IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 17, 2020, accepted February 5, 2020, date of publication February 11, 2020, date of current version February 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973291

A Refined Analysis of Zcash Anonymity

ZONGYANG ZHANG[1], WEIHAN LI[1], HAITAO LIU[2], AND JIANWEI LIU[1]
[1]School of Cyber Science and Technology, Beihang University, Beijing 100083, China
[2]Chinese Flight Test Establishment, Xi'an 710089, China

Corresponding author: Zongyang Zhang (zongyangzhang@buaa.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1400702, in part by the National Natural Science Foundation of China under Grant 61972017, Grant 61972018, Grant 61932014, and Grant 61972310, in part by the Beijing Natural Science Foundation under Grant 4182033, in part by the National Cryptography Development Fund under Grant MMJJ20180215, and in part by the Beihang University Innovation & Practice Fund for Graduate under Grant YCSJ-02-2019-011.

ABSTRACT With the continuous development and popularity of blockchain technology, anonymity of cryptocurrency has attracted wide attention. Zcash is an altcoin of Bitcoin aiming to protect blockchain anonymity. Its anonymity is highly guaranteed by zero-knowledge proofs. However, it is still practicable to decrease Zcash's anonymity. In this paper, we provide a refined empirical analysis of Zcash anonymity. We improve current address clustering methods and increase the clustering rate by 9%. We also analyze the whole process of distributing mining reward and identify 87.5% addresses and 25.7% transactions. Besides, we simplify Zcash transaction network and then pick out nodes (edges) which play important roles in network connectivity. We show that these nodes are mostly mining pools. In particular, users participating in shieldedpool are mostly founders, miners and mining pools, although shieldedpool itself is designed for protecting anonymity of users with high privacy requirements. Our results, to an extent, are opposite to the original intention of Zcash.

INDEX TERMS Zcash, anonymity, transaction network, address clustering.

I. INTRODUCTION

Bitcoin [1] is a peer-to-peer digital cash system proposed by Nakamoto in 2008. The entire transaction history of Bitcoin is stored in a distributed public ledger denoted as blockchain. Bitcoin system guarantees the pseudonymity [2] of transactions in two aspects. Firstly, the addresses, in form of hashed cryptographic keys, used for sending and receiving BTCs, are created pseudo-randomly. Secondly, one user can create any number of Bitcoin addresses in order to protect his identity. However, a series of previous studies [3]–[7] indicate that the anonymity in Bitcoin system can be greatly reduced. It is possible to track Bitcoin transaction flow, cluster different Bitcoin addresses belonging to the same user and match Bitcoin addresses to users' real identities.

Several techniques are proposed to improve the anonymity of Bitcoin, such as mixing services [8] and joint transaction [9]. A series of altcoins have also been created to improve anonymity such as Dash [10], Monero [11] and Zcash [12]. Among these altcoins, Zcash has its own unique advantage. Zcash's anonymity relies on *shieldedpool*,[1] where partial transaction information such as input/output addresses and transaction value is no more directly available from blockchain compared with Bitcoin. The theoretical basis for shieldedpool is practical zero-knowledge proofs called zk-SNARKs [12].

Several researchers [13]–[15] consider Zcash anonymity in practice. On the one hand, they use similar methods in Bitcoin to analyze Zcash, mainly aiming at transactions between t-addresses (addresses not related to shieldedpool). They cluster and tag addresses, then match them with actual identities. On the other hand, researchers study how to use shieldedpool for deanonymization. After establishing some cluster heuristics related to shieldedpool, they investigate how coins are deposited into and withdrawn from shieldedpool. However, current address clustering methods only consider part of all transaction types. Some other Bitcoin deanonymization methods not used in Zcash before are also suitable for investigating Zcash anonymity. Thus, in this paper we improve the deanonymization results by considering more transaction

The associate editor coordinating the review of this manuscript and approving it for publication was Jiafeng Xie.

[1] Users may choose whether or not to use shieldedpool in a transaction. If not, then the transaction is purely in *valuepool*. Details are shown in Section II.

types and using more deanonymization methods such as user behavior identification and complex network analysis.

A. OUR CONTRIBUTION

In this paper, we give a refined analysis of Zcash anonymity and improve current Zcash deanonymization results. The main contributions of this paper are as follows:

1) We improve address clustering methods and take more transaction types into account. We propose a refined address clustering heuristic and coalesce 36% addresses into multiple entities. The clustering rate is increased by 9% compared to previous research [15].
2) We study the whole process of mining reward distribution. We focus on intermediate transactions that have not been thoroughly studied before and obtain improved results. We discriminate 87.5% of all addresses involving in this process and 25.7% of all transactions serving for it.
3) We build a transaction network and analyze its basic topological properties such as degree, clustering coefficient and Pagerank. We conclude that it is a heterogeneous and sparse network, which is consistent with the actual trading situation. Furthermore, we also simplify this transaction network according to results in 1) and 2). We compare relevant properties and pick out important nodes (edges) using the new simplified network. Note that few studies focus on the topological properties of transaction network itself and deanonymization on the level of complex network.
4) We find that users participating in shieldedpool are mostly founders, mining pools and miners. This is opposite to the intention of Zcash where shieldedpool is designed to transfer coins with high privacy need. To the best of our knowledge, this is the first detailed research on identity and proportion of users inside shieldedpool.

B. ORGANIZATION

In Section II, we introduce how Zcash works. In Section III, we analyze the general statistics of Zcash blockchain. We give our deannonymization methods and results in Section IV. Section V gives the conclusion.

C. RELATED WORK

1) DEANONYMIZATION OF BITCOIN

There have been multiple studies focusing on deanonymizing Bitcoin transactions. Reid and Harrigan [6] firstly analyzed Bitcoin anonymity. They built two types of networks (i.e., transaction network and user network) and analyzed their topological characteristics. However, research only on network topological properties lacks practical significance. Several other researchers explored the development of transaction network and tracked the flow of transactions [3], [6], [7]. For instance, Ron and Shamir [7] tracked 364 transactions over 50,000 BTCs and gave a detailed transaction flow analysis. However, research above lacks analysis of connections among different addresses.

Thus, some researchers applied clustering heuristics [5], [7], [16] which cluster different addresses belonging to the same user. One common assumption is multi-input heuristic, which means all the input addresses of one transaction belong to the same user. Another assumption is change heuristic, which means input addresses and change addresses in a single transaction also belong to the same user. Androulaki *et al.* [16] applied the above two heuristics to build an anonymity attack model and made an experiment in a college. They found that the clustering simulation results of the model were close to the actual situation.

Besides, there also exists TCP/IP layer analysis. Koshy *et al.* citeK:AAB:14 analyzed the matching relationship between Bitcoin addresses and IP addresses. The main idea is that the first node to inform the receiving node of a transaction is the source of the transaction.

2) DEANONYMIZATION OF ZCASH

Since Zcash was proposed as an altcoin of Bitcoin, many researchers in Zcash borrow techniques from Bitcoin research. For instance, Kappos *et al.* [15] ran multi-input heuristic to analyze the deanonymization of Zcash, although this heuristic is only appropriate between transparent transactions. Several studies are unique in Zcash, aiming at the deanonymization of transactions related to shieldedpool. Jeffrey [14] found a common regularity of transactions related to shieldedpool. This regularity is performed as round-trip transactions (RTT for short). First, coins are sent from a t-address[2] to a z-address. Shortly afterwards, coins with the same or very similar value (usually with a gap of common fee value) is moved from shieldedpool back to valuepool. Jeffrey believes that the two t-addresses of RTT are likely to be controlled by one entity. It is found that 31.5% of the coins sent to shieldedpool may be involved in RTT, and this regularity is likely due to the behavior of miners and mining pools.

Alex and Daniel [13] analyzed how mining rewards are distributed from mining pools to miners. In Zcash, mining rewards need to be put into shieldedpool before being given to miners. They found two main patterns of paying mining rewards to miners. In the first pattern, called Pattern T, reward is moved from a z-address to a t-address of the mining pool, and then distributed to miners. In the second pattern, called Pattern Z, reward is distributed directly from z-addresses to miners' t-addresses. By analyzing Pattern T and Pattern Z, Alex and Daniel identified 96% of mining rewarding transactions.

Kappos *et al.* [15] proposed a new heuristic in Zcash, linking t-addresses and z-addresses. The main idea of this heuristic is from change heuristic. They identified and classified

[2] The t-addresses can be thought as addresses applied in valuepool and z-addresses can be thought as addresses applied in shieldedpool. The detailed explanation of t-address and z-address is in Section II.

TABLE 1. In-address and out-address represent input and output address, respectively. In-infor and out-infor refer to input and output information, respectively. The notation ∘ represents that the information of current grid is attainable, and × represents the information is unattainable.

in-address	out-address	transaction-type	in-infor	out-infor
t	t	Transparent (t-t)	∘	∘
t	z	Shielded (t-z)	∘	×
z	t	Deshielded (z-t)	×	∘
z	z	Private (z-z)	×	×
tz	tz	Cross (tz-tz)	×	×

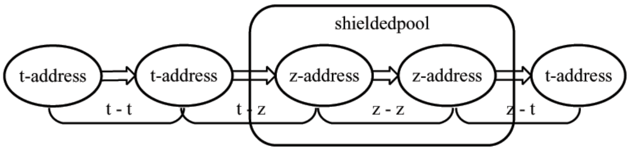


FIGURE 1. Types of Zcash transaction. ***t*-*t*** transaction denotes a transaction from ***t***-address(es) to ***t***-address(es). ***z***-addresses in shieldedpool are not available.

various participants in Zcash, analyzed the transaction characteristics and gave an in-depth analysis of all interactions with (and within) shieldedpool.

II. BACKGROUND

In this section, we introduce how Zcash works. Zcash was launched on 29th October, 2016 [17]. The currency in Zcash blockchain is called ZEC. Since the original version of Zcash is planned to be a fork of Bitcoin, the structure of transactions in Zcash is similar to Bitcoin. There are two types of addresses in Zcash. One is transparent address and the other is shielded address. Transparent transactions (i.e., the sending and receiving addresses are both transparent addresses) are nearly the same as transactions in Bitcoin. That is, one can easily obtain transaction information such as value, fee, input (output) number and senders' (receievers') addresses from blockchain. As the public keys of these transparent addresses always start with a letter t, we denote them as t-addresses.

In order to protect anonymization, shielded address is used in Zcash system. As public keys of these shielded addresses always start with a letter z, we refer to these addresses as z-addresses below. Next we explain how transactions with z-address become "shielded". z-addresses are not exposed in blockchain and the coins sent to or received by a z-address are also not revealed. In addition, any number of t-addresses and z-addresses are permitted in one transaction.

In Table 1, we distinguish several kinds of transactions in Zcash system. The t-t transactions, as mentioned before, are nearly the same as those in Bitcoin. However, in a z-t transaction where all the input addresses are z-addresses and all the output addresses are t-addresses, one can only collect little input information from the blockchain as the input address of this transaction is "null", the number of input addresses is "zero" and the value of unspent inputs is also "zero". Similarly, in a t-z transaction, output information is hard to attain and in a z-z transaction both input and output information is unattainable. A simple view of transaction types in Zcash is shown in Figure 1.

Although the value of ZECs sent to or received by a particular z-address is not attainable, the variation of coins after a transaction can be obtained. This is why *valuepool* and *shieldedpool* are brought in. Shieldedpool describes the value variation of z-addresses and valuepool describes the value variation related to t-addresses. In detail, in a t-t or z-z transaciton, the value of valuepool and shieldedpool will not change. In a t-z transaction, the value of valuepool will decrease and the value of shieldedpool will increase. Thus, the t-z transactions may be vividly considered as putting ZECs from valuepool to shieldedpool. Similarly, the z-t transactions can be thought as transferring ZECs from shieldedpool into valuepool. Two parameters V_{pub}^{old} and V_{pub}^{new} are used in blockchain script to describe the value in valuepool and shieldedpool. V_{pub}^{old} (V_{pub}^{new}) means the value of valuepool before (after) the operation of the current transaction. Then the variation of shieldedpool's value V_{sld} can be obtained by Equation (1),

$$V_{sld} = V_{pub}^{old} - V_{pub}^{new} \tag{1}$$

A. PARTICIPANTS IN ZCASH

Zcash's main users include founders, miners, and mining pools formed by a number of miners. Each coinbase transaction generates about 12.5 ZECs, of which 2.5 ZECs are returned to founders and 10 ZECs are distributed to the miners or mining pools as rewards for generating blocks. Block rewards will be halved every several years. We emphasize that in Zcash's protocol [17], new coins must be put into shieldedpool before subsequent transactions are executed, which, to an extent, strengthens the anonymity.

III. ANALYSIS OF BLOCKCHAIN DATA: POOR USE OF SHIELDEDPOOL

In this section, we give a general analysis of Zcash Blockchain. We download Zcash blockchain, and mainly use Python to achieve data processing. We collect blockchain data from Oct 29th, 2016 to Feb 28th, 2019. The total value of Zcash at that time is 5,258,353 ZECs. There are more than 474,822 blocks, about 4,000,000 transactions and about 200 GB of transaction data.

We pay special attention to shieldedpool, as this is the main difference between Bitcoin and Zcash. Recall that all coins in Zcash are assigned in valuepool and shieldedpool. A comparison of the total value in valuepool and the total value in shieldedpool is shown in Figure 2. The total value of valuepool increases basically at a linear rate due to the continuous generation of new blocks. However, its peak total value is only 366,417 ZECs, which accounts for only 6.9% of the total value. Therefore, we believe that few users use shieldedpool.

We then further investigate different types of transactions in Zcash. The total number of each transaction type is listed

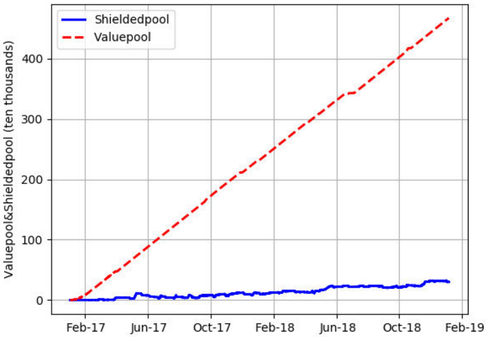


FIGURE 2. The total value of shieldedpool and valuepool.

TABLE 2. The total number and percentage of each transaction type.

Type	Number	Percentage
Transparent (t-t)	2,819,884	76.3%
Coinbase	383,990	10.3%
Shielded (t-z)	206,617	5.6%
Deshielded (z-t)	262,026	7.1%
Cross (tz-tz)	16,141	0.4%
Private (z-z)	14,066	0.3%

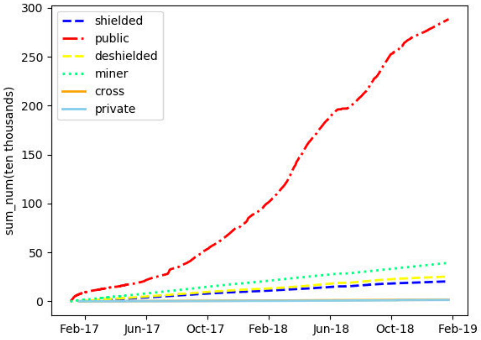


FIGURE 3. The total number of each transaction type.

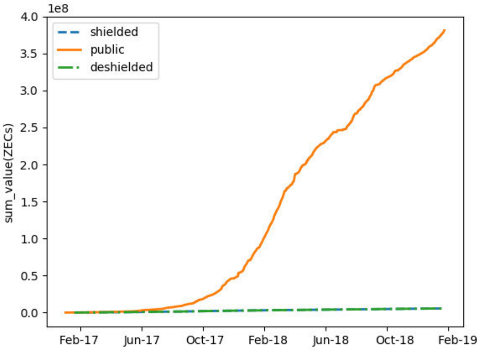


FIGURE 4. The total value of each transaction type.

in Table 2. We find that the majority types of transactions are transparent transactions and coinbase transactions, accounting for 76.3% and 10.3% of all transactions, respectively. None of them is related to z-address and shieldedpool. Only 13.6% transactions include z-address. This can be further obtained in Figure 3 and Figure 4 . It seems that most transactions including z-addresses are shielded and deshielded transactions instead of private transactions.

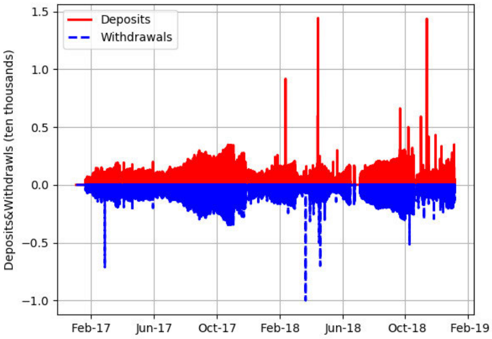


FIGURE 5. All deposits and withdrawals related to shieldedpool.

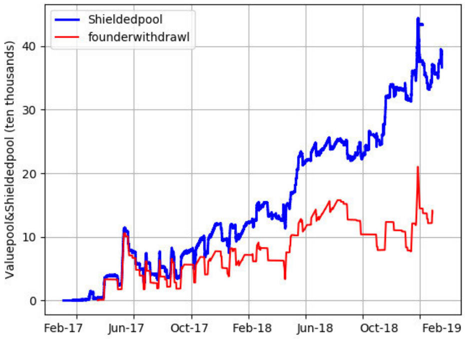


FIGURE 6. Total transaction value of shieldedpool operated by founders.

Note that the shielded transactions and the deshielded transactions are close in terms of number, number percentage and total input value. This may indicate that after ZECs are moved into shieldedpool, they are withdrawn within a few hours. Similar analysis is available in previous research [14] and this kind of ‘‘deposit and withdrawal’’ mode is called ‘‘RTT’’ (Round-trip transactions). Figure 5 gives details of RTT. According to [18], 31.5% of all the transactions related to shieldedpool belongs to RTT.

However, identities of users inside shieldedpool lack further research. Previous work paid more attention to ‘‘who is in the shieldedpool’’ but we focus on ‘‘the proportion of different members in shieldedpool.’’ This question is naturally drawn up by an interesting experiment below. From Figure 6, we observe the total value of shieldedpool over time (the blue and thicker line). The red and thinner line represents the total value operated by founders according to the heuristic in previous research [15]. We find that at the early stage of Zcash, ZECs involved in shieldedpool are almost operated by founders. However, as time goes by, the disparity between these two lines gradually widens. This means there are other entities contributing the value of shieldedpool.

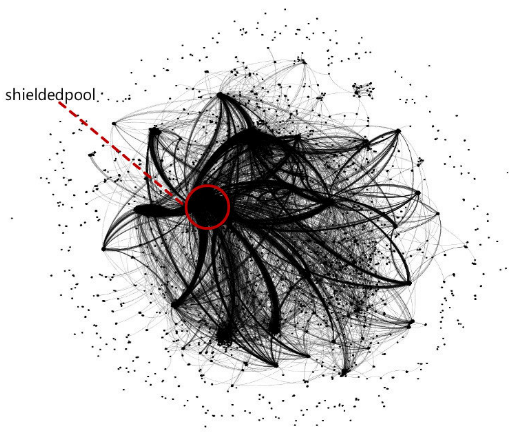


FIGURE 7. The general structure of TN. The size of each node is proportional to its degree. The largest node represents shieldedpool and connects with many addresses with large degree due to frequent deposits and withdrawals.

We give a possible explanation of this "disparity" in Section IV-E.

IV. OUR WORK

We present our deanonymization results in this section. In Section IV-A, we build a transaction network and analyze its topological properties. We introduce the new clustering heuristic in Section IV-B and analyze the whole process of mining reward in Section IV-C. We simplify the transaction network in Section IV-D and give conclusions in Section IV-E.

A. BUILDING TRANSACTION NETWORK

We choose 1,000 blocks from height 29400 to 29500 and build a transaction network using Gaphi. Note that we do not focus on the whole Zcash blockchain as the data processing will be greatly slowed down particularly in the establishment and visibility of our transaction network. Besides, there are also deanonymization research on partial blockchain data [19]. In fact, due to the large number of users and transactions in Zcash, a sample of data is also quite representative.

The transaction network is built as follows. Every t-address is seen as a node. If one node acts as input and another node acts as output in a transaction, then a directed edge is established between these two nodes. Considering that one transaction may include multiple nodes, there may be multiple edges in one transaction. Due to the invisibility of z-addresses, it is hard to refer to these addresses as nodes. However, the in&out information of shieldedpool itself is available. As all z-addresses are in shieldedpool, we consider shieldedpool as a unique node, representing the set of z-addresses.

By applying the rules in the last paragraph, we obtain a transaction network with 98,554 nodes and 183,771 edges. We call this network TN. In Figure 7, we show a general structure of TN, and the biggest node is shieldedpool. The degree of a node in a network is the number of connections to other nodes. TN has an average degree of 3.7, which indicates that one address has connection with 3-4 addresses in average. Pagerank[3] is used to measure the relative importance of network nodes [20]. Figure 8 shows the top 10 nodes of degree and Pagerank in TN. Shieldedpool has a degree of over 70,000, indicating an important role in connectivity. Meanwhile, nodes with top 10 degree and nodes with top 10 Pagerank are the same. Besides, although these 10 nodes only account for 0.01% in number, they contribute 33.3% edges of all the network. These two aspects both imply that in TN, the connectivity and importance of nodes have a positive correlation to some extent. This means nodes with high connectivity tend to be more important.

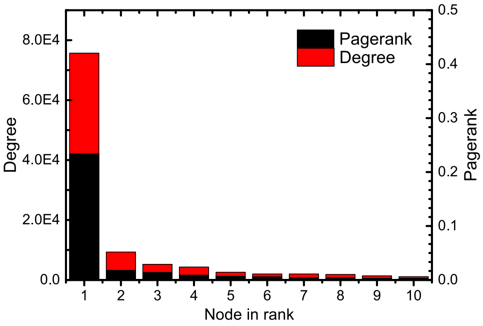


FIGURE 8. Top 10 nodes with highest degree and Pagerank.

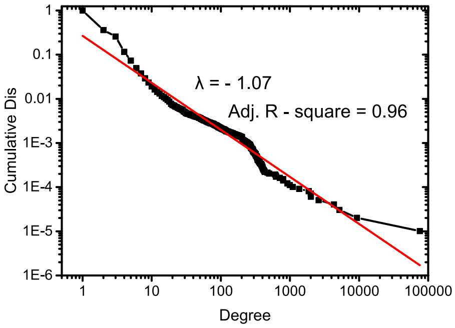


FIGURE 9. The cumulative degree distribution of TN.

A clustering coefficient is a measure of the degree to which nodes in a network tend to cluster together [21], [22]. TN has an average clustering coefficient of 0.185, which means it is a sparse network. Roughly speaking, even if an address $addr_A$ is involved in two transactions at the same time, there is often no transaction between the addresses connected with $addr_A$ in two transactions. For example, in transaction t_1, $addr_A$ and $addr_{t1}$ connect. In transaction t_2, $addr_A$ and $addr_{t2}$ connect. A low clustering coefficient means that there is often no transaction between $addr_{t1}$ and $addr_{t2}$. The cumulative

[3] We use a simplified version of Pagerank adapted from [20]. Let u be a node in a complex network and $R(u)$ be the Pagerank of node u. Then let $\mathbf{Deg}_u^{out}$ be the set of nodes u points to and $\mathbf{Deg}_u^{in}$ be the set of nodes which points to u. Let $N_u = |\mathbf{Deg}_u^{out}|$ be the number of links from u and let c be a factor used for normalization. Then $R(u) = c\sum_{v\in\mathbf{Deg}_u^{in}} R(v)/N_v$.

degree distribution of TN is shown in Figure 9. We find that the degree distribution of TN network basically presents a power-law distribution ($P(X \geq x) \propto x^{\lambda}$), which is very similar to many complex social networks [23].

In conclusion, TN is a heterogeneous network with power-law degree distribution and low clustering coefficient, where a few major addresses play crucial roles. However, we can only obtain network topological properties from the current TN and it is hard to use these properties to link Zcash addresses with users' identities. Therefore, other deanonymizing methods are needed to simplify the network, extract important nodes (edges) and deanonymize users. We will show them shown in the following sections.

B. ADDRESS CLUSTERING

In Zcash, there are two main address clustering methods. One is multi-input heuristic (Heuristic 1) and the other is change heuristic (Heuristic 2). Heuristic 1 holds because a sender, who knows the private key corresponding to each input user's public key, would not reveal his private keys to others [5], [16]. Therefore, input addresses in a transaction might be linked. Heuristic 2 means that when a sender would not put all his ZECs into shieldedpool, he might transfer part of them to a t-address. So the sending address and this t-address might be linked.

Heuristic 1 (Multi-Input Heuristic) [15]: If two or more t-addresses are inputs in the same transaction (whether that transaction is transparent, shielded, or cross), then they are controlled by the same entity.

Heuristic 2 (Change Heuristic) [15]: If one (or more) address is an input t-address and a second address is an output t-address in the same tz-tz transaction, then if this is the only transparent output address, the second address belongs to the same user who controls the input addresses.

We apply these two heuristics and get 735 entities including 26,406 addresses. So the clustering rate is 27%, close to the result 26% in previous research [15]. We emphasize that Heuristic 1 contributes the majority of entities but Heuristic 2 only contributes 11 entites and 34 nodes. In fact, Heuristic 2 only involves shielded transactions. If we take other transaction types into account, the clustering rate might be improved.

We improve Heuristic 2 based on the observation on transaction data. For example, a transparent transaction has one input address and two output addresses.[4] All of them are t-addresses and the transaction value is 566.355519 ZECs. Considering the two output addresses, one received 566.2228206 ZECs and the other one received 0.1326927 ZECs. That is to say, the fee of this transaction is 6×10^{-6} ZECs. This fact strongly indicates that the second address is a change address. This is because the two output value have a huge gap and it is hard for one single account to meet the above two conditions at the same time. The transaction only has one input and two outputs. This strengthens our guess as two outputs usually mean the transaction is pure value-transferring instead of functional ones such as mixing service or procedural ones such as mining reward distribution. Our variable change heuristic (Heuristic 3) is based on this special circumstance.

TABLE 3. Entities with top 10 number addresses.

Rank	Number	Rank	Number
1	9321	6	1233
2	4802	7	972
3	2183	8	645
4	1992	9	635
5	1657	10	550

Heuristic 3 (Variable Change Heuristic): If a transparent transaction has one input and two outputs, and the value of one of the two outputs is more than 20 times that of the other, then the address with smaller value is the change address.[5]

After applying Heuristic 3, we obtain a group of 4,472 entities, among which 580 entities (13% of all) are the same as the result by applying Heuristic 1. This result, to some extent, reflects the reliability of Heuristic 3.

We then merge the above two entities obtained after applying Heuristic 1,2 and Heuristic 3, respectively. If two entities have a common node, then nodes in these two entities are merged into one entity. Repeat this process until there are no duplicate nodes in any two entities. Finally, we get 593 entities with 36,169 nodes, increasing the clustering rate from 27% to 36%. The top 10 large entities is shown in Table 3. The top 10 entities include 23,990 nodes, accounting for 66% of all the clustering nodes, which once again indicates that Zcash transaction network is a highly heterogeneous network.

C. IDENTIFING THE WHOLE TRANSACTION PROCESS OF MINING REWARD

Currently, every coinbase transaction generates 12.5 ZECs. Among them, 2.5 ZECs is transferred to founders and 10 ZECs is allocated to miners as reward. Similar to Bitcoin, miners often gather together to form a mining pool and mining rewards are distributed to its miners. According to Zcash protocol, the mining reward must be put into the shieldedpool before use [17]. This actually gives us another way to identify related transactions.

Previous studies present two possible patterns for mining pools to distribute mining rewards [13]. One is Pattern T, which means that after coins are put into the shieldedpool (z-addresses of mining pools). Coins are firstly transferred to t-addresses of mining pools, and then allocated to t-addresses of miners. The other is Pattern Z, where ZECs are directly allocated to miners from the mining pools' z-addresses.

[4] Txid of this transaction is "fffacc59f3dc6e48b50bcc79199ef96ee135fbbd db261f4639fefa1069260136".

[5] We just think the address with smaller value is the change address since one user usually have multiple addresses and we believe if the converse (the larger value is the change) holds, the user may prefer to choose another address with close value for trading.

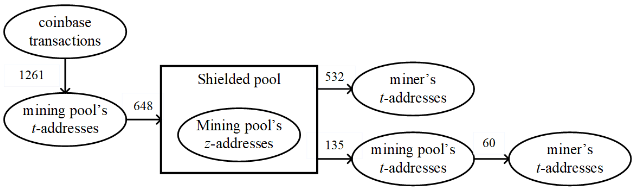


FIGURE 10. **Transactions in mining reward process.**

By applying these two patterns, transactions with miners as receivers can be found [13].

The process of reward distribution is as follows. First, in a coinbase transaction, mining rewards are transferred to mining pools' addresses. Second, rewards are deposited into shieldedpool as required by Zcash protocol [17]. Finally, after a series of intermediate procedures, rewards are distributed to miners. In fact, not only transactions in the last procedure (transactions where miners take as receivers), but also intermediate transactions can also be explored and used for deanonymization. The whole process is shown in Figure 10. Mining rewards are transmitted to mining pool's t-addresses and mining pool's z-addresses. Then it goes in two ways. One directly goes to miner's t-addresses and the other one goes to mining pool's t-addresses and then to miner's t-addresses. Numbers above arrows mean the number of transactions we identify.

We build two lists, including miners list L_m and mining pools list L_p, and then present Heuristic 4. L_p is updated according to the first to third items and L_m is updated according to the second and third items. We repeat the update process until no new address is added to these two lists.

Heuristic 4 (Mining Heuristic): 1. In a coinbase transaction, if the output address with a value about 10 ZECs belongs to a mining pool, then L_p can be updated.

2. If the input of a transaction is a t-address, and its output contains 50 or more t-addresses (the output may have some addresses belonging to L_p), then the output t-addresses except mining pools' t-addresses belongs to miners, and the input t-address belongs to the mining pool. Thus, L_p and L_m can be updated. This is a part of Pattern T.

3. If the input of a transaction is a z-address, and its output contains 50 or more t-addresses (the output may have some addresses belonging to L_p), then the output t-addresses except mining pools' t-addresses belongs to miners. Thus, L_m can be updated. This is a part of Pattern Z.

4. If the input of a transaction is a z-address, and its output is a t-address in L_p, then this transaction is a part of Pattern T.

After running Heuristic 4, we obtain L_p with 44 addresses and L_m with 86,176 addresses. At first glance, we are surprised at so many L_m addresses since the total number of addresses is 98,554 and 36,169 addresses are clustered. This suggests that many entities such as exchanges, services also participate in mining. Besides, we also obtain 1,261 coinbase transactions, 648 transactions from mining pools' t-addresses to shieldedpool, 532 transactions from shieldedpool to t-addresses belonging to miners, 135 transactions from shieldedpool to t-addresses of mining pool and 60 transactions from t-addresses of mining pools to t-addresses of miners. These results are shown in Figure 10.

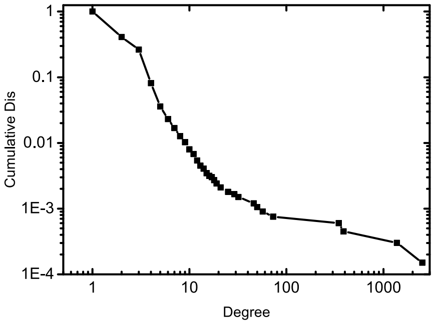


FIGURE 11. **The cumulative degree distribution of TN_S.**

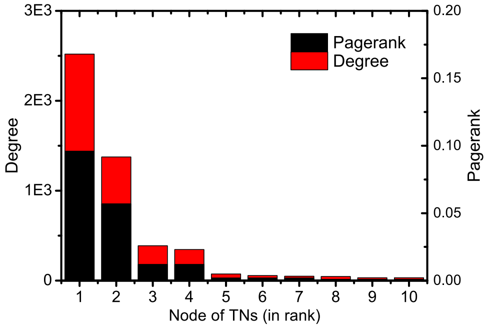


FIGURE 12. **The top 10 nodes with highest degree and Pagerank in TN_S.**

D. A SIMPLIFIED TRANSACTION NETWORK TN_R

In this section, we neglect secondary nodes (edges), extract primary nodes (edges) and construct a simplified network TN_s. We compare the two networks and show that shieldedpool, mining pools and miners instead of common users play a crucial role in Zcash trading. The construction of TN_s is as follows. Firstly, as there is no need to handle transactions inside one entity, we denote an entity as one node. For other users who have transactions with the same user, there is no difference on the activated target address. Secondly, the miner addresses of the miner list L_m not appearing in the clustering results can also be seen as one node since transactions among miners are sparse and less important. Depending on the above two rules, we get a simplified transaction network (TN_s). A general structure of TN_s is shown in Figure 13. We find that shieldedpool, mining pools and miners instead of common users play a crucial role in the network. Besides, the connection between miners and shieldedpool is far less important than the other crucial parts. This indicates that most mining rewards are distributed by mining pools.

A comparison of properties in TN and TN_s is showed in Table 4. TN_s is a network with 6,636 nodes and 8,674 edges, accounting for 6% and 4% of TN. This suggests a great extent of simplification. TN_s has a lower average

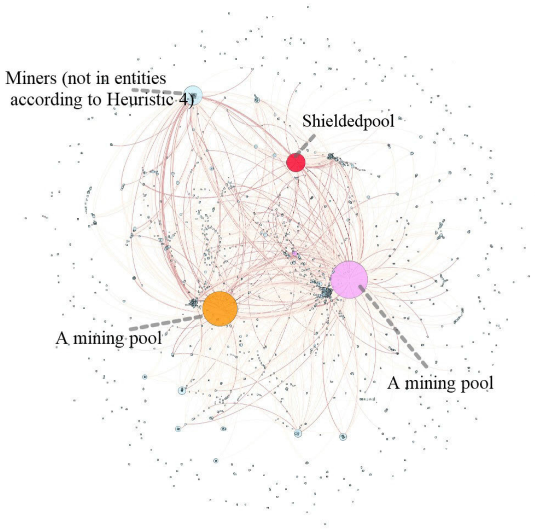


FIGURE 13. The general structure of TN_S. The red node represents shieldedpool. The pink and orange node represent mining pools. The blue node represents the set of miners (according to Heuristic 4) which are not divided into entities. The size of nodes is proportional to Pagerank. The width of edges is proportional to the frequency of edge.

TABLE 4. A comparison of TN and TN_S.

Parameteres	TN	TN_s
nodes	98554	6636
edges	183771	8674
average degree	3.7	2.6
average clustering coefficient	0.185	0.28

degree and higher average clustering coefficient. This means that after our replace and simplification, correlation among nodes is strengthened, which is our target of deanonymization.

Figure 11 describes the cumulative degree distribution of TN_s. The largest degree in TN_s is 2,519, and is much smaller than that in TN. As shown in Figure 12, the nodes in TN_s with top 10 degree and Pagerank are still in high consistency. Nodes with top 10 degree are also with top 10 Pagerank. Most of them consist of addresses belonging to mining pools and miners. That is to say, other entities are less involved. Compared to TN, the much smaller Pagerank indicates that we do highlight the important nodes. Note that 6 of the largest degree nodes in TN are regarded as entities in TN_s, which to some extent reflects the superiority of our address clustering hypothesis. TN_s highlights the important users and transactions and provides a reference for the analysis of the whole transaction network.

E. USERS IN SHIELDEDPOOL

In Section IV, we identify 87.5% nodes regarded as mining pools or miners, which implies that pure users who do not participate in mining only make up a small fraction. We also identify 25.7% transactions in the whole process of mining reward distribution.

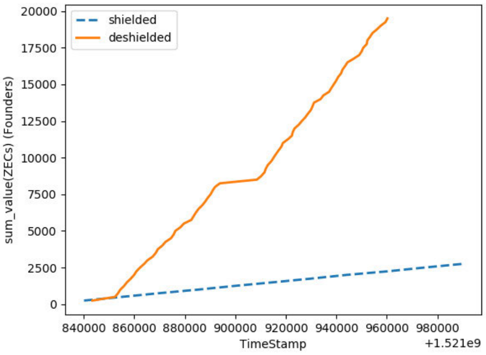


FIGURE 14. Shieldedpool value statistics operated by founders.

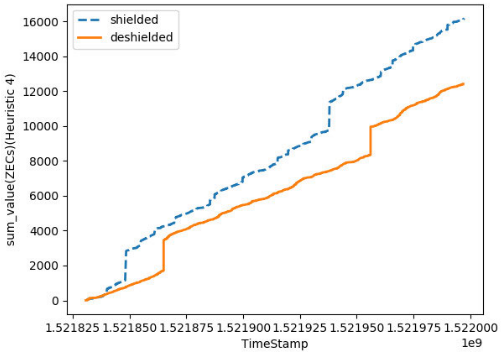


FIGURE 15. Shieldedpool value statistics according to Heuristic 4.

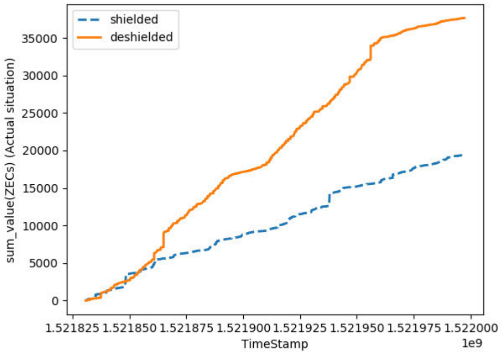


FIGURE 16. The actual shieldedpool value statistics.

Next we discuss the participation of various entities in shieldedpool. Figure 14 and Figure 15 respectively show the deposits and withdrawals of shieldedpool operated by founders and mining pools (miners). Figure 16 shows statistics of actual shieldedpool. We identify 95% of deposits and 87.5% of withdrawals. It means that even among very low proportions of transactions involving *z*-addresses, the majority users are founders, mining pools and miners, instead of changes, services or individual users. This further shows that few users actually use shieldedpool.

V. CONCLUSION

In this paper, we give a refined deanonymization analysis in Zcash. We first build a transaction network TN. Then we use deanonymization methods to simplify TN. These deanonymization methods include an improved address clustering heuristic (Heuristic 3) and mining heuristic (Heuristic 4). We compare and analyze several characteristics of TN and the simplified network TN_s. In particular, we find that users participating in shieldedpool are mostly founders, miners and mining pools after investigating regular behaviors of these participants. Future work may improve our heuristics (Heuristic 3 and 4) from a more rigorous perspective. Other methods of studying complex network, such as community division and dynamic analysis might also be useful for denanonymization.

REFERENCES

[1] S. Nakamoto. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System.* [Online]. Available: https://bitcoin.org/bitcoin.pdf

[2] A. Pfitzmann and M. Köhntopp, "Anonymity, unobservability, and pseudonymity—A proposal for terminology," in *Proc. Int. Workshop Design Issues Anonymity Unobservability Designing Privacy Enhancing Technol.*, Berkeley, CA, USA, Jul. 2000, pp. 1–9.

[3] D. Kondor, M. Pósfai, I. Csabai, and G. Vattay, "Do the rich get richer? An empirical analysis of the bitcoin transaction network," 2014, *arXiv:1308.3892*. [Online]. Available: https://arxiv.org/abs/1308.3892

[4] P. Koshy, D. Koshy, and P. D. McDaniel, "An analysis of anonymity in bitcoin using P2P network traffic," in *Proc. 18th Int. Conf. Financial Cryptogr. Data Secur. (FC)*, Christ Church, Barbados, Mar. 2014, pp. 469–485.

[5] 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," *Commun. ACM*, vol. 59, no. 4, pp. 86–93, 2016.

[6] F. Reid and M. Harrigan, "An analysis of anonymity in the bitcoin system," in *Proc. Privacy, Secur., Risk and Trust (PASSAT), IEEE 3rd Int. Conf. Int. Conf. Social Comput. (SocialCom)*, Boston, MA, USA, Oct. 2011, pp. 1318–1326.

[7] D. Ron and A. Shamir, "Quantitative analysis of the full bitcoin transaction graph," in *Proc. Financial Cryptogr. Data Secur. (FC)*, Okinawa, Japan, Apr. 2013, pp. 6–24.

[8] G. Maxwell. (2013). *CoinSwap: Transaction Graph Disjoint Trustless Trading*. [Online]. Available: https://bitcointalk.org/index.php?topic=321228.0

[9] T. Ruffing, P. Moreno-Sanchez, and A. Kate, "Coinshuffle: Practical decentralized coin mixing for bitcoin," in *Proc. 19th Eur. Symp. Res. Comput. Secur. Comput. Secur. (ESORICS)*, Wroclaw, Poland, Sep. 2014, pp. 345–364.

[10] E. Duffield and D. Diaz. (2014). *Dash: A Privacy Centric Cryptocurrency*. [Online]. Available: https://github.com/dashpay/docs/raw/master/binary/DashWhitepaper-V2.pdf

[11] N. Saberhagen. (2013). *Cryptonote V 2.0*. [Online]. Available: https://cryptonote.org/whitepaper.pdf

[12] E. B. Sasson, A. Chiesa, C. Garman, M. Green, I. Miers, E. Tromer, and M. Virza, "Zerocash: Decentralized anonymous payments from bitcoin," in *Proc. IEEE Symp. Secur. Privacy (SP)*, Berkeley, CA, USA, May 2014, pp. 459–474.

[13] B. Alex and F. Daniel. (2018). *Deanonymization of Hidden Transactions in ZCash*. [Online]. Available: https://www.cryptolux.org/images/d/d9/Zcash.pdf

[14] Q. Jeffrey. (2018). *An Analysis of Anonymity in the ZCash Cryptocurrency*. [Online]. Available: https://deepblue.lib.umich.edu/handle/2027.42/143130

[15] G. Kappos, H. Yousaf, M. Maller, and S. Meiklejohn, "An empirical analysis of anonymity in ZCash," in *Proc. 27th USENIX Secur. Symp. (USENIX Secur.)*, Baltimore, MD, USA, Aug. 2018, pp. 463–477.

[16] E. Androulaki, G. Karame, M. Roeschlin, T. Scherer, and S. Capkun, "Evaluating user privacy in bitcoin," in *Proc. 17th Int. Conf. Financial Cryptogr. Data Secur. (FC)*, Okinawa, Japan, Apr. 2013, pp. 34–51.

[17] H. Daira, B. Sean, H. Taylor, and W. Nathan. (2018). *Zcash Protocol Specification*. [Online]. Available: https://cryptorating.eu/whitepapers/Zcash/protocol.pdf

[18] J. Quesnelle, "On the linkability of Zcash transactions," 2017, *arXiv:1712.01210*. [Online]. Available: http://arxiv.org/abs/1712.01210

[19] H. Yousaf, G. Kappos, and S. Meiklejohn, "Tracing transactions across cryptocurrency ledgers," in *Proc. 28th USENIX Secur. Symp. (USENIX Secur.)*, Santa Clara, CA, USA, Aug. 2019, pp. 837–850.

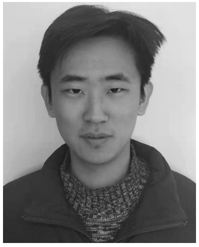
[20] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw.*, vol. 56, no. 18, pp. 3825–3833, Dec. 2012.

[21] P. W. Holland and S. Leinhardt, "Transitivity in structural models of small groups," *Comparative Group Stud.*, vol. 2, no. 2, pp. 107–124, 1971.

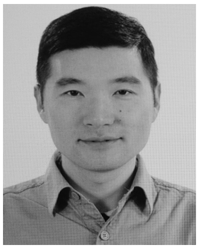
[22] D. J. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[23] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, Nov. 2009.

ZONGYANG ZHANG received the Ph.D. degree in computer software and theory from Shanghai Jiao Tong University, in 2012. He is currently an Assistant Professor with the School of Cyber Science and Technology, Beihang University. His research interests include public-key cryptography and blockchains.



WEIHAN LI received the bachelor's degree from Beihang University, in 2019, where he is currently pursuing the master's degree with the School of Cyber Science and Technology.



HAITAO LIU received the master's degree in physical electronics from Xidian University, in 2008. He is currently a Senior Engineer with the Department of Avionics, Chinese Flight Test Establishment. His research fields include aviation communication and navigation.

JIANWEI LIU received the Ph.D. degree from the Communication and Electronic System Department, Xidian University, in 1998. He is currently a Full Professor with the School of Cyber Science and Technology, Beihang University. His research interests include wireless communication networks, cryptography, information security, communication network security, channel coding, and modulation technology.

• • •