. They found that the degree distribution of the whole graph and its subgraph of Ethereum conform to the power-law distribution, whereas the graph representation method based on a random walk can effectively keep the structural characteristics of the power-law distribution graph.

Lischke and Fabian[47] summarized and compared the development history of the Bitcoin system in the first four years in terms of time, transaction, and country and found that the Bitcoin network follows the power-law distribution, although not the whole network scope. Somin et al.[90] regard a blockchain as a social network. They analyzed the network properties of the ERC20 protocol's cryptocurrency transaction data and proved that the network shows strong power-law properties, consistent with the expectations of the current network theory.

5.2.5 Matthew effect

When judging whether the network has power-law distribution characteristics, priority connection is a common way in complex network analysis. The priority connection can further capture the dynamic change activities of entities by various network characteristics, explore the dynamic change law of the entity behavior, and discuss whether the flow direction of coins among users conforms to the economic law. In economics, the priority connection is also called the "Matthew effect", the phenomenon that the rich get richer.

Kondor et al.[19] analyzed the evolution of a network and the dynamic process occurring on the network (i.e., the flow and accumulation of Bitcoin). The evolution of basic network characteristics, such as degree distribution, degree correlation, and clustering with time, was also studied. They found that the wealth balance of wealthy users increased faster than that of ordinary users, and a positive correlation exists between wealth and node degree.

Di Francesco Maesa et al.[51] verified the rich hypothesis and measured the concentration of the rich from the perspective of balance and connectivity. They thought that a user is richer compared with other users in the network, and his balance and the number of input transactions are very high. The user at time t is richer than the user at time $t+1$, and the richest user at time t is still the richest at time $t+1$. As time goes by, wealth becomes increasingly concentrated. In the following work, there are conclusions that users with a high degree value have a high value in other centrality measures and high economic weight in the ecosystem (i.e., they belong to the richest users)[24, 91, 92].

5.2.6 Peeling chain

In the peeling process, the change address always sends a small number of coins to other addresses as the input of the next transaction and then generates the change address, thus forming a long chain structure. The appearance of peeling chain transactions is often used as a cover for certain behaviors. For example, money laundering that disturbs the line of sight is achieved by generating multiple transactions, a pseudo-spam is generated, airdrop or advertising is performed using a peeling chain, and anonymity is enhanced.

By studying the small-world phenomenon, Di Francesco Maesa et al.[24] found that the network diameter did not increase much with the evolution of time. This feature of the Bitcoin user graph is because some transactions are also used for merging and splitting user funds, not just for payment, such as the peeling chain for splitting funds. As further analysis of this typical event, they detected an abnormal transaction chain by analyzing abnormal values in the degree distribution in the later work. The following phenomena were observed: the input is 1; each output is greater than or equal to 2; the amount transferred to the output address is very small, approximately 0.000 01 BTC; and the remaining large amount of balance is transferred to another address, then transferred out from the address where the balance was received last time during the next transfer, and then circulated in turn until the money is dispersed. This transaction chain is called a pseudo-spam distribution[51]. They also pointed out the conjecture of the reasons behind the transaction mode of "pseudo-screen transaction", that is, anonymous attack and advertising[91]. Similar work has been done in the literature[93].

In the early analysis of Bitcoin, Ron and Shamir[70] analyzed Bitcoin transactions before May 2012 and found a large number of small transactions in the network, but some of them transferred a large number of funds. The manipulator behind it uses the change address of each transaction to disperse the coins and then concentrates the coins scattered in many addresses into one account.

Reyes-Macedo et al.[94] focused on analyzing the transactions derived from cybercrime (ransomware attacks) or activities involving the use of cryptocurrencies based on public chains. They found that cybercriminals use stripping chains in Bitcoin systems and hide their traces by means of currency exchange companies or converting the collected digital currency into other kinds of digital currency.

Meiklejohn et al.[71] explored various alternative strategies developed by thieves to hide the source of stolen Bitcoins. To prove the effectiveness of using change address heuristics to track the flow of coins, they tracked illegally acquired Bitcoin flows and discovered some flows of Bitcoin directly from stolen to exchanges or other known institutions, providing an opportunity for institutions with subpoena power to know whose account was deposited, and potentially to know the identity of the thief.

5.2.7 Price manipulation

Chen et al.[67] used data leaked by exchanges to build a payment network, mining the exchange transaction network to identify possible manipulation patterns and verify whether there is manipulation in the market. Analysis results show that there are many abnormal transaction patterns between abnormal accounts (e.g., self-circulation, two-way, and star). This result shows that these accounts are controlled by the same group and are strong evidence of price manipulation. Based on Ref. [70], many addresses did not transfer out transactions after receiving Bitcoins, there were a large number of small transactions that only moved a small part of a single Bitcoin, but there were also hundreds

of transactions with the amount exceeding 50 000 BTC. It also mentioned that many large active entities either had the largest amount of external transfers or the largest number of transactions. By building a transaction graph, we can notice that many subgraphs contain these large transactions and their neighborhoods that try to hide the existence and relationship structure between these transactions.

The statistics of main knowledge discovery categories from the perspective of complex network are shown in Table 5.

5.3 Privacy disclosure risk analysis

The purpose of an entity analysis is to find the address owned by an entity and identify the user's true identity.

This is a very meaningful but very difficult thing, which will make the analysis work difficult to break through on a large scale. The existing work is represented as a type of user by tabling or by some descriptive features, without identifying the true identity of each user. However, if the user's wallet address is leaked, such as the Mt. Gox event, then you can compare the transaction amount recorded in the wallet address with the transaction amount recorded at the same time and find all the wallets owning all addresses[78]. Although the only matching address is less than one-third of the total match, these data are enough for further analysis.

Based on studies, de-anonymization in blockchains has caused the risk of privacy leaks to a certain extent. The privacy leak risk on blockchains is derived from the leakage of the transaction behavior and the disclosure of the user's identity. The disclosure of the transaction behavior refers to the inference of individual behavior or certain financial activities from key elements, such as the addresses of both parties in the transaction, transaction time, transaction remarks, and transaction contents (if used for payment) during the transaction. The leakage of a transaction behavior may be proactive, such as a web platform for receiving a rewarded Bitcoin address and working by writing a reptile to climb the Bitcoin address existing as part of the dataset[98]. However, in most cases, the transaction behaviors are passive, and analyzing and mining the vast amount of transaction data on the blockchain are a big attractive for investors.

The analysis of transaction behaviors can obtain the profit and loss status of users in a certain period of time. That is, the user's buying and selling amount at a certain time is compared with the price at that time. As mentioned in Ref. [74], we can infer the type of investor belonging to a transaction, whether it is an individual or an exchange, based on the approximate amount, transaction frequency, and transaction time of the trader and whether the individual trader is a common investment. Further analysis of the transaction amount and transaction frequency can determine abnormal capital flows, and the analysis of the type of traders can speculate whether they are long-term holders or frequent traders or in the middle. Based on the characteristics of UTXO, the analysis of the transaction behavior of Bitcoin traders can calculate their current or historical positions, analyze whether they are holding coins or transactions normally, and infer whether they are retail investors or large individual investors or exchanges based on the size of their positions.

The disclosure of user identity refers to the matching of the user's identity information (e.g., ID number, bank card number, and mobile phone number when registering the exchange account) and address with the transaction information in the blockchain, which is also the ideal goal for entity analysis. Di Luzio et al.[99] explored the possibility of information disclosure caused by using ripple network payment, indicating that there is a risk of personal information disclosure in the process of payment with cryptocurrency.

As a confrontation between data analysis on the chain and data privacy protection on the chain, a new way of information hiding has emerged. The hidden information is transmitted in the form of a transaction, in which there are signs of transmission frequency, transaction time, and amount. Accordingly, two or more addresses

Table 5 Incomplete statistics of knowledge discovery in the literature research.

Knowledge discovery	Cryptocurrencies	Reference
Small world	BTC & ETH & Namecoin	[20, 24, 30, 31, 47, 51, 56, 88, 89]
Assortativity	BTC & ETH & Namecoin & LTC	[29, 30, 39, 46, 48, 88]
Connectivity	BTC & ETH	[24, 39, 51, 89, 95]
Power-law	BTC & ETH	[24, 30–32, 39, 47, 89, 95]
Matthew Effect	BTC & ETH	[19, 24, 51, 91, 92]
Peeling chain	BTC	[24, 51, 70, 71, 93, 94, 96]
Price manipulation	BTC	[67, 70, 97]

will be found for communication within the appropriate analysis time window. Given the transaction amount and other reasons, the handling fee may be higher than the transmission amount. One possible reason for this kind of transaction is that it is used to test the network. If it is too frequent and relatively fixed, then it can be called "spam". If it is judged whether it is a transaction, such as intelligence, it can be considered a "secret information transmission" according to prior knowledge. Suspicious transactions can be mainly judged by the transaction amount, handling fee, transaction time, frequency, and outgoing and incoming conditions of both parties in the transaction field, among others. Further exploration may require clustering them, using the Louvain algorithm to analyze the relationship between users and addresses, and then using Gephi to visually display them for future tracking and supervision. Concealed communication[100] is a further information privacy protection strategy, which hides not only the transaction information but also the transaction behavior and enhances the security of the transaction behavior. In the final analysis, users should understand the privacy protection strategy, have a good understanding of privacy leakage problems that may be involved in Ethereum, and use related services, such as privacy protection services, to reduce the risk of privacy leakage through a simple analysis.

5.4 Illegal behavior detection

In a blockchain network, to enhance privacy, people often use one-time addresses to transfer money. These addresses are called "pseudonyms", and it is worth promoting to change the pseudonyms frequently because it can effectively protect the property of their accounts from being discovered by others and protect the identity information of the other party and because it is almost free. The anonymity inherent in blockchain technology promotes the development of privacy protection, and it provides masks and umbrellas for malicious acts and even cybercrime in cyberspace. For example, many black markets use blockchain technology to launder money to escape attacks and use cryptocurrency as their payment method for illegal business activities and then realize the cryptocurrency obtained. The behavior of criminals destroyed the benign development of blockchain ecology and led to many social and economic problems. However, the public chain data are open and cannot be tampered with, which provides an opportunity to identify illegal activities based on blockchain data analysis. In addition to the common analytical tools mentioned in this paper, causal or motivational reasoning is also a useful tool for exploring the underlying behavioral characteristics of blockchain data[101]. This section introduces two main illegal behaviors and puts forward the scheme, process, and idea of detecting illegal behaviors.

5.4.1 Ponzi schemes

As a classic scam, the Ponzi scheme borrows blockchain technology to bring small profits and big losses to participants. Machine learning and data mining technology are widely used to detect Ponzi schemes. Chen et al.[102] collected real-world samples and obtained 200 Ponzi schemes of intelligent contracts by manually checking more than 3000 open-source intelligent contracts on the Ethereum platform. Then, two features were extracted from the transaction history and operation code of the smart contract. Finally, a classification model for detecting intelligent Ponzi schemes was proposed. A large number of experiments show that the performance of this model is better than that of many traditional classification models, and it can achieve higher accuracy in practical applications. In their other work[103], we also used the method of extracting features from the operation code and transaction network to detect contracts related to the Ponzi scheme.

Bartoletti et al.[104] analyzed the use of smart contracts to engage in Ponzi schemes and avoided participating in illegal projects by paying attention to high-return project advertisements, and analyzing contract codes and transaction records.

5.4.2 Money laundering

The natural anonymity of cryptocurrencies, such as Bitcoin, makes them potential money laundering tools. Hu et al.[105] used data collected in the past three years to create a transaction chart and made an in-depth analysis of various chart features to distinguish money laundering transactions from conventional transactions. Exchange, as a platform for cryptocurrency transactions, provides the connection between pseudonyms and real identities. Inspired by this, Ranshous et al.[106] conducted research on the transaction mode of the address under the exchange as an important step in anti-money laundering work. The address owned by the exchange was determined by setting the network theme, and the transaction activity was characterized.

5.4.3 General process of illegal behavior detection

Illegal behavior detection is generally divided into three steps: data acquisition, classification, and clustering. Taking Bitcoin data as an example, we first downloaded data from the full node, analyzed the data, and obtained

the basic characteristics of the transaction ID, inputs, outputs, number of transactions, transaction time, and amount. Then, the transactions were clustered, and addresses belonging to the same user were assigned to the same user name and marked. We extracted and analyzed data characteristics, performed data statistics, classified different types of user characteristics, and distinguished between normal users and abnormal users. Finally, the machine learning method was used to cluster normal and abnormal users. According to the characteristics of user classification, suspected abnormal users were found, and the suspicious users were finally verified by examples[107].

In the clustering process, we need to adopt a combination of a variety of heuristic clustering algorithms to divide user addresses into user address sets as much as possible to improve the accuracy of subsequent clustering. Then, we conducted detailed feature mining on users and presented clear distinguishing conditions. In the verification phase, user transactions in the cluster can be compared with known cases of Bitcoin criminal activities to verify the results.

Shen et al.[108] took a different approach. They used the analysis of transaction motives as the entry point, specified behavior judgment rules based on traders' behavior motives, and then converted the two judgment rules into transaction patterns that can be used to identify abnormal transaction behaviors, i.e., the airdrop candy behavior transaction mode and greedy capital injection behavior transaction mode mentioned in the article. Thus, a method of identifying abnormal Bitcoin transaction behaviors based on a motivation analysis is proposed. Finally, the validity of the identification method was verified through truth-value matching and the identification of real cases.

Both methods can draw a network structure diagram of suspicious users, detect whether changes in the diagram organization over a period of time are related to illegal events, explore the internal connections between suspicious users, and assist market supervision.

6 Open Issues

Although fruitful results have been achieved, the research on cryptocurrency transactions based on graph mining and graph analysis needs further improvement and expansion in methodology and research issues. This section introduces the shortcomings and future development requirements of graph-based analysis and graph mining methods in the existing blockchain data analysis work and puts forward corresponding solutions to these problems.

6.1 Data acquisition and parsing

Data acquisition is the primary task of data analysis. The efficiency and scalability of large-scale data analysis methods bear the challenge of the rapidly growing amount of data on the blockchain. In the face of different sources and increasingly rich data sources, data preprocessing is particularly important. Although there are powerful blockchain data processing tools, such as BlockSci, they still fail to achieve the function of annotating and analyzing auxiliary blockchain data. Based on the work mentioned above, we should work toward the goal of a comprehensive, accurate, real-time, and good classification when dealing with data.

6.2 Deanonymization

Deanonymization is the greatest challenge in blockchain data analysis. If the identity of participants cannot be identified, then it is difficult to understand blockchain datasets and analyze meaningful results, and blockchain analysis can only wander in the initial stage. However, deanonymization does not necessarily need to know the true identity of each address in the ledger, and some jobs can be divided into particular categories to which they belong. Examples include exchanges, miners, ordinary traders, or illegal elements.

In the current work on deanonymization, the main methods are common input methods based on heuristic methods, address change and output transactions, and methods based on behavioral characteristics, including transaction attribute characteristics and user behavior characteristics. There is also the combination of supervised learning and unsupervised learning. With the help of out-of-chain information assistance methods, this kind of information includes, but is not limited to, addresses published on web pages or relevant forums for reward, node IP addresses, exchange of leaked data, and websites that provide tagging services. Visual analysis methods mainly use system tools, which are behind the application of the previous methods. The work on deanonymization requires the combination of a variety of methods to improve the comprehensiveness and accuracy of data as much as possible while ignoring the reduction of efficiency.

6.3 Network dynamic modeling and analysis

In most of the research considered here, the researchers modeled the transaction records as static graphs to

calculate the macro attributes of the network, ignoring the current situation of the dynamic development of the network. It is difficult to describe the dynamic transactions on the network. In network modeling, multiple attributes of nodes and connected edges should be considered. For example, the edges between nodes in Figs. 1 and 6 can also be assigned multiple attributes, such as transaction time and amount.

For a network that is constantly changing and developing, the analysis of the transaction network should reflect its dynamic characteristics and evolution law, such as snapshot networks and time and scale progressive cumulative networks.

6.4 Detection and supervision of illegal behaviors

Cryptocurrency transactions, which are complex, difficult to analyze, and have no central authority to control, are extremely challenging for illegal behavior detection based on blockchains. Criminals often use the anonymity of blockchains to engage in illegal activities, such as the "Silk Road" online transaction platform based on Bitcoin, making Ponzi schemes using smart contracts, and publishing phishing links. In addition to relevant departments specifying regulatory policies, blockchain data analysts should choose a better user address clustering method to determine all users' addresses comprehensively and accurately. They also need to distinguish the characteristics of normal and abnormal users to better complete the work of anonymity, transaction tracking, identity recognition, and capital flow tracking and provide technical support for regulatory authorities to initiate industry supervision.

Accordingly, we should be alert to token issuance activities in the primary market, identify various blockchain investment projects, prevent large-scale and high-value financial fraud, and help investors avoid illegal projects.

6.5 Solutions

In blockchain data analysis, especially on-chain cryptocurrency transaction data analysis, first, we need to acquire and parse data, then perform necessary entity analysis and clustering, and finally conduct metric analysis after modeling the data as a network according to the demand. Faced with the challenges in the existing research, future research can address the above problems by building a blockchain data analysis platform that complies with the ETL process and applies complex network methods to achieve relatively good results. Taking Bitcoin blockchain data as an example, to obtain comprehensive data, this platform should run full-node download data, design analysis algorithms capable of adaptive updates based on the topology and properties of the recent transaction network, or take regular updates to obtain incremental data stored in the database. These steps will help to analyze newly emerged nodes and transactions using the Python-provided blockchain parser package, parse data on demand and use indexes provided by synchronized data files in the Bitcoin wallet, and access and fetch data in index order. Based on the acquired Bitcoin data, the system parses the Bitcoin data structure and computes and stores the data as required.

We considered the fusion of results and the accuracy and comprehensiveness of entity identification and clustering results after the emergence of new data. However, as a result, the algorithm had low efficiency. The platform used an incremental clustering method based on Bitcoin transaction data and synthesized multiple heuristic methods to achieve dynamic updates of data at regular intervals and improve the accuracy and comprehensiveness of clustering results at the expense of certain clustering times. Specifically, block data were analyzed to obtain clustered transactions of wallet addresses, constitute clustered address groups, and extract the relationships among clustered entities by looking up the address index table. The block wallet address data were incrementally clustered using the concatenated set algorithm to obtain new Bitcoin entity relationships and thus infer entity types. At the same time, the entities were identified and labeled to achieve a visual analysis of the entity transaction behavior. The method can accurately perform incremental clustering of addresses, which achieves the effect of deanonymization to a certain extent and reflects the evolution of Bitcoin entities.

In terms of transaction data modeling, the huge cryptocurrency transaction network has inherent time-series dynamics and multidimensional correlation. We can rely on the complex network modeling method based on the dynamic characteristics of complex networks to build a complex network model with time-varying dynamic characteristics. Moreover, modeling the transaction data as a network, we can analyze the network structure using relevant functions, such as the network package for relevant metrics. Specifically, the single- and two-way interactions between users constitute a complex transaction network. For the transaction record participants, the network or graph G can be described by $G = (N, L)$, where the sets N

and L denote the sets of nodes (vertices) and links (edges), respectively. Graph G consists of N nodes interconnected by L links. Examples of graphs with three nodes and two links can be specified as follows: $N = 1, 2, 3$ and $L=(1, 2), (1, 3)$. Nodes can be referenced by their rank i in the set of nodes N. If two nodes i and j are directly connected through a link, then they are called neighbors.

However, while constructing the network, the performance of the experimental device has to be considered, whether to obtain the first-order or second-order neighbors of the transaction address. However, obtaining only the first-order neighbors of the address will give a relatively large amount of data, and the second-order neighbors will be even larger. Random sampling can be considered to reduce the amount of data used to construct the network. Although a larger amount of data is more reflective of the true picture of the network, performing network analysis on a subgraph will be more efficient than analyzing the entire network, i.e., analyzing the way the network grows over a certain period and a smaller network window will be more useful. The dynamics of the network in time can also be implemented with cumulative networks or sliced networks.

In terms of specific applications, the platform can train predictive models using a big data approach, combine multiple methods mentioned in Section 5.1 to overcome anonymity challenges, solve the identity attribution problem of anonymous users, and better perform detection tasks on cryptocurrency transaction networks to identify illegal criminal activities.

On the flip side, although cryptocurrencies can provide anonymity, when we top up or withdraw funds, we need to do so through payment methods requiring users' real names, such as bank cards, and this association with regulated payment methods also provides a regulatory entry point for financial regulation in terms of anti-money laundering detection.

7 Conclusion

Complex network theory has been widely considered a powerful tool for modeling and describing various complex systems. This paper summarizes and analyzes the research work on graph analysis and graph mining for cryptocurrency transaction data to obtain knowledge discovery from the perspective of complex networks.

With the huge cryptocurrency economic ecology, mining and analyzing cryptocurrency transaction data stored on blockchains can not only evaluate transaction behavior problems in complex economic environments and understand the distribution of coins among users or addresses but also help combat cryptocurrency that is used as a payment medium and hence help curb illegal financial activities, such as drug trafficking and money laundering. Accordingly, law enforcement agencies can effectively supervise the cryptocurrency market.

In this article, we also propose a blockchain data analysis framework and present a detailed process for acquiring and analyzing blockchain data, the blockchain data model, the acquisition and analysis of on-chain data, and existing analysis tools. The available insights and application scenarios are summarized and analyzed from five aspects, focusing on the analysis of blockchain data from the perspective of complex networks, especially cryptocurrency transaction data, introducing emerging research challenges and new applications, and discussing future development directions.

To date, due to the nature of blockchain technology and transaction data, cryptocurrency transaction network analysis and knowledge discovery are still considered challenging tasks. Future work can further explore the dynamic characteristics and evolutionary laws of the cryptocurrency transaction network and, based on the effective analysis and mining of the transaction network, propose more deanonymization and entity detection methods. With the interdisciplinary research involving multiple fields, the research on blockchain data from the perspective of complex networks, especially cryptocurrency transaction data, is a very promising field.

Acknowledgment

This work was partially funded by the Key Research and Development Program of Shandong Province (Nos. 2017GGX10142, 2019GNC106027, and 2019JZZY010134); and the Natural Science Foundation of Shandong Province (Nos. ZR2020MF058 and ZR2020MF029).

References

[1] L. Hughes, Y. K. Dwivedi, S. K. Misra, N. P. Rana, V. Raghavan, and V. Akella, Blockchain research, practice and policy: Applications, benefits, limitations, emerging research themes and research agenda, *Int. J. Inf. Manage.*, vol. 49, pp. 114–129, 2019.

[2] S. Nakamoto, Bitcoin: A peer-to-Peer Electronic Cash, https://bitcoin.org/bitcoin.pdf, 2008.

[3] J. Wu, J. Liu, Y. Zhao, and Z. Zheng, Analysis of cryptocurrency transactions from a network perspective: An overview, *J. Netw. Comput. Appl.*, vol. 190, p. 103139,

2021.

[4] L. H. Zhu, F. Gao, M. Shen, Y. D. Li, B. K. Zheng, H. L. Mao, and Z. Wu, Survey on privacy preserving techniques for blockchain technology, (in Chinese), *J. Comput. Res. Dev.*, vol. 54, no. 10, pp. 2170–2186, 2017.

[5] S. Xu, X. Chen, and Y. He, EVchain: An anonymous blockchain-based system for charging-connected electric vehicles, *Tsinghua Science and Technology*, vol. 26, no. 6, pp. 845–856, 2021.

[6] S. Saxena, B. Bhushan, and M. A. Ahad, Blockchain based solutions to secure IoT: Background, integration trends and a way forward, *J. Netw. Comput. Appl.*, vol. 181, p. 103050, 2021.

[7] F. J. de Haro-Olmo, Á. J. Varela-Vaca, and J. A. Álvarez-Bermejo, Blockchain from the perspective of privacy and anonymisation: A systematic literature review, *Sensors*, vol. 20, no. 24, p. 7171, 2020.

[8] C. G. Akcora, M. F. Dixon, Y. R. Gel, and M. Kantarcioglu, Blockchain data analytics. *J. IEEE Intell. Inf.*, vol. 20, no. 1, pp. 1–7, 2019.

[9] W. L. Chen and Z. B. Zheng, Blockchain data analysis: A review of status, trends and challenges, (in Chinese), *J. Comput. Res. Dev.*, vol. 55, no. 9, pp. 1853–1870, 2018.

[10] R. Xin, J. Zhang, and Y. Shao, Complex network classification with convolutional neural network, *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 447–457, 2020.

[11] S. S. Adebola, L. A. Gil-Alana, and G. Madigu, Gold prices and the cryptocurrencies: Evidence of convergence and cointegration, *Phys. A: Stat. Mech. Appl.*, vol. 523, pp. 1227–1236, 2019.

[12] Z. Wei, P. Wan, L. Xiao, and D. Shen, The inefficiency of cryptocurrency and its cross-correlation with Dow Jones Industrial Average, *Phys. A: Stat. Mech. Appl.*, vol. 510, pp. 658–670, 2018.

[13] A. Sward, I. Vecna, and F. Stonedahl, Data insertion in Bitcoin's blockchain, *Ledger*, vol. 3, pp. 1–23, 2018.

[14] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara, and K. Sadatdiynov, A survey of data partitioning and sampling methods to support big data analysis, *Big Data Mining and Analytics*, vol. 3, no. 2, pp. 85–101, 2020.

[15] S. Hong and H. Kim, Analysis of Bitcoin exchange using relationship of transactions and addresses, in *Proc. 21st Int. Conf. Advanced Communication Technology* (*ICACT*), Pyeongchang, Korea, 2019, pp. 67–70.

[16] R. Galici, L. Ordile, M. Marchesi, A. Pinna, and R. Tonelli, Applying the ETL process to blockchain data. Prospect and findings, *Information*, vol. 11, no. 4, p. 204, 2020.

[17] C. Kinkeldey, J. D. Fekete, and P. Isenberg, BitConduite: Visualizing and analyzing activity on the Bitcoin network, in *Proc. 19th Eurographics Conf. Visualization*, Barcelona, Spain, 2017, pp. 25–27.

[18] M. Spagnuolo, F. Maggi, and S. Zanero, BitIodine: Extracting intelligence from the Bitcoin network, in *Proc. 18th Int. Conf. Financial Cryptography and Data Security*, Christ Church, Barbados, 2014, pp. 457–468.

[19] D. Kondor, M. Pósfai, I. Csabai, and G. Vattay, Do the rich get richer? An empirical analysis of the Bitcoin transaction network, *PLoS One*, vol. 9, no. 2, p. e86197, 2014.

[20] B. Zheng, L. Zhu, M. Shen, X. J. Du, and M. Guizani, Identifying the vulnerabilities of Bitcoin anonymous mechanism based on address clustering, *Sci. China Inf. Sci.*, vol. 63, no. 3, p. 132101, 2020.

[21] B. Zheng, L. Zhu, M. Shen, X. Du, J. Yang, F. Gao, Y. Li, C. Zhang, S. Liu, and S. Yin, Malicious Bitcoin transaction tracing using incidence relation clustering, in *Proc. 9th Int. Conf. Mobile Networks and Management*, Melbourne, Australia, 2017, pp. 313–323.

[22] J. Liang, L. Li, W. Chen, and D. Zeng, Targeted addresses identification for Bitcoin with network representation learning, in *Proc. 2019 IEEE Intelligence and Security Informatics*, Shenzhen, China, 2019, pp. 158–160.

[23] P. Tasca, A. Hayes, and S. Liu, The evolution of the Bitcoin economy: Extracting and analyzing the network of payment relationships, *J. Risk Finance*, vol. 19, no. 2, pp. 94–126, 2018.

[24] D. Di Francesco Maesa, A. Marino, and L. Ricci, Uncovering the Bitcoin blockchain: An analysis of the full users graph, in *Proc. 2016 IEEE Int. Conf. Data Science and Advanced Analytics* (*DSAA*), Montreal, Canada, 2016, pp. 537–546.

[25] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage, A fistful of Bitcoins: Characterizing payments among men with no names, in *Proc. 2013 Conf. Internet Measurement Conf.*, Barcelona, Spain, 2013, pp. 127–140.

[26] F. Zola, J. L. Bruse, M. Eguimendia, M. Galar, and R. O. Urrutia, Bitcoin and cybersecurity: Temporal dissection of blockchain data to unveil changes in entity behavioral patterns, *Appl. Sci.*, vol. 9, no. 23, p. 5003, 2019.

[27] M. Fleder, M. S. Kester, and S. Pillai, Bitcoin transaction graph analysis, arXiv preprint arXiv: 1502.01657, 2015.

[28] H. Sun, N. Ruan, and H. Liu, Ethereum analysis via node clustering, in *Proc. 13th Int. Conf. Network and System Security*, Sapporo, Japan, 2019, pp. 114–129.

[29] T. Chen, Z. Li, Y. Zhu, J. Chen, X. Luo, J. C. S. Lui, X. Lin, and X. Zhang, Understanding ethereum via graph analysis, *ACM Trans. Internet Technol.*, vol. 20, no. 2, p. 18, 2020.

[30] X. T. Lee, A. Khan, S. S. Gupta, Y. H. Ong, and X. Liu, Measurements, analyses, and insights on the entire ethereum blockchain network, in *Proc. Web Conf. 2020*, Taipei, China, 2020, pp. 155&166.

[31] S. Ferretti and G. D'Angelo, On the ethereum blockchain structure: A complex networks theory perspective, *Concurr. Comput.: Pract. Exp.*, vol. 32, no. 12, p. e5493, 2020.

[32] D. Lin, J. Wu, Q. Yuan, Z. Zheng, Modeling and understanding ethereum transaction records via a complex network approach, *IEEE Trans. Circuits Syst. II: Express Briefs*, vol. 67, no. 11, pp. 2737–2741, 2020.

[33] D. Y. Huang, K. Levchenko, and A. C. Snoeren, Estimating profitability of alternative cryptocurrencies (short paper), in *Proc. 22nd Int. Conf. Financial Cryptography and Data Security*, Nieuwpoort, Belgium, 2018, pp, 409–419.

[34] T. Chen, Z. Li, Y. Zhang, X. Luo, A. Chen, K. Yang,

B. Hu, T. Zhu, S. Deng, T. Hu, et al., DataEther: Data exploration framework for ethereum, in *Proc. 39th Int. Conf. Distributed Computing Systems* (*ICDCS*), Dallas, TX, USA, 2019, pp. 1369–1380.

[35] Y. Li, U. Islambekov, C. Akcora, E. Smirnova, Y. R. Gel, and M. Kantarcioglu, Dissecting ethereum blockchain analytics: What we learn from topology and geometry of the ethereum graph?, in *Proc. 2020 SIAM Int. Conf. Data Mining*, Cincinnati, OH, USA, 2020, pp. 523–531.

[36] S. Farrugia, J. Ellul, and G. Azzopardi, Detection of illicit accounts over the Ethereum blockchain, *Expert Syst. Appl.*, vol. 150, p. 113318, 2020.

[37] D. Lin, J. Wu, Q. Yuan, and Z. Zheng, T-EDGE: Temporal WEighted MultiDiGraph embedding for ethereum transaction network analysis, *Front. Phys.*, vol. 8, p. 204, 2020.

[38] Q. Bai, C. Zhang, Y. Xu, X. Chen, and X. Wang, Poster: Evolution of ethereum: A temporal graph perspective, in *Proc. 2020 IFIP Networking Conf.* (*Networking*), Paris, France, 2020, pp. 652–654.

[39] D. Guo, J. Dong, and K. Wang, Graph structure and statistical properties of Ethereum transaction relationships, *Inf. Sci.*, vol. 492, pp. 58–71, 2019.

[40] M. Poongodi, A. Sharma, V. Vijayakumar, V. Bhardwaj, A. P. Sharma, R. Iqbal, and R. Kumar, Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system, *Comput. Electr. Eng.*, vol. 81, p. 106527, 2020.

[41] J. Tigani and S. Naidu, *Google BigQuery Analytics*. Indianapolis, IN, USA: Wiley, 2014.

[42] W. Zheng, Z. Zheng, H. N. Dai, X. Chen, and P. Zheng, XBlock-EOS: Extracting and exploring blockchain data from EOSIO, *Inf. Process. Manage.*, vol. 58, no. 3, p. 102477, 2021.

[43] P. Zheng, Z. Zheng, J. Wu, and H. N. Dai, XBlock-ETH: Extracting and exploring blockchain data from ethereum, *IEEE Open J. Comput. Soc.*, vol. 1, pp. 95–106, 2020.

[44] H. L. Mao, Z. Wu, M. He, J. Q. Tang, and M. Shen, Heuristic approaches based clustering of Bitcoin addresses, (in Chinese), *J. Beijing Univ. Posts Telecomm.*, vol. 41, no. 2, pp. 27–31, 2018.

[45] L. T. Leong, Snapshot samplings of the Bitcoin transaction network and analysis of cryptocurrency growth, arXiv preprint, arXiv: 2003.06068, 2020.

[46] J. Liang, L. Li, and D. Zeng, Evolutionary dynamics of cryptocurrency transaction networks: An empirical study, *PLoS One*, vol. 13, no. 8, p. e0202202, 2018.

[47] M. Lischke and B. Fabian, Analyzing the Bitcoin network: The first four years, *Future Internet*, vol. 8, no. 1, p. 7, 2016.

[48] A. P. Motamed and B. Bahrak, Quantitative analysis of cryptocurrencies transaction graph, *Appl. Netw. Sci.*, vol. 4, no. 1, p. 131, 2019.

[49] A. Baumann, B. Fabian, and M. Lischke, Exploring the Bitcoin network, in *Proc. 10th Int. Conf. Web Information Systems and Technologies*, Barcelona, Spain, 2014, pp. 369–374.

[50] W. Chen, T. Zhang, Z. Chen, Z. Zheng, and Y. Lu, Traveling the token world: A graph analysis of ethereum ERC20 token ecosystem, in *Proc. Web Conf. 2020*, Taipei, China, 2020, pp. 1411–1421.

[51] D. Di Francesco Maesa, A. Marino, and L. Ricci, An analysis of the Bitcoin users graph: Inferring unusual behaviours, in *Complex Networks & Their Applications V*, H. Cherifi, S. Gaito, W. Quattrociocchi, and A. Sala, eds. Cham, Germany: Springer, 2016, pp. 749–760.

[52] X. F. Wang, X. Li, and G. R. Chen, *Network Science: An Introduction*, (in Chinese), Beijing, China: Higher Education Press, 2012.

[53] A. Kumar, A. Kumar, P. Nerurkar, M. R. Ghalib, A. Shankar, Z. Wen, and X Qi, Empirical Analysis of Bitcoin network (2016-2020), in *Proc. 2020 IEEE/CIC Int. Conf. Communications in China*, Chongqing, China, 2020, pp. 96–101.

[54] I. Alqassem, I. Rahwan, and D. Svetinovic, The anti-social system properties: Bitcoin network data analysis, *IEEE Trans. Syst., Man, Cybernetics: Syst.*, vol. 50, no. 1, pp. 21–31, 2020.

[55] WalletExplorer, https://www.walletexplorer.com/info, 2021.

[56] Hubwiz.com, http://sc.hubwiz.com/codebag/blocketl-java/, 2021.

[57] GitHub, https://github.com/blockchain-etl/bitcoin-etl, 2021.

[58] H. Kalodner, M. Möser, K. Lee, S. Goldfeder, M. Plattner, A. Chator, and A. Narayanan, BlockSci: Design and applications of a blockchain analysis platform, in *Proc. 29th USENIX Conf. Security Symp.*, Berkeley, CA, USA, 2017, p. 153.

[59] Blockchain.info, http://blockchain.info/, 2021.

[60] Chainalysis, https://www.chainalysis.com/, 2021.

[61] GitHub, https://github.com/stefanolande/blockapi, 2021.

[62] G. Di Battista, V. Di Donato, M. Patrignani, M. Pizzonia, V. Roselli, and R. Tamassia, Bitconeview: Visualization of flows in the Bitcoin transaction graph, in *Proc. 2015 IEEE Symp. Visualization for Cyber Security* (*VizSec*), Chicago, IL, USA, 2015, pp. 1–8.

[63] X. Yue, X. Shu, X. Zhu, X. Du, Z. Yu, D. Papadopoulos, and S. Liu, BitExTract: Interactive visualization for extracting Bitcoin exchange intelligence, *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 162–171, 2019.

[64] F. Oggier, S. Phetsouvanh, and A. Datta, BiVA: Bitcoin network visualization & analysis, in *Proc. 2018 IEEE Int. Conf. Data Mining Workshops* (*ICDMW*), Singapore, 2018.

[65] Y. Boshmaf, H. Al Jawaheri, and M. Al Sabah, BlockTag: Design and Applications of a Tagging System for Blockchain Analysis, in *ICT Systems Security and Privacy Protection*, G. Dhillon, F. Karlsson, K. Hedström, and A. Zúquete, eds. Cham, Germany: Springer, 2019, pp. 299–313.

[66] Etherscan, https://cn.etherscan.com/, 2021.

[67] W. Chen, J. Wu, Z. Zheng, C. Chen, and Y. Zhou, Market manipulation of Bitcoin: Evidence from mining the Mt. Gox transaction network, in *Proc. IEEE INFOCOM 2019–IEEE Conf. Computer Communications*, Paris, France, 2019, pp. 964–972.

[68] E. Brinckman, A. Kuehlkamp, J. Nabrzyski, and I. J. Taylor, Techniques and applications for crawling, ingesting and analyzing blockchain data, in *Proc. 2019 Int. Conf. Information and Communication Technology Convergence* (*ICTC*), Jeju, Republic of Korea, 2019, pp. 717–722.

[69] Q. Hou, M. Han, and Z. Cai, Survey on data analysis in social media: A practical application aspect, *Big Data Min. Anal.*, vol. 3, no. 4, pp. 259–279, 2020.

[70] D. Ron and A. Shamir, Quantitative analysis of the full Bitcoin transaction graph, in *Proc. 17th Int. Conf. Financial Cryptography and Data Security*, Okinawa, Japan, 2013, pp. 6–24.

[71] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. A. Savage, A fistful of bitcoins: Characterizing payments among men with no names, *Commun. ACM*, vol. 59, no. 4, pp. 86–93, 2016.

[72] J. S. Wang, Z. M. Lü, Z. N. Zhao, and H. W. Zhang, Address incremental clustering method for visual analysis of blockchain transaction, (in Chinese), *Comput. Eng.*, vol. 46, no. 8, pp. 14–20, 2020.

[73] F. Victor, Address clustering heuristics for ethereum, in *Proc. 24th Int. Conf. Financial Cryptography and Data Security*, Kota Kinabalu, Malaysia, 2020, pp. 617–633.

[74] J. V. Monaco, Identifying Bitcoin users by transaction behavior, in *Proc. SPIE 9457, Biometric and Surveillance Technology for Human and Activity Identification XII*, Baltimore, MD, USA, 2015, p. 945704.

[75] F. Chen, H. Wan, H. Cai, and G. Cheng, Machine Learning in/for blockchain: Future and challenges, *Can. J. Stat.*, vol. 49, no. 4, pp. 1364–1382, 2021.

[76] J. Zhu, P. Liu, and L. He, Mining information on Bitcoin network data, in *Proc.* 2017 *IEEE Int. Conf. Internet of Things (Ithings) and IEEE Green Computing and Communications (Greencom) and IEEE Cyber, Physical and Social Computing (Cpscom) and IEEE Smart Data (Smartdata)*, Exeter, UK, 2017, pp. 999–1003.

[77] F. Bres, I. A. Seres, A. A. Benczr, and M. Quintyne-Collins, Blockchain is watching you: Profiling and deanonymizing ethereum users, arXiv preprint arXiv: 2005.14051, 2020.

[78] Y. Xing, Research on de-anonymization techniques of Bitcoin trading network, (in Chinese), Master dissertation, Southeast University, Nanjing, China, 2017.

[79] A. Biryukov, D. Khovratovich, and I. Pustogarov, Deanonymisation of clients in bitcoin P2P network, in *Proc. ACM SIGSAC Conf. Computer and Communications Security*, Scottsdale, AZ, USA, 2014, pp. 15–29.

[80] S. Kairam, N. H. Riche, S. Drucker, R. Fernandez, and J. Heer, Refinery: Visual exploration of large, heterogeneous networks through associative browsing, *Comput. Graph. Forum*, vol. 34, no. 3, pp. 301–310, 2015.

[81] D. McGinn, D. Birch, D. Akroyd, M. Molina-Solana, Y. Guo, and W. J. Knottenbelt, Visualizing dynamic Bitcoin transaction patterns, *Big Data*, vol. 4, no. 2, pp. 109–119, 2016.

[82] A. B. Turner, S. McCombie, and A. J. Uhlmann, Discerning payment patterns in Bitcoin from ransomware attacks, *J. Money Laund. Control*, vol. 23, no. 3, pp. 545–589, 2020.

[83] F. Zola, J. L. Bruse, M. Eguimendia, M. Galar, and R. O. Urrutia, Bitcoin and cybersecurity: Temporal dissection of blockchain data to unveil changes in entity behavioral patterns, *Appl. Sci.*, vol. 9, no. 23, p. 5003, 2019.

[84] Z. Guo and S. Zhang, Sparse deep nonnegative matrix factorization, *Big Data Mining and Analytics*, vol. 3, no. 1, pp. 13–28, 2020.

[85] W. Shu and Y. H. Chuang, The perceived benefits of six-degree-separation social networks, *Internet Res.*, vol. 21, no. 1, pp. 26–45, 2011.

[86] X. F. Liu, X. J. Jiang, S. H. Liu, and C. K. Tse, Knowledge discovery in cryptocurrency transactions: A survey, *IEEE Access*, vol. 9, pp. 37229–37254, 2021.

[87] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[88] Y. Zhao, J. Liu, Q. Han, W. Zheng, and J. Wu, Exploring EOSIO via graph characterization, in *Proc. 2nd Int. Conf. Blockchain and Trustworthy Systems*, Dali, China, 2020, pp. 475–488.

[89] G. H. Nguyen, J. B. Lee, R. A. Rossi, N. K. Ahmed, E. Koh, and S. Kim, Continuous-time dynamic network embeddings, in *Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 969–976.

[90] S. Somin, G. Gordon, and Y. Altshuler, Network analysis of ERC20 tokens trading on ethereum blockchain, in *Unifying Themes in Complex Systems IX*, A. J. Morales, C. Gershenson, D. Braha, A. A. Minai, and Y. Bar-Yam, Eds. Cham, Germany: Springer, 2018, pp. 439–450.

[91] D. Di Francesco Maesa, A. Marino, and L. Ricci, Detecting artificial behaviours in the Bitcoin users graph, *Online Soc. Netw. Media*, vols. 3&4, pp. 63–74, 2017.

[92] D. Di Francesco Maesa, A. Marino, and L. Ricci, Data-driven analysis of Bitcoin properties: Exploiting the users graph, *Int. J. Data Sci. Anal.*, vol. 6, no. 1, pp. 63–80, 2018.

[93] H. H. S. Yin, K. Langenheldt, M. Harlev, R. R. Mukkamala, and R. Vatrapu, Regulating cryptocurrencies: A supervised machine learning approach to de-anonymizing the Bitcoin blockchain, *J. Manage. Inf. Syst.*, vol. 36, no. 1, pp. 37–73, 2019.

[94] V. G. Reyes-Macedo, M. Salinas-Rosales, and G. G. Garcia, A method for blockchain transactions analysis, *IEEE Lat. Am. Trans.*, vol. 17, no. 7, pp. 1080–1087, 2019.

[95] J. C. Pan, D. M. Han, F. Z. Guo, W. T. Zheng, J. H. Yu, and W. Chen, Visual exploration of topological structure for Bitcoin trading network, (in Chinese), *J. Softw.*, vol. 30, no. 10, pp. 3017–3025, 2019.

[96] P. Nerurkar, D. Patel, Y. Busnel, R. Ludinard, S. Kumari, and M. K. Khan, Dissecting Bitcoin blockchain: Empirical analysis of Bitcoin network (2009–2020), *J. Netw. Comput. Appl.*, vol. 177, p. 102940, 2021.

[97] N. Gandal, J. T. Hamrick, T. Moore, and T. Oberman, Price manipulation in the Bitcoin ecosystem, *J. Monetary Econ.*, vol. 95, pp. 86–96, 2018.

[98] D. Ermilov, M. Panov, and Y. Yanovich, Automatic Bitcoin

address clustering, in *Proc. 16th IEEE Int. Conf. Machine Learning and Applications* (*ICMLA*), Cancun, Mexico, 2017, pp. 461–466.

[99] A. Di Luzio, A. Mei, and J. Stefa, Consensus robustness and transaction de-anonymization in the ripple currency exchange system, in *Proc. 39th Int. Conf. Distributed Computing Systems* (*ICDCS*), Atlanta, GA, USA, 2017, pp. 140–150.

[100] F. Gao, L. Zhu, K. Gai, C. Zhang, and S. Liu, Achieving a covert channel over an open blockchain network, *IEEE Netw.*, vol. 34, no. 2, pp. 6–13, 2020.

[101] Z. Wang, C. Wang, X. Ye, J. Pei, and B. Li, Propagation history ranking in social networks: A causality-based approach, *Tsinghua Science and Technology*, vol. 25, no. 2, pp. 161–179, 2020.

[102] W. Chen, Z. Zheng, E. C. H. Ngai, P. Zheng, and Y. Zhou, Exploiting blockchain data to detect smart Ponzi schemes on ethereum, *IEEE Access*, vol. 7, pp. 37575–37586, 2019.

[103] W. Chen, Z. Zheng, J. Cui, E. Ngai, P. Zheng, and Y. Zhou, Detecting Ponzi schemes on ethereum: Towards healthier blockchain technology, in *Proc. 2018 World Wide Web Conf.*, Lyon, France, 2018, pp. 1409–1418.

[104] M. Bartoletti, S. Carta, T. Cimoli, and R. Saia, Dissecting Ponzi schemes on Ethereum: Identification, analysis, and impact, *Future Gener. Comput. Syst.*, vol. 102, pp. 259–277, 2020.

[105] Y. Hu, S. Seneviratne, K. Thilakarathna, K. Fukuda, and A. Seneviratne, Characterizing and detecting money laundering activities on the Bitcoin network, arXiv preprint arXiv: 1912.12060, 2019.

[106] S. Ranshous, C. A. Joslyn, S. Kreyling, K. Nowak, N. F. Samatova, C. L. West, and S. Winters, Exchange pattern mining in the Bitcoin transaction directed hypergraph, in *Proc. 2017 Int. Conf. on Financial Cryptography and Data Security*, Sliema, Malta, 2017, pp. 248–263.

[107] L. Yang, X. Dong, S. Xing, J. Zheng, X. Gu, and X. Song. An abnormal transaction detection mechanim on Bitcoin. in *Proc. 2019 Int. Conf. Networking and Network Applications* (*NaNA*), Daegu, Republic of Korea, 2019, pp. 452–457.

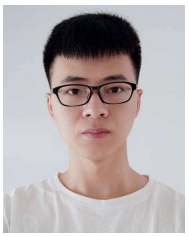
[108] M. Shen, A. Q. Sang, L. H. Zhu, R. G. Sun, and C. Zhang. Abnormal transaction behavior recognition based on motivation analysis in blockchain digital currency. *Chin. J. Comput.*, vol. 44, no. 1, pp. 193–208, 2021.

Wanshui Song received the BS degree from Weifang University of Science and Technology in 2019, and he is currently studying for the master degree in Shandong Normal University. He is a reviewer of *Journal of Supercomputing*. His research mainly focuses on blockchain data analysis, blockchain illegal behavior detection, and blockchain transaction pattern recognition.

Wenyin Zhang is a doctor of Chengdu Institute of Computing, Chinese Academy of Sciences, a postdoctoral fellow of the School of Automation, Tianjin University, a doctoral supervisor of Suwon University, Republic of Korea, a professor, dean, and master supervisor of the School of Information Science and Engineering, Linyi University, and a member of CCF and ACM. He is mainly engaged in computer vision processing, digital watermarking, network information security, blockchain technology application, and other research work.

Linbo Zhai received the BS and MS degrees from the School of information Science and Engineering, Shandong University, in 2004 and 2007, respectively, and the PhD degree from School of Electronic Engineering, Beijing University of Posts and Telecommunication, in 2010. Since then, he has been a teacher with Shandong Normal University. His current research interests include cognitive radio, crowdsourcing, and distributed network optimization.

Jiuru Wang received the MS degree from Anhui University of Science and Technology in 2009, and PhD degree from Harbin Engineering University in 2013. Now, He is working as an associate professor at Linyi University. His research interests mainly include information security, key management, and blockchain application.



Pengkun Jiang is currently studying for the master degree at Linyi University. His research mainly focuses on information hiding in blockchain.

Shanyun Huang received the bachelor degree in 2019, and he is currently studying for the master degree in Shandong Normal University. His research mainly focuses on information hiding in blockchain.

Bei Li received the master degree in 2021. Her research mainly focuses on blockchain privacy protection.