Received: 3 November 2021 | Revised: 22 February 2022 | Accepted: 1 May 2022

IET Blockchain

DOI: 10.1049/blc2.12014

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Bitcoin address clustering method based on multiple heuristic conditions

Xi He[1] | Ketai He[1] | Shenwen Lin[2] | Jinglin Yang[2] | Hongliang Mao[2]

[1]School of Mechanical Engineering, University of Science and Technology Beijing, Beijing, China

[2]Internet Financial Security Technology Key Laboratory, National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China

Correspondence
Ketai He, School of Mechanical Engineering, University of Science and Technology Beijing, Beijing, China.
Email: heketai@ustb.edu.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2019QY(Y)0601

Abstract

Single heuristic method and incomplete heuristic conditions were difficult to cluster a large number of addresses comprehensively and accurately. Therefore, this paper analysed the associations between Bitcoin transactions and addresses and used six heuristic conditions to cluster addresses and entities. We proposed an improved change address detection algorithm and compared it with the original change address algorithm to prove the effectiveness of the improved algorithm. By adding conditional constraints, the identified change address was more accurate, and the convergence speed of the algorithm was accelerated. Our work presented the pseudo-anonymity mechanism of the Bitcoin system, which could be used by the law enforcement agencies to track and crack down illegal transactions.

1 | INTRODUCTION

Since Bitcoin became a peer-to-peer digital payment way, it has attracted the attention of a legion of researchers. Currently there is a lot of research which has focused on various areas of the Bitcoin system, such as privacy security [1], transaction pattern [2], network analysis [3, 4], and price prediction [5]. In addition, many authors have also made detailed summaries about this field in their literature reviews [6–8]. They tended to expound the research issues, summarise the methods, and analyse the results and findings. Also, they presented the main challenges and several future directions in this area. Another thrust of Bitcoin entity research, closer to our own interest, has focused on the community discovery. At present, there is not much literature on the Bitcoin network community. Researchers mainly focused on analysing the network graph property of various communities [9, 10] and then speculating on the characteristics of large entities in the Bitcoin network [11, 12]. Therefore, studying the network of the community, obtaining the structure of the community, and then explaining the transaction pattern of the large Bitcoin community are our research goals in the future. Blockchain technology is favoured by many people, because it breaks the traditional centralisation and establishes a trust mechanism. The birth of the Bitcoin system makes it easier and faster to transact among people. The anonymity and decentralisation of the Bitcoin system play an important role when people use it to conduct peer-to-peer transactions. Meanwhile, the anonymity of Bitcoin also provides protection for some illegal transactions. Since anyone can create multiple Bitcoin accounts for transactions on the network where their identity is represented by a Bitcoin address consisting of numbers and letters, Bitcoin address cannot reflect the identity of the entity. At the same time, some illegal entities rarely reuse addresses to avoid tracking from law enforcement agencies, and instead, they create new addresses for each transaction to reduce their presence on the network. However, previous studies [13] indicated that with the combination of off-chain information [14], addresses could be clustered to the same transaction entity by analysing the transaction records of addresses in the full ledger data of blockchain, which is beneficial to further identifying entities and inferring relationships between entities. On the other hand, it also exposes the problems of anonymous Bitcoin system mechanisms. Researchers have used these heuristics and off-chain information to develop blockchain analysis software, such as Bitlodine [15], BitConeview [16], BitConduite [17] and BitExtract [18]. In this study, we aimed to cluster the addresses using heuristic algorithms comprehensively based on the previous work, and to further explore the

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2022 The Authors. *IET Blockchain* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

potential relationships between entities using community detection algorithm. The main contributions of this paper are as follows.

1. The methods of Bitcoin address clustering are analysed and summarised in detail. This paper aims to explore the relationship among Bitcoin addresses comprehensively, so as to improve the degree of address aggregation.
2. We analyse the reasons for the low accuracy of detection of Bitcoin change address, combine with a variety of limited conditions for detection of change address, and propose an improved change address detection process based on different number of output address transactions.
3. We use the Louvain community detection algorithm, combined with the idea of complex network to analyse the relationship between transaction entities, and further detect the potential relationship between entities.
4. The clustering rules of the well-known wallet analysis website (WalletExplorer, a smart Bitcoin block explorer) are revealed for the first time. We obtain the rules for marking the change address of the website. And all of these rules are confirmed by the website developer.
5. Through a large amount of data and experiments, we have proved the effectiveness of applying multiple heuristic algorithms, and we can analyse Bitcoin transaction activities using the research methods, thus laying a good foundation for the identification of transaction entities.

The remainder of this paper is organised as follows. In Section 2, we briefly introduce the traceability of Bitcoin transaction entities. In Section 3, we summarise the current heuristic algorithms, analyse the problems of the change address detection method in detail, and put forward the change address algorithm process after classifying in accordance with the number of output addresses. In Section 4, we use a variety of heuristic algorithms mentioned in Section 3 to carry out experimental implementation. We elaborate the experimental environment and data sources, give the algorithm flow of address clustering and entity community division, and analyse the performance of the experimental results and method. Section 5 is the related work. Finally, Section 6 is the conclusion and prospect of this paper.

2 | TRACEBILITY OF BITCOIN TRANSACTION ENTITIES

Blockchain technology makes the Bitcoin ledger public, immutable and stable, which provides a solid basis for inferring the identity of the entity.

Bitcoin transactions have the following important characteristics.

1. Each transaction acquires the sender's private key to sign.
2. Every transaction will be recorded in the blockchain ledger and anyone can view its transaction details.
3. Each full node of Bitcoin maintains an unspent transaction outputs (UTXO) set, making each transaction traceable.

Transactions are conducted through Bitcoin address by Bitcoin holders. For scattered Bitcoin users, it is difficult for us to label the users who belong to these Bitcoin addresses. This is one of the reasons why it is difficult to trace the entity when a single user conducts illegal transactions. However, except for a single user, there are many large-scale transaction entities in the Bitcoin system, such as Exchanges, gambling companies, mining pools and Bitcoin mixing service providers. Since large entities involve many transactions, it is possible for us to extract the transaction characteristics of these large entities through machine learning, and then use the model to identify other unknown entities that are similar but not labelled.

In general, the traceability of the transaction activities of Bitcoin transaction entities can be started from two aspects. On the one hand, cluster analysis of transaction addresses could be used to find the address groups controlled by each transaction entity. On the other hand, collecting off-chain data is also conducive to linking Bitcoin addresses with entity identities, such as the contextual information with Bitcoin addresses appearing on various web pages, personal information filled in by users when registering in Bitcoin forums, and website data that specifically analyse Bitcoin address tags. Through the methods mentioned above, the transaction activities in which the entity participates can be traced and the information dimension of the entity's identity increased as well.

3 | METHODOLOGY

Multiple heuristics were used to address clustering to improve the degree of entity aggregation. In this section, we described a total of six heuristic algorithms, and some of the existing methods were improved. Louvain Community Detection Algorithm was used to divide the entities into communities for further analysis of the relationship among entities. We introduce and compare six heuristic algorithms, as shown in Table 1.

3.1 | Multiple heuristic address clustering

Satoshi Nakamoto mentioned in his paper [19] *Bitcoin: A Peer-to-Peer Electronic Cash System* that in order to reduce transaction fees, no one is willing to make small transfers many times, but tends to combine the balances in his multiple accounts to meet the amount required for the transaction. Based on this statement, many scholars have carried out clustering studies on Bitcoin addresses. One of the most important methods is called the common-input-ownership-heuristic. Because the clustering accuracy of this method is truly high, most current analysis of Bitcoin transaction still relies on this method for clustering study of transaction addresses.

Heuristic algorithm 1 **(H1)**: A heuristic algorithm based on transaction input address, common-input-ownership heuristic.

TABLE 1 Methodology

Algorithm	Name	Function	Clustering type	Accuracy	Remark
H1	Common-input-ownership heuristic	Based on transaction input address	Address clustering	100%	On the premise that there is no mixing service
H2	Change address detection heuristic	Based on transaction input address and output address	Address clustering	Influenced by heuristic conditions	–
H3	Coinbase transaction mining address clustering heuristic	Based on transaction output address	Address clustering	100%	–
H4	Multiple mining pool address clustering heuristic	Based on the number of output addresses	Address clustering	Influenced by heuristic conditions	–
H5	Mixing transaction recognition heuristic	Based on equal currency multi-input multi-output	Address clustering	Influenced by heuristic conditions	–
H6	Louvain community detection algorithm	Cluster of addresses belonging to the same entity	Community division	Influenced by heuristic conditions	–

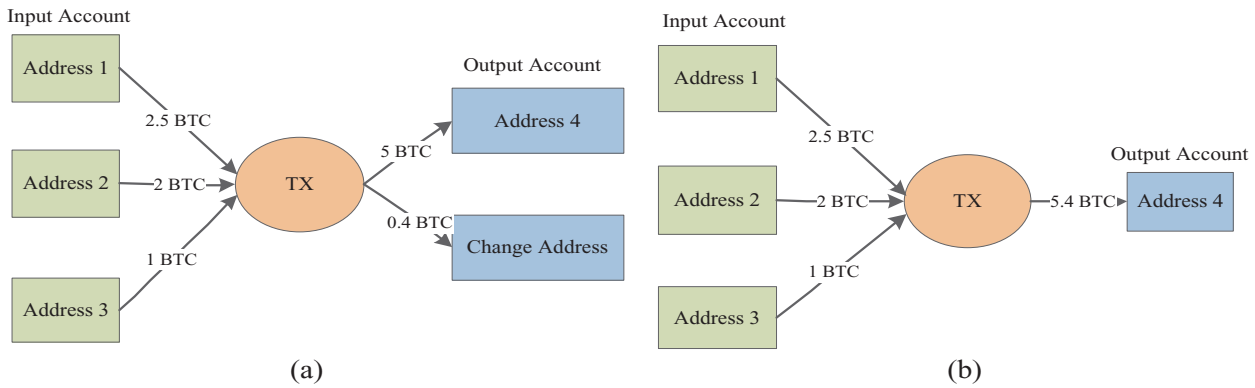

FIGURE 1 Common-input-ownership Schematic diagram

According to the protocol of the Bitcoin system, if you want to use the Bitcoin of one address, the private key of that address must be provided, which means that the user of the address must sign for the transaction. Consequently, when multiple addresses are used as the input of a transaction together, we believe that all the input addresses of the transaction can be clustered into an address group. In other words, all the input addresses are controlled by the same transaction entity. The clustering accuracy of H1 can reach 100% without considering the fact that users use mixing services to avoid clustering analysis intentionally. A simple example illustrates the transaction process, as shown in Figure 1. In Figure 1(a), the three input addresses belong to the same transaction entity. In the output addresses, the transaction receiver's account is Address 4, the change amount is 0.4 BTC, and the handling fee for this transaction is 0.1 BTC. If a transaction requires no change, as shown in Figure 1(b), the input amount equals the output amount plus the transaction fee.

In addition to the common-input-ownership heuristic, another heuristic for address clustering is the change address detection heuristic algorithm. In the Bitcoin system, every time a transaction occurs, a node will package the transaction and record it on the blockchain ledger. After removing the transfer amount and transaction fees, the remaining Bitcoins will be stored in the change address. Obviously, the Bitcoin in the change address belongs to the entity where the input address in the transaction is located. Thus, there is an association between the current change address and the input address. That is to say, they are both controlled by the same entity. At the same time, the Bitcoins in the change address will be used as the input amount for a later transaction. So, if you can detect the change address in a transaction record over time, you can link multiple transactions together, increasing the degree of aggregation between multiple different addresses. The research of Androulaki *et al.* [20] and Meiklejohn *et al.* [21] also indicates that change address is an important mechanism to enhance user privacy. The vulnerability existing in the anonymous mechanism of Bitcoin once again proves that the recognition of change address can improve the degree of address aggregation and play a good role in entity

ALGORITHM 1 H1 + H2 + H3 + H4

Input: Query_addr, iteration count: m;
Output: Addresses in the same cluster.

Initialise transaction data set txLsit;
Initialise temporary address set tempList;
Initialise Cluster result set ClusterList;
Initialise New query addresses set newquerylist;
Add Acquery_addr to queryList;
Add Acquery_addr to ClusterList;
while m **do**
for acquery in queryList **do**
for TX in txLsit **do**
if TX is Coinbase transaction **then**
Extract all output addresses from TX and add them to the collection tempList;
elif TX is mining pool transaction
Extract all output addresses from TX and add them to the collection another tempList;
else
Extract all input addresses from TX and add them to the collection tempList;
Extract change address from outputs and add to tempList;
if acquery in tempList **then**
tempList append to newquerylist;
if newqueryList is null **then**
return ClusterList;
else
queryList equals newqueryList;
newqueryList append to ClusterList;
m–;
return ClusterList.

ALGORITHM 2 The entity community division algorithm

Input: The result of addresses clustering by H1 + H2 + H3 + H4;
Output: Community divided results.

Extract all transactions for the experimental address;
Build a transaction network diagram;
Initialize best_partition= V_i;
while true **do**
while true **do**
for V_i in V **do**
Add node V_i to neighbor node $V_j$and calculate ΔQ;
Move V_i into the community where ΔQ is largest;
if no vertex moves to new community **then**
break
Rebuild the community, compress all nodes in the same community into a new node;
if best_partition not change **then**
break
return communities.

identification, especially for the identification of illegal transaction entities.

Heuristic algorithm 2 (H2): A heuristic algorithm based on transaction input address and output address change address detection heuristic.

In previous studies, there were not many researches on Bitcoin change address recognition alone, because the accuracy of the change address detected by the four constraints proposed by Sarah Meiklejohn is not high enough to be clustered into the same entity with the input address of the transaction. The four constraints for an output address to be determined as a change address are as follows [21]:

1. The address can only be used as the output of transaction once;
2. The transaction that this address participates in is not a Coinbase transaction,that is, each block on the blockchain corresponds to a Coinbase transaction. There is no input address for Coinbase transactions, only output addresses;
3. The output address is different from the input address, that is, it is not a 'self-change' transaction;
4. There is no other address in the output accounts which is different from the change address and only appears once in the blockchain ledger.

It should be emphasised that 'self-change' refers to the fact that in the Bitcoin protocol, the system specifies the change address for the transaction automatically. The common practice is to provide a change address that is the same as the input address. We should eliminate this kind of 'self-change' transaction when we detected the change address.

Consequently, we concluded that the cause of low accuracy of change address clustering mainly lies in:

1. The heuristic condition is based on empirical observation and has strong subjectivity;
2. Bitcoin transaction happens all the time. The first of the four qualifications proposed by Sarah Meiklejohn *et al.* required traversal of the entire blockchain ledger transaction data to ensure the address is only used as the output of a transaction once. This process is quite time-consuming. And the data set that people used to conduct the experiment could not have included the transaction records that were generated during the experiment. People usually tend to take a certain time node as the criterion and only analyse the transaction data within that time period;
3. The change address can be used as input for a subsequent transaction. When a non-change address is associated with the input address of multiple other transactions, the program cannot detect such an error. Then a large number of addresses are clustered through multiple iteration cycles, and they are wrongly assigned to the same entity. This makes it

more troublesome to find and eliminate such false positives when we conduct data inspection in the later stage.

On the basis of previous studies about the detection method of change address, this paper synthesises the limiting conditions of many scholars on the change address. At the same time, we classify the number of the output address of the transaction, and put forward the process of change address identification.

The change address detection algorithm proposed in this paper can be divided into the following two situations for discussion.

1. There are only two output addresses for a transaction.
   a. In the blockchain ledger, address A1 appears only once, while address A2 appears more than once;
   b. The amount of A1 has more than three decimal places than A2.

   If A1 meets the above two conditions, A1 is considered to be a change address.

2. There are more than two output addresses for a transaction.

When the number of output addresses exceeds two, it can be marked as a change address if a certain output address meets the four qualification conditions proposed by Meiklejohn *et al*. The four conditions are described above and will not be repeated here.

In summary, this paper proposed the improved detection process of change address, as shown in Figure 2.

Next, combined with the common input ownership heuristic, the change address is used as a link to connect multiple transactions, and then the correlation clustering analysis is conducted for more addresses.

Heuristic algorithm 3 (H3): A heuristic algorithm based on transaction output address, Coinbase transaction mining address clustering heuristic.

In the Bitcoin system, when a transaction occurs, a node will package the transaction and record it on the blockchain ledger. In this process, the full node is calculated by a random number to obtain the right to billing. When a full node completes transaction packaging first and passes the entire network verification, the node will receive a mining reward. This is a minting transaction belonging to the current block, also known as a Coinbase transaction or a mining transaction. Specifically, in the early days when the Bitcoin system was launched, mining could be divided into two situations: pit mining and single miner mining. However, with the development of technology, the situation of self-mining by users is disappearing gradually, and the trend of mining is evolving towards the emergence of large mining pools. For pool mining, the owner of the pool will gather the miners together and use a full node to drive a number of mining machines, in which each miner is only responsible for calculating the hash value, and the income from mining belongs to the owner. In the later stage, the owner will conduct secondary distribution of the income according to the proof of work of each miner. As a result, more and more miners are inclined to join the mining pool to reduce the financial and time costs and obtain more stable income. To sum up, we can assume that the output address of a Coinbase transaction is controlled by the same entity. An example of a Coinbase transaction is shown in Figure 3.

Heuristic algorithm 4 (H4): A heuristic algorithm based on the number of output addresses, multiple mining pool address clustering heuristic.

The difference between H4 and H3 is that the clustering objects of H4 are for multiple mining pools. The heuristic clustering rule is that if there are more than 100 output addresses in a transaction, and one of the output addresses is known to belong to a certain mining pool, then we assign all output addresses to the mine owner of the certain mining pool [22, 23].

As shown in Figure 4, if the Mining pool 1 in the output address is known to be AntPool, then the Mining income of all other Mining pools belongs to AntPool.

3.2 | Entity relationship recognition

Heuristic algorithm 5 (H5): A heuristic algorithm based on equal currency multi-input multi-output, mixing transaction recognition heuristic.

The anonymity of Bitcoin brings convenience for transactions and protects the privacy of both parties to the transaction at the same time. Although several heuristic algorithms mentioned above can identify the address group controlled by a single entity, this is a judgment without considering the mixing transactions. The mixing service exposes the pseudo-anonymity of Bitcoin. The mechanism of the mixing service is that if someone wants to transfer illegally through Bitcoin, they can use the mixing service to hide the illegal proceeds. Mixing services enable users to mix with other users' funds quickly and efficiently and create random mapping relationships between existing user accounts and new accounts to achieve anonymity. Thus, it is necessary for regulators to identify mixing transactions to narrow the scope of investigation of illegal entities. Judging from the transaction characteristics, we assume that if there are more than four input addresses and output addresses of a transaction, there will be mixing transactions in the transaction [24]. The schematic diagram of mixing service is shown in Figure 5. Four users participate in mixing service and send Bitcoin to the platform. Mixing service platform provides four addresses to receive Bitcoin, and four users provide the receiving addresses after the mixing service is completed, respectively. As long as the platform does not use the same address to receive and return Bitcoin, it cuts off the flow of funds and realises the requirement of anonymity. Here, the mixing service platform uses four addresses to receive Bitcoin, but in fact, the current mixing mechanism can use one address to receive all inputs in the same transaction, increasing the effectiveness of service.

Heuristic algorithm 6 (H6): A heuristic algorithm for identifying relationships between Bitcoin entities, Louvain community detection algorithm.

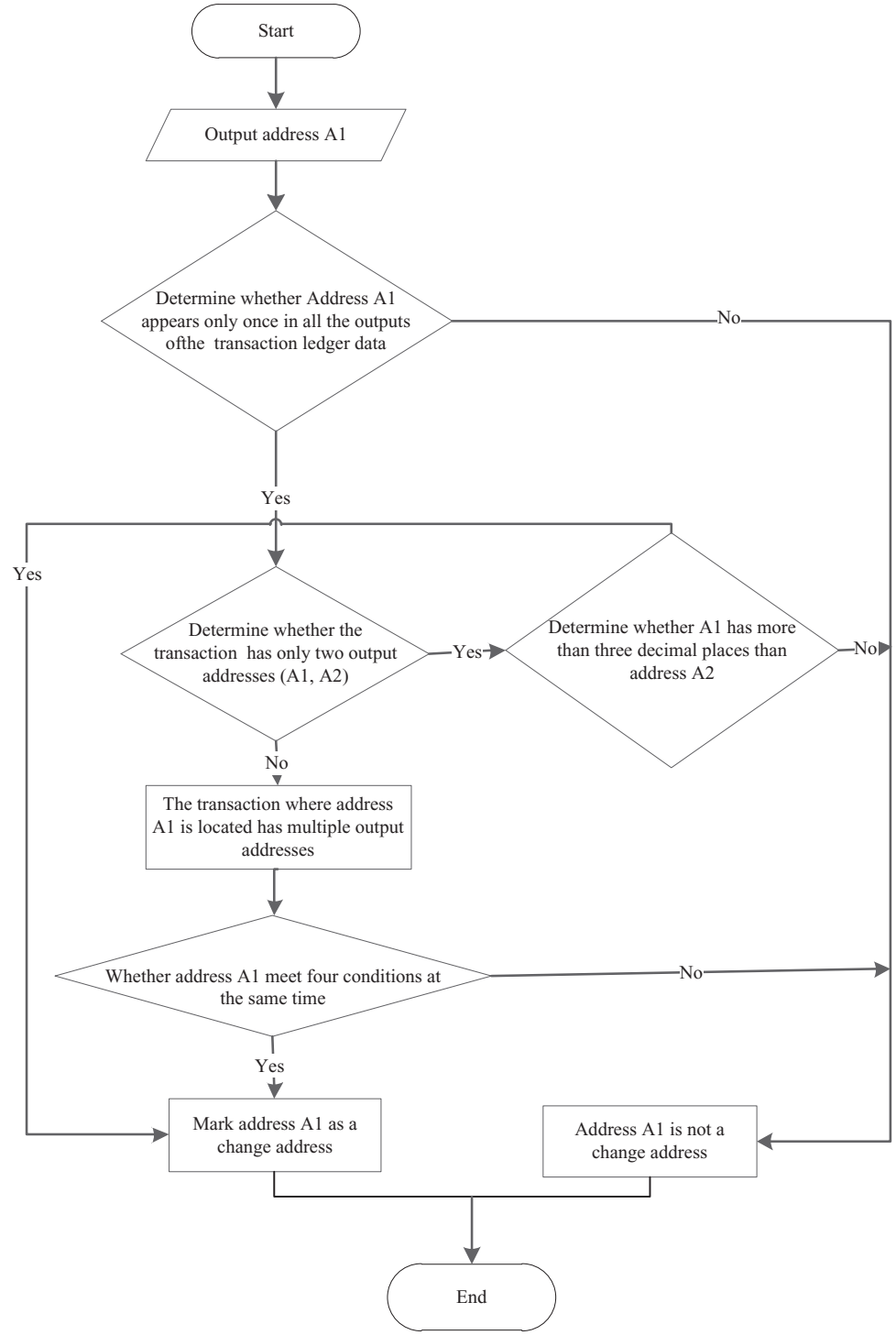

FIGURE 2 Improved Change address detection process

Louvain algorithm [25] is considered to be one of the best-performing community discovery algorithms. Louvain community detection algorithm can find highly modular partitions for large networks in a short time, and present a complete hierarchical community structure for complex networks. Louvain algorithm is a modularity-based community discovery algorithm. In terms of its recognition efficiency, Louvain-based community division method can complete community recognition for a complex network containing hundreds of millions of transactions within 150 min. Considering the characteristics of blockchain Bitcoin transaction network, such as high complexity, large scale and large number of transactions, as well as the good performance of Louvain community detection algorithm in identifying large complex networks, this paper uses Louvain

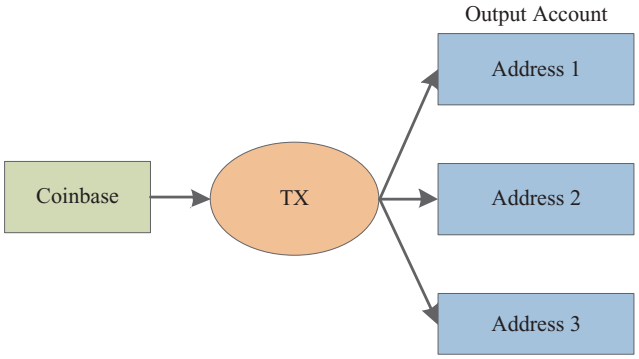

FIGURE 3 Coinbase schematic diagram

algorithm to divide the community of Bitcoin transaction network. Therefore, based on a large number of address groups clustered by the four heuristic algorithms (H1 to H4), we used Louvain community detection algorithm to further explore the relationship between entities, such as group activities between two different entities through intermediary, so as to achieve the purpose of community division of the transaction entity.

1. Evaluation model
   - Modularity

Louvain algorithm uses modularity Q to evaluate the quality of a community network division [26]. The physical meaning of modularity is the difference between the number of connected edges of nodes in the community and the number of edges under random conditions. And $Q = 0.3$ is generally taken as a measure of obvious community structure in the network. Q can be calculated as follows:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta\left(c_i, c_j\right),$$

Where A_{ij} represents the weight of edges between i and j. When the network is not a weighted graph, the weight of all edges can be regarded as 1. $k_i = \sum_j A_{ij}$ refers to the sum of the weights of all the edges connected to node i. $m = \frac{1}{2}\sum_{ij} A_{ij}$ represents the sum of the weights of all the edges in the network. c_i is the community to which node i belongs. Function $\delta(c_i, c_j)$ indicates that if node i is in the same cluster as j, the return value is 1, otherwise returns 0.

   - The gain of modularity ΔQ

When a new node joins the community, for example, when node i is assigned to community c where neighbour node j is located, the Louvain algorithm will recalculate the modularity of the community. ΔQ can be calculated as follows:

$$\Delta Q = \left[\frac{\sum_{in} + k_{in}}{2m} - \left(\frac{\sum_{\text{tot}} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right].$$

$\sum_{in}$ represents the sum of the weights of all edges in the current community. $\sum_{\text{tot}}$ represents the sum of the weights of all external edges connected to the community C. For a single node, the sum of the weights of the edges connected to the external community is equal to the sum of the weights of all the edges connected to it, that is, k_i.

2. Iterative process

Step 1: Each node in Figure 6 is treated as an independent community. Corresponding to the Bitcoin network, the address group belonging to the same entity is regarded as an independent community, and the initial number of communities is the same as the number of nodes.

Step 2: For each node i, we try to assign the node i to the community where each of its neighbour nodes is located, calculate ΔQ before and after the assignment, and record the neighbour node with the largest ΔQ. If Max $\Delta Q > 0$, node i is assigned to the community where the neighbour node of ΔQ is the largest; otherwise, it remains unchanged.

Step 3: We repeat step 2 until the communities of all nodes no longer change.

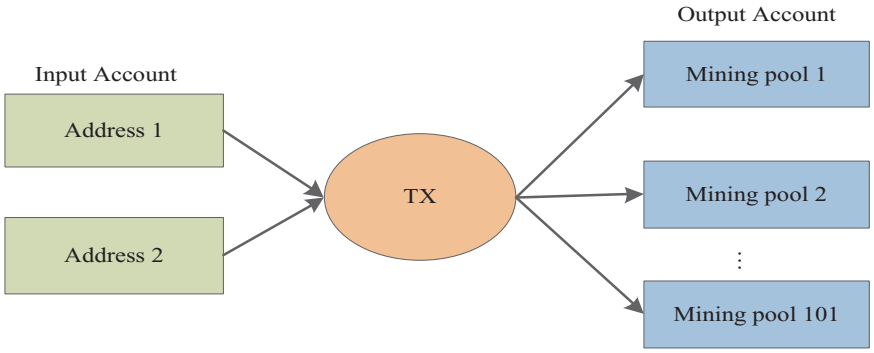

FIGURE 4 Multiple mining pool address cluster schematic diagram

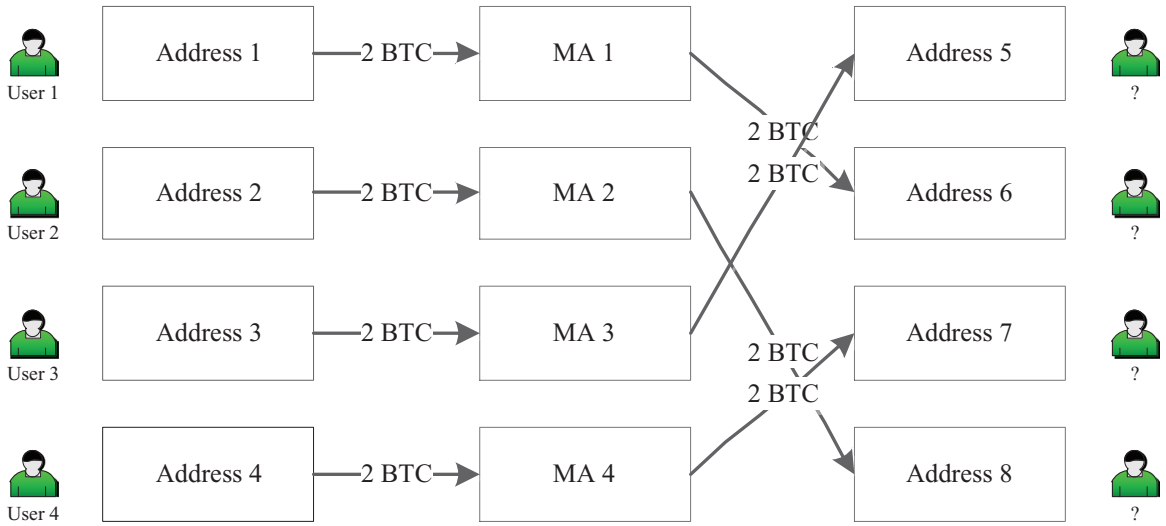

FIGURE 5 Mixing service transaction schematic diagram

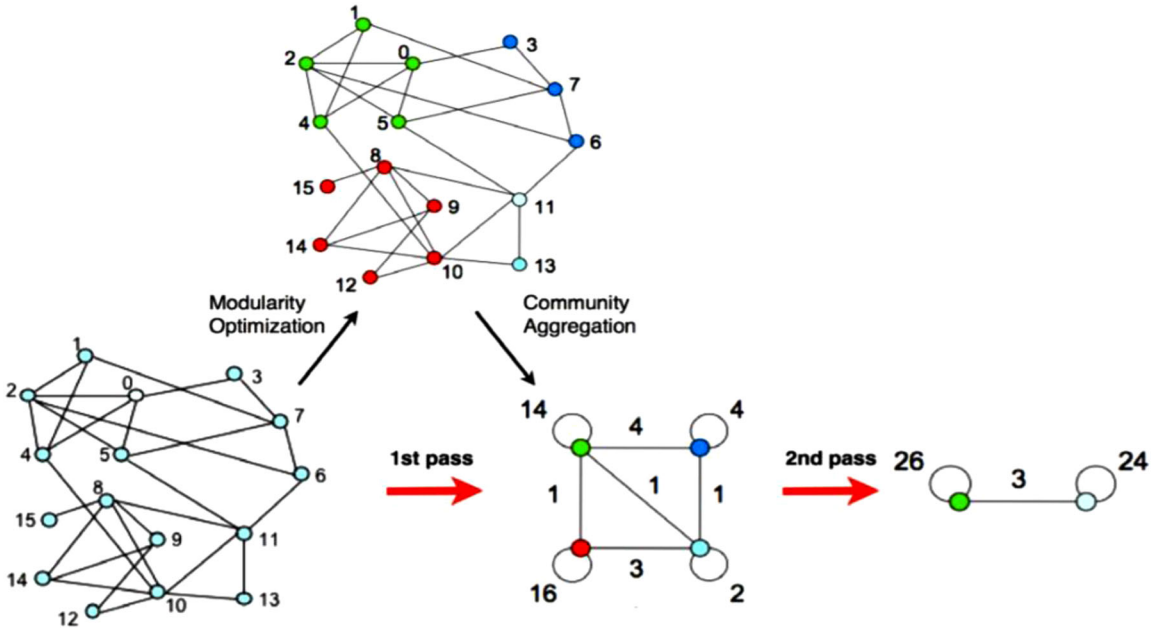

FIGURE 6 Visialisation of the steps of Louvain Method

Step 4: We compress all nodes in the same community into a new node and recalculate the weights of edges between nodes.

Step 5: We repeat step 1 until there is no change in the modularity of the entire diagram.

We use the clustering results of H1, H2, H3 and H4 methods as the input of H6.

Phase 1: Groups of addresses belonging to the same entity were clustered by H1, H2, H3, and H4 used as nodes of the network. At this point, a community corresponds to an entity.

Phase 2: To consider the relationship among the entities in the experimental data set, we added an edge between the entities if they meet the following two conditions [27].

- There are less than 10 entities in the output of the transaction;
- All recipients are different from the sender.

4 | EXPERIMENTS AND RESULTS

In this section, we explained the experimental environment and data sources. We discussed the specific process of the

experiment and analysed the results of the experiment. This included the results of the address aggregation analysis, the effectiveness of the change address detection algorithm, and the entity relationship recognition results.

4.1 | Experimental environment and data preparation

The experimental data used in this experiment is divided into two parts. One is the Bitcoin addresses to be clustered, and the other one is all transaction records in which these addresses participate. Firstly, we designed a web crawler to crawl Bitcoin addresses. We obtained the Bitcoin address filled in by the user from the user interface in the Bitcointalk forum (https://Bitcointalk.org/). The user interface with the Bitcoin address is shown in Figure 7.

As of February 2021, the forum has approximately 3.1 million user registrations. We crawled the information of the first 30,000 registered users since the establishment of the forum and saved it in csv format.

The record of all transactions that a Bitcoin address participates in comes from the blockchain ledger. We extracted the key fields of information such as the transaction hash of the experimental address, the transaction amount, the hash of the transaction record in which the experimental address participated, and the input and output addresses. And we saved the information in json format.

The operating environment of our experiment is shown in Table 2.

4.2 | Experiment process

In this section, we explained the algorithm process of experimental scheme clustering, including the multi-heuristic address clustering algorithm process and entity community division process, respectively.


Summary - Yaunfitda

Name:	Yaunfitda
Posts:	2917
Activity:	1554
Merit:	472
Position:	Sr. Member
Date Registered:	November 25, 2015, 02:45:01 PM
Last Active:	October 27, 2020, 05:20:17 PM
ICQ:	
AIM:	
MSN:	
YIM:	
Email:	*hidden*
Website:	
Current Status:	Offline
Skype:	3KFGnvPnk3mbgs9YobidGwSySCwcEcGw2u
Bitcoin address:	bc1qp5j40v65glz5smf7530setta7yj39l4nf903gx
Other contact info:	1CaBDSUQHzi2HKTy88UVgwWJegvnKZRWWT
Gender:	
Age:	N/A
Location:	
Local Time:	October 28, 2020, 11:41:26 AM

FIGURE 7 Profile page in Bitcointalk

TABLE 2 Introduction of operation environment

Experimental operating environment	
CPU	AMD Ryzen 5–2500U
RAM	8 GB
Operating system	Windows 10
Programming language	Python

4.2.1 | Address clustering process

We combined H1 with the improved change address detection algorithm H2, H3 and H4 to cluster the addresses we collected from the Bitcoin Forum. The clustering algorithm flow is as follows.

Since Bitcoin transactions occur all the time, to reduce the complexity of the experiment, we obtained all transactions involving addresses to be identified in the blockchain ledger (as of 10 September 2020). In the process of detecting the change address in the output of the transaction, we put the change address founded by the program into the list separately for subsequent analysis of the accuracy of the change address.

4.2.2 | Community division process

In the initial stage, we regarded a single entity clustered by H1 + H2 + H3 + H4 as an independent community. The number of initial communities is the same as the number of entities. Then, using the complex network idea, edges are added between entities to form an entity transaction network under the two conditions mentioned above. Then we use Louvain algorithm to divide the community. The algorithmic process for dividing the entity community is as follows.

4.3 | Experimental results

4.3.1 | Address clustering results analysis

We selected an address randomly from the experimental data set for analysis. We take the address 18yVghacMaDU8SzG487h2eQvj2SaUCbtXj as an example and iterate twice under the methods H1, H1 + H2, H1 + H2 + H3, and H1 + H2 + H3 + H4, respectively. Through experiments, we found that under different heuristic conditions, the number of addresses obtained was 35, 39, 120 and 6469. The result was shown in Figure 8. We plotted them on a logarithmic 10-scale.

From the experimental results, it can be seen that if one more heuristic condition is used, the number of clustered addresses increased as well. This is due to the fact that with each iteration, new addresses associated with the destination address are incorporated into the entity. Also, we need to traverse the

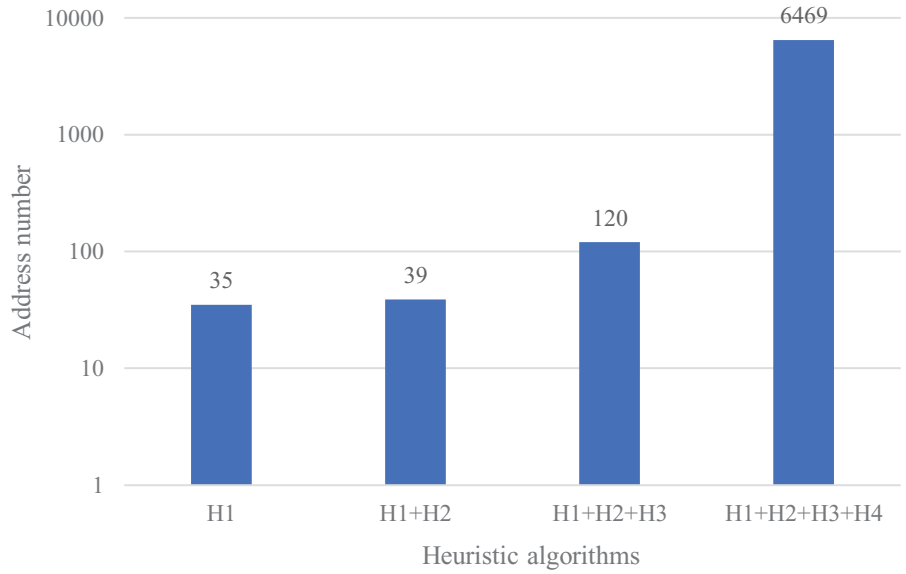

FIGURE 8 Address clustering results

transaction records of these new addresses in order to continue to associate with other addresses in the next iteration. By comparison, the number of clustered addresses in our method after two iterations is far more than the number of addresses marked for the wallet belonging to the experimental address in Wallet Explorer.

In addition, we noticed that many previous works did not explain the clustering rules of WalletExplorer when comparing their experimental results with WalletExplorer. Therefore, we revealed the rules for address clustering of the website. We found that WalletExplorer adopt the first heuristic algorithm to cluster address that we mentioned in the previous section, so the clustering results of this website can only be used to verify the effectiveness of H1. You can find the interpretation from the website (https://www.walletexplorer.com/info). Besides, we found that the site also marked the change address in each transaction (if already present). Meanwhile, we also inferred the rules of the website for marking the change address through experiments, and obtained his confirmation by contacting the Author (Aleš Janda) of the website. The content of this part will be presented in the following chapters.

4.3.2 | Validity analysis of change address detection algorithm

Since there is no perfect standard for the detection of change address, however, based on the previous work, we classified the number of different output addresses of transactions and hope it will be helpful to further improve the accuracy of change address detection. In order to prove whether the change address detection algorithm we proposed is more effective than the previous algorithm, we used two methods to verify.

TABLE 3 Method 1 Experimental results

Target address	WalletExplorer	Our method
18y**j	16	18

TABLE 4 Method 2 experimental results

Algorithm	Change Address
Original change address algorithm	8821
Improved method	5751

Method 1:Compare and analyse the marking results of WalletExplorer.

Through a large number of experimental analysis on the change address marked in WalletExplorer and the data of its transaction, we found that the website's judgment on the change address is based on only one principle, that is, output is a change address if it belongs to the same wallet as sender. We still analyse the transactions involved in the experimental address selected. After iterating twice through H1 + H2, we compared the experimental results with the change address marked in Walltexplorer, as depicted in Table 3. Obviously, our detection results for change addresses are more comprehensive than WalletExplorer's detection results, and the results identified by our method for each address all include the change address marked by WalletExplorer, and the matching rate of the coincident address is 100%.

Method 2: Compare the address reduction rate of algorithms.

We compare the experimental results of the change address between the original change address algorithm and our improved algorithm. The results (Table 4) indicated that the original change address algorithm found 8821 change addresses,

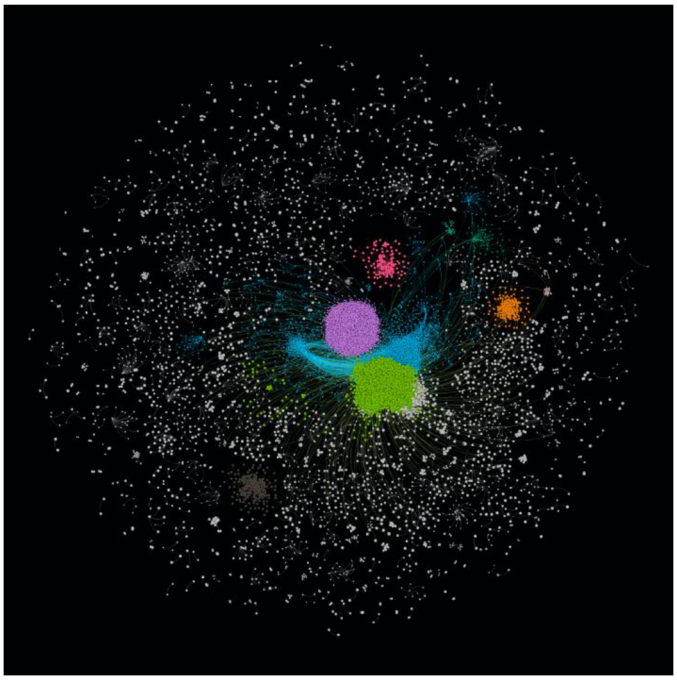

FIGURE 9 Entity transaction relationship diagram

and our integrated algorithm detected 5751 change addresses. The average contribution of our algorithm to the address reduction rate is 34.8%.

By using the two methods above to analyse the effectiveness of change address detection, the experimental results indicated that our integrated algorithm is more effective than the original change address algorithm.

4.3.3 | Entity relationship recognition results

Mixing transaction recognition results

We identified a total of 3252 transactions using the heuristic conditions defined by H5 in the experimental data that may contain mixing services.

Analysis of community division results

Using the Louvain algorithm combined with the complex network, the entities clustered by H1 + H2 + H3 + H4 are further divided into communities. Through experiments, we divided 74,286 entity nodes into 1247 communities. We used Gephi (https://gephi.org/), a complex network analysis software, to map transaction networks to further visualise the relationships between entities. The entities incidence relation diagram is acquired as depicted in Figure 9. Each node in the graph represents an entity, and the connections between nodes represent transactions. The cluster formed by the connection between nodes is a community, and the entities in the same community are coloured with the same color. Due to the memory limitation of the experimental environment, some simple transaction data that are scattered in the entire transaction graph are represented in white uniformly.

5 | RELATED WORK

The anti-anonymity of the Bitcoin system has been the focus of scholars for a long time. People usually analyse the data of the blockchain ledger and combine with the off-chain information to conduct anti-anonymity of the transaction entity.

5.1 | Address relationship analysis

There are many studies about address relationship analysis. As mentioned above, the relationship analysis of Bitcoin addresses mainly focuses on address clustering. At present, the mainstream address clustering method is common-input-ownership heuristic and change address detection method. The common-input-ownership heuristic is valid for transactions other than mixing service transactions. The change address detection method can find the relationship between the output address and the input address and then group them into the same wallet. Specifically, Reid and Harrigan reported the common-input-ownership heuristic [13], and Nakamoto also mentioned the idea that multiple input addresses in a transaction may belong to the same entity [19]. Reid and Harrigan also mentioned the heuristic algorithm for change address detection for the first time in their research. Subsequently, Androulaki *et al.* and Meiklejohn *et al.* also carried out experimental applications and further expansion of this method, respectively [20, 21]. Mao *et al.* proposed an algorithm integrating three kinds of heuristic conditions and designed the clustering scheme [28]. Conti *et al.* studied the relevant transactions involved in ransomware and proposed a lightweight framework to analyse and identify all Bitcoin addresses belonging to the ransomware criminals [29].

5.2 | Community detection algorithm

Community detection algorithm is used for community discover in Bitcoin network, mainly for clustering users and further discovering the potential relationship between users. At present, the common algorithms for community discovery include hierarchical clustering algorithm, modularity optimisation algorithm, and label propagation algorithm, etc. In previous studies, researchers mainly used Louvain algorithm to divide communities. Blondel *et al.* proposed a modularity-based heuristic method, Louvain algorithm, for the first time in 2008, and applied it to extract community structures for large networks.[25] Since many data processing algorithms can be accelerated through randomisation, subsequently, Kirianovskii *et al.* proposed a random version of the Louvain algorithm, which enables the Louvain algorithm to process the complex network with more nodes faster [30]. Based on Louvain's processing principle for complex networks, Remy *et al.* tried to track Bitcoin users with Louvain community detection algorithm, and used the algorithm to re-identify multiple addresses belonging to the same user effectively [27]. On the basis of clustering a large number of addresses, Zheng *et al.* further applied the improved Louvain community detection algorithm to discover the connections between Bitcoin users, and combined address and user clustering to improve the accuracy of entity speculation [22, 31].

6 | CONCLUSIONS AND FUTURE WORK

This paper reviewed and summarised six heuristic algorithms for address clustering and entity relationship analysis on the basis of previous work. The comprehensive use of multiple heuristic methods to cluster addresses is conducive to clustering the addresses of large entities. We analysed the reasons for the low accuracy of the detection results of the existing change address algorithm and improved it. The experimental results indicated that our algorithm is more effective than the original change address algorithm. In addition, we divided the address groups that have been clustered by the first four heuristic algorithms into communities. For future work, we will further expand the experimental data set and consider the conditions of the heuristic algorithm strictly to further improve the accuracy of address clustering. Moreover, we will collect more off-chain information to add more identity dimensions for the transaction entities, so as to achieve the purpose of tracing the transaction entities and some illegal transaction activities.

ACKNOWLEDGEMENT

This work was supported by the National Key Research and Development Program of China [grant No. 2019QY(Y)0601]. The author would like to thank Aleš Janda, the developer of WalletExplorer, for his confirmation and help to this work.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Xi He https://orcid.org/0000-0001-9705-2744

REFERENCES

1. Conti, M., Sandeep Kumar, E., Lal, C., Ruj, S.: A survey on security and privacy issues of Bitcoin. IEEE Commun. Surv. Tutor. 20(4), 3416–3452 (2018)
2. Chang, T.-H., Svetinovic, D.: Improving Bitcoin ownership identification using transaction patterns analysis. IEEE Trans. Syst. Man Cybern. Syst. 50(1), 9–20 (2020)
3. Nerurkar, P., Patel, D., Busnel, Y., Ludinard, R., Kumari, S., Khan, M.K.: Dissecting Bitcoin blockchain: Empirical analysis of Bitcoin network (2009-2020). J. Netw. Comput. Appl. 177, 102940 (2021)
4. Javarone, M.A., Wright, C.S.: From bitcoin to bitcoin cash: A network analysis. In: Proceedings of the 1st Workshop on Cryptocurrencies and Blockchains for Distributed Systems, pp. 77–78 (2018)
5. Chen, Z., Li, C., Sun, W.: Bitcoin price prediction using machine learning: An approach to sample dimension engineering. J. Comput. Appl. Math. 365, 112395 (2020)
6. Wu, J., Liu, J., Zhao, Y., Zheng, Z.: Analysis of cryptocurrency transactions from a network perspective: AN overview. J. Netw. Comput. Appl. 190(C), 103139 (2021)
7. He, X., Zhang, F., Lin, S., Mao, H., He, K.: A review on data analysis of Bitcoin transaction entity. In: 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 159–164. IEEE, Piscataway, NJ (2020)
8. Chen, W., Zheng, Z.: Blockchain data analysis: Status, trends and challenges. Comput. Res. Dev. 55(09), 1853–1870 (2018)
9. Alqassem, I., Rahwan, I., Svetinovic, D.: The anti-social system properties: Bitcoin network data analysis. IEEE Trans. Syst. Man Cybern. Syst. 50(1), 21–31 (2020)
10. Damiano, D.M.F., Marino, A., Ricci, L.: Data-driven analysis of Bitcoin properties: Exploiting the users graph. Int. J. Data Sci. Anal. 6(1), 63–80 (2018)
11. Ferretti, S., D'Angelo, G.: On the Ethereum blockchain structure: A complex networks theory perspective. Concurr. Comput. 32(12), e5493.1–e5493.12 (2020)
12. Lischke, M., Fabian, B.: Analyzing the bitcoin network: The first four years. Future Internet 8(1), 7 (2016)
13. Reid, F., Harrigan, M.: An analysis of anonymity in the Bitcoin system. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 1318–1326. IEEE, Piscataway, NJ (2011)
14. Ermilov, D., Panov, M., Yanovich, Y.: Automatic Bitcoin address clustering. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 461–466. IEEE, Piscataway, NJ (2017)
15. Spagnuolo, M., Maggi, F., Zanero, S.: Bitlodine: Extracting intelligence from the Bitcoin network. Financial Cryptograh. Data Secur. 8437, 457–468 (2014)
16. Di Battista, G., Di Donato, V., Patrignani, M., Pizzonia, M., Roselli, V., Tamassia, R.: Bitconeview: Visualization of flows in the Bitcoin transaction graph. In: 2015 IEEE Symposium on Visualization for Cyber Security (VizSec), pp. 1–8 (2015)
17. Kinkeldey, C., Fekete, J.-D., BitConduite, I.P.: Visualizing and analyzing activity on the Bitcoin network. EuroVis 2017-Posters, pp. 25–27 (2017)
18. Yue, X., Shu, X., Zhu, X., Du, X., Yu, Z., Papadopoulos, D., et al.: BitExtract: Interactive visualization for extracting Bitcoin exchange intelligence. IEEE Trans. Visual Comput. Graphics 25, 162–171 (2019)
19. Nakamoto, S., Bitcoin: a peer-to-peer electronic cash system. Technical Report (2008). https://bitcoin.org/bitcoin.pdf.
20. Androulaki, E., Karame, G.O., Roeschlin, M., Scherer, T., Capkun, S.: Evaluating user privacy in Bitcoin. In: Financial Cryptography and Data Security, pp. 34–51 (2013)
21. Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G.M., et al.: A fistful of Bitcoins. Commun. ACM 59(04), 86–93 (2016)

22. Zheng, B., Zhu, L., Shen, M., Du, X., Guizani, M.: Identifying the vulnerabilities of Bitcoin anonymous mechanism based on address clustering. Sci. China Inform. Sci. 63(3), 99–113 (2020)
23. Lewenberg, Y., Bachrach, Y., Sompolinsky, Y.: Bitcoin mining pools: A cooperative game theoretic analysis. In: Proceedings of International Conference on Autonomous Agents and Multiagent Systems, pp. 919–927 (2015)
24. Athey, S., Parashkevov, I., Sarukkai, V., Xia, J., Bitcoin Pricing, Adoption, and Usage: Theory and Evidence. Research Papers (2016)
25. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008(10), P10008 (2008)
26. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E 70(6), (2004)
27. Remy, C., Rym, B., Matthieu, L.: Tracking Bitcoin users activity using community detection on a network of weak signals. In: International Workshop on Complex Networks and their Applications, pp. 166–177 (2018)
28. Mao, H., Wu, Z., He, M.: Bitcoin address clustering method based on heuristics. J. Beijing Univ. Posts Telecommun. 41(2), 27–31 (2018)
29. Conti, M., Gangwal, A., Ruj, S.: On the economic significance of ransomware campaigns: A Bitcoin transactions perspective. Comput. Secur. 79, 162–189 (2018)
30. Kirianovskii, I., Granichin, O., Proskurnikov, A.: A new randomized algorithm for community detection in large networks. IFAC-PapersOnLine 49(13), 31–35 (2016)
31. Zheng, B., Zhu, L., Shen, M., Du, X., Yang, J., Gao, F., et al.: Malicious Bitcoin transaction tracing using incidence relation clustering. In: International Conference on Mobile Networks and Management, pp. 313–323 (2018)

How to cite this article: He, X., He, K., Lin, S., Yang, J., Mao, H.: Bitcoin address clustering method based on multiple heuristic conditions. IET Blockchain 2, 44–56 (2022). https://doi.org/10.1049/blc2.12014