




Characterizing key agents in the cryptocurrency economy through blockchain transaction analysis

Xiao Fan Liu^{1*} , Huan-Huan Ren², Si-Hao Liu² and Xian-Jian Jiang²

*Correspondence:
xf.liu@cityu.edu.hk

¹Web Mining Laboratory,
Department of Media and
Communication, City University of
Hong Kong, 18 Tat Hong Avenue,
Kowloon, Hong Kong SAR, China
Full list of author information is
available at the end of the article

Abstract

The cryptocurrency economy provides a comprehensive digital trace of human economic behavior: almost all cryptocurrency users' activities are faithfully recorded in transactions on public blockchains. However, the user identifiers in the transaction records, i.e., blockchain addresses, are anonymous. That is, they cannot be associated with any real "off-chain" identify of actual users. Nonetheless, identifying the economic roles of the addresses from their past behaviors is still feasible. This paper analyzes Ethereum token transactions, characterizes key economic agents' behavior from their transaction patterns, and explores their identifiability through interpretable machine learning models. Specifically, six types of most active economic agents are considered, including centralized cryptocurrency exchanges, decentralized exchanges, cryptocurrency wallets, token issuers, airdrop services, and gaming services. Transaction patterns such as trading volume, transaction tempo, and structural properties of transaction networks are defined for individual blockchain addresses. The results showed that cryptocurrency exchanges and online wallets have signature behavior patterns and hence can be accurately distinguished from other agents. Token issuers, airdrop services, and gaming services can sometimes be confused. Moreover, transaction networks' features provide the richest information in the economic agent's identification.

Keywords: Cryptocurrency; Ethereum; Deanonimization; Network analysis; Machine learning

1 Introduction

The cryptocurrency economy is a complex yet transparent socioeconomic system. Bitcoin, Ethereum, and more than 270,000 other cryptocurrencies and tokens have been issued on dedicated or host blockchains as of June 2020 [1]. The most common ways for users to obtain cryptocurrencies are through coin mining and trading in cryptocurrency exchanges or over the counter. Individual investors and venture capital institutions can also purchase tokens from business teams in exchange for shares of their projects or companies. The cryptocurrencies obtained by users can be further used as money, merchandise, equity, and gaming tokens in various economic activities.

© The Author(s) 2021. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

The cryptocurrency economy not only replicates real-world economic systems but also further records every economic activity in public databases. A blockchain network is a decentralized computer system comprising a number of computers that independently verify, store, compute, and synchronize information generated by their end users [2]. Each record takes the form of a transaction, i.e., the transfer of cryptocurrency from one set of blockchain addresses to another, with an optional piece of auxiliary information. As of June 2020, Bitcoin blockchain network stored more than 500 million transactions among 600 million addresses [3], and Ethereum blockchain network stored more than 700 million transactions among 100 million addresses [1]. Because the blockchain networks are publicly accessible, the transaction records can be downloaded, audited, and analyzed by any interested party.

However, an obstacle to understanding the cryptocurrency transaction records is the anonymity of blockchain addresses. Compared to the user of a traditional online service who has to register and obtain an identity from a service provider, a cryptocurrency user can generate their identity, i.e., a pair of public and private keys, using ellipse encryption algorithms purely offline [2]. In this case, since the public key is not obtained from any party apart from the user, it is impossible to associate this “username” with any other information that can be used to infer the user’s real identity, e.g., their IP address.

Nonetheless, Satoshi Nakamoto warned in his proposal of the Bitcoin system that due to the transparency of the transaction records, repeatedly used blockchain addresses may reveal user behavior and hence user identity [4]. Although end users can create a new address for each transaction, the service providers, e.g., exchanges and wallets, typically cannot because they have to maintain stable service portals. As a result, the activeness of the millions of blockchain addresses is highly uneven. The cryptocurrency holdings of and the numbers of transactions initiated and received by the addresses all follow long tail distributions [5, 6]. The most active addresses, therefore, are naturally the entry points towards a comprehensive understanding of the cryptocurrency transaction records.

This research examines the most active blockchain addresses. Specifically, six types of most visible cryptocurrency economic agents are considered, including centralized and decentralized cryptocurrency exchanges, cryptocurrency wallets, token issuers, airdrop services, and gaming services. Transactional patterns such as volume features, temporal features, and structural features of the transaction network of blockchain addresses are used to characterize and differentiate agents in different roles.

The remainder of the paper is organized as follows: Sect. 2 provides a retrospect of the existing research on the identification and classification of blockchain addresses; Sect. 3 introduces the data sources, feature extraction methods, and machine learning models used to study the addresses; Sect. 4 presents the signature transactional features and the identifiability of the key agents; and Sect. 5 concludes the research and discusses future perspectives.

2 Related work

Early efforts of cryptocurrency address de-anonymization mainly based on heuristic address clustering on Unspent Transaction Output (UTXO) blockchain data models, e.g., Bitcoin. Two typical examples are multiple input and change address heuristics [7]. Multiple input heuristics consider that in a Bitcoin transaction with more than one input address, the input addresses are highly likely to belong to the same user. Change address

heuristics consider that when a transaction has multiple outputs, one of the outputs could be a change address, which belongs to the initiator of the transaction. With the addresses clustered, the addresses inside the same cluster can be considered to bear the same identity [7]. Heuristic methods are useful but prone to error. For example, 156,722 addresses were successfully associated with the largest cryptocurrency exchange, Mt. Gox, using Bitcoin transactions up to 2012 [5]. However, only approximately 69% of the addresses can be correctly associated with individual end users [8].

Another line of effort toward cryptocurrency address de-anonymization takes address identification as a classification problem. Machine learning algorithms are used to derive computational models from patterns extracted in transaction records. Transactional patterns used to describe a blockchain address include the amount, time, and frequency of its transactions; the cryptocurrency/token balance; and the active days [9–11]. Transaction networks can also be constructed among blockchain addresses, in which nodes are individual addresses or sets of addresses clustered by heuristics, and edges are transactions between addresses. The structural features of nodes in the networks include various centrality measures [10], motifs [12], and network representation learning derived embeddings in vector spaces [13, 14]. For smart contracts in Ethereum-like blockchains, their codes and bytecodes are also useful features [15, 16]. The above mentioned features are effective in binary classification tasks, e.g., determining whether an address is a Ponzi scheme [15, 17, 18], phishing address [14], or other kind of scam [19, 20], and multi-classification tasks, e.g., differentiating between cryptocurrency exchanges, gambling services, mining pools, and darknet markets [9–11, 21].

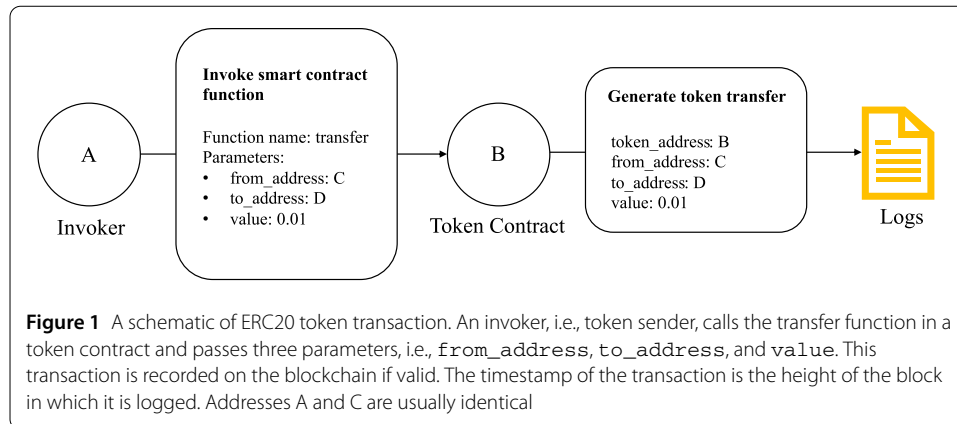
This paper follows the latter research direction, in which we systematically define a spectrum of features in transaction patterns and explore the identifiability of several key agents in the cryptocurrency economy. In contrast to the previous multi-classification tasks, we not only report the precision of classification but also elaborate the transactional patterns of the blockchain addresses with regards to their economic roles and explain in-depth the reasons for their identifiability or lack thereof.

3 Data and methods

3.1 Blockchain data

As of June 2020, Ethereum blockchain network stored the largest number of cryptocurrency transactions among all public blockchains. These transactions can be briefly classified into three types: ether (the original Ethereum currency) transfers, token transfers, and smart contract calls. The transactions of the more than 270,000 tokens account for 56% (414 million) of the total transactions (745 million). As many of the cryptocurrency economic activities, such as fundraising, deal only with tokens rather than ether, we use token transactions to study economic agent behaviors in this research.

A token transaction from one address to another is accomplished by invoking the `transfer()` function, in the token smart contract, with three parameters, namely `from_address`, `to_address`, and `value`, which stands for the sender, receiver, and amount of this transaction, respectively (Fig. 1). The token sender and receiver can be either a user-owned address (Externally Owned Account, EOA) or a contract address (CoA). ERC20 is the most common standard for creating customized tokens on Ethereum. ERC 20 tokens are fungible; that is, a token can be divided into small proportions, which can circulate in the economy independently. All ERC 20 tokens' transactions up to June 2019 were obtained using an Ethereum blockchain client.



3.2 Known key agent identities

Although technically anonymous, the identities of cryptocurrency addresses are sometimes publicly disclosed online. For example, some forum users, e.g., *Reddit* and *Bitcointalk* users, post personal Bitcoin or Ethereum addresses in their forum profiles. Addresses owned by cryptocurrency exchanges, wallet services, and gambling services can be identified by proactively trading or interacting with them [7]. Online intelligent platforms, such as *Walletexplorer.com* [22] and *Etherscan.io* [1], post known labels for Bitcoin and Ethereum addresses and allow users to tag addresses that they can recognize. We collected 3364 labels from Etherscan.io and retained addresses that belong to six key agent roles: centralized cryptocurrency exchange, decentralized exchange, wallet, token issuer, airdrop service, and gaming service, and that have participated in more than 100 transactions as of June 2019 as the study samples.

Centralized and decentralized exchanges are both cryptocurrency exchanges in which users can buy and sell different types of cryptocurrencies with fiat money or other cryptocurrencies. However, they bear a significant difference. In centralized exchange, a seller first deposits tokens into the exchange's addresses and open a sell order. The sell order is then matched with a buy order, either by the exchange or by the users themselves (over the counter). After clearing and settlement, the buyer can withdraw the token from the exchange. In this case, the exchange address serves as an escrow between the buyer and seller. Typical examples of centralized exchanges are *Binance* and *Kraken*. However, decentralized exchange users deal with the exchange directly. A decentralized exchange maintains a pool of different cryptocurrencies and sets the listing prices algorithmically. Buyers buy tokens from the pool, and sellers sell tokens to the pool. Typical examples of decentralized exchanges are *Bancor*, *KyberNetwork*, and *Uniswap*.

For the remaining types, *wallet* stands for online cryptocurrency banking services in which users deposit their cryptocurrencies and tokens, *token issuer* stands for the addresses that were used to sell tokens to investors through fundraising activities, e.g., Initial Coin Offering, Initial Exchange Offering, and Security Token Offering, *airdrop service* stands for the addresses that disseminate tokens freely to cryptocurrency users for advertisement purposes, and *gaming service* stands for the addresses used by gambling or recreational gaming organizers.

As shown in Table 1, the transactions of the selected addresses span three years and have exchanged billions of USD worth of tokens. Therefore, we believe that these addresses

Table 1 Six key agent roles and their basic transaction statistics

Type	Number of addresses (EOA/CoA)	Average number of transactions	Average total transaction volume (in USD)	Time span of transactions
Centralized exchange	124 (118/6)	166,110	17,173,430	May 2, 2016 to Jun 29, 2019
Decentralized exchange	201 (0/201)	17,877	665,844	Jun 10, 2017 to Jun 29, 2019
Wallet	316 (0/316)	1316	290,995	Aug 13, 2017 to Jun 29, 2019
Token issuer	210 (14/196)	14,269	801,416	Apr 18, 2017 to Jun 29, 2019
Airdrop service	149 (1/148)	35,158	66,996	Sep 11, 2017 to Jun 29, 2019
Gaming service	25 (3/22)	35,018	1	Nov 23, 2017 to Jun 29, 2019
Total	1025 (136/889)	32,895	2,471,780	May 2, 2016 to Jun 29, 2019

Table 2 Four groups of transaction pattern features for each blockchain address

Group	Symbol	Description	
Volume	M_{min}^{in}	Minimum dollars received in a transaction	
	M_{max}^{in}	Maximum dollars received in a transaction	
	M_{mean}^{in}	Average dollars received in a transaction	
	M_{std}^{in}	Standard deviation of dollars received in all transactions	
	M_{sum}^{in}	Total dollars received in all transactions	
	M_{min}^{out}	Minimum dollars sent in a transaction	
	M_{max}^{out}	Maximum dollars sent in a transaction	
	M_{mean}^{out}	Average dollars sent in a transaction	
	M_{std}^{out}	Standard deviation of dollars sent in all transactions	
	M_{sum}^{out}	Total dollars sent in all transactions	
	$M_{balance}$	$M_{sum}^{in} - M_{sum}^{out}$	
	Temporal	I_{min}^{in}	Minimum interval of received transactions
		I_{max}^{in}	Maximum interval of received transactions
I_{mean}^{in}		Average interval of received transactions	
I_{std}^{in}		Standard deviation of received transaction intervals	
I_{min}^{out}		Minimum interval of sent transactions	
I_{max}^{out}		Maximum interval of sent transactions	
I_{mean}^{out}		Average interval of sent transactions	
I_{std}^{out}		Standard deviation of sent transaction intervals	
Network size		T_{in}	Number of received transactions
		T_{out}	Number of sent transactions
	N_{in}	Number of transaction recipients, i.e., out-degree	
	N_{out}	Number of transaction senders, i.e., in-degree	
	N^{ego}	Number of nodes in the ego network	
Network structure	R	Reciprocity of the current address	
	C	Cluster coefficient of current address	
	D^{ego}	Density of the ego network	
	R^{ego}	Reciprocity of the ego network	

with disclosed identities can be considered representatives in the Ethereum ecosystem. Evidently, these six types are a non-exhaustive list of the key economic roles in the cryptocurrency economy. Some other major roles are also of interest. For example, the mining pools coin all the new original cryptocurrency in the blockchain system. However, they are not included in the current study because of their obvious marks, i.e., the addresses are stated explicitly in each mined block and thus do not need further characterization.

3.3 Transaction feature extraction

Four groups of transaction features (Table 2) are considered when characterizing blockchain addresses. Volumes and temporal features capture the patterns of transactions in which the addresses directly participate. Transaction network's structural features capture the higher order interaction patterns among the address and its counterparties.

Volume features include the mean, maximum, minimum, and total value of token transactions initiated and received by the node, respectively (giving eight variables), as well as the balance on an address. Token values are measured in US dollars using their daily exchange rates published on the online cryptocurrency market intelligence platform *Coinmarketcap.com*. If a token is not listed at the time of the transaction, its price is treated as 0. Temporal features include the mean, maximum, minimum, and standard deviation of the time intervals between the consecutive edges connecting to an address, i.e., the transactions initiated and received, giving another eight variables.

We use directed network $G = (V, E)$ to denote the transaction network constructed from token transfers. The set of nodes V represents blockchain addresses. $E = \{e | (V_s, V_t, t), V_s, V_t \in V\}$ is the set of directed edges, in which each e represents an ERC20 token transfer from addresses V_s to V_t in a block with height h . The block height can also be considered as the timestamp when the transaction occurs.

Network size features of a node include the numbers of incoming and outgoing edges, i.e., transactions, T_{in} and T_{out} , in-degree N_{in} , out-degree N_{out} , and the size N_{ego} of its 2-depth ego network. For each node v , its 2-depth ego network $G_v^{ego} = (V_v^{ego}, E_v^{ego})$ is defined as the collection of nodes V_v^{ego} , including the center node v and its direct and indirect neighbors that can be connected to within a distance of 2, by the edge set E_v^{ego} . Duplicated edges between nodes are combined, i.e., there is at most one edge connecting a pair of nodes in the ego networks.

Network structural features of a node v include the reciprocity between the node and its neighbors, i.e., the existence of bi-directional edges between two adjacent nodes,

$$R_v = \frac{|(u, v) | (u, v) \in E_v^{ego} \text{ and } (v, u) \in E_v^{ego}|}{|(u, v) | (u, v) \in E_v^{ego} \text{ or } (v, u) \in E_v^{ego}|};$$

the clustering coefficient, i.e., the existence of edges between the nodes' neighbors,

$$C_v = \frac{1}{\text{deg}^{tot}(v)(\text{deg}^{tot}(v) - 1) - 2 \text{deg}^{\leftrightarrow}(v)} T(v),$$

where $T(v)$ is the number of triangles that contains node v , $\text{deg}^{tot}(v)$ is the summation of its in-degree and out-degree, and $\text{deg}^{\leftrightarrow}(v)$ is its reciprocal degree; the density of its 2-depth ego network

$$D_v^{ego} = \frac{m}{n(n-1)},$$

where n is the number of nodes in the ego network and m is the number of edges in the ego network; and the reciprocity of its 2-depth ego network

$$R_v^{ego} = \frac{|(u, w) | (u, w) \in E_v^{ego} \text{ and } (w, u) \in E_v^{ego}|}{m}.$$

3.4 Training process of machine learning models

Five classifiers are trained: *Logistic Regression* (LR), *Support Vector Machine* (SVM), *Multilayer Perceptron* (MLP), *Random Forest* (RF), and *LightGBM*. These algorithms are the

most common choices and best performing ones in previous blockchain addresses identity classification tasks (mostly on Bitcoin addresses) [23]. The choices and tuning of the algorithm parameters are given in Sect. 1 of Additional file 1.

We train the model for five times and use the average accuracy, macro-precision, macro-recall, and macro-F1 as the final results. In each iteration, 100 addresses are randomly selected from each of the centralized exchange, decentralized exchange, wallet, token issuer, and airdrop services type, and all 25 gaming services addresses are used as the training sample. The 2-depth ego networks are constructed using the transactions in the active period of the center node v . The samples are further divided into a 80% training set and a 20% testing set. Models are trained, using stratified four-fold cross validation on the training set, i.e., 60% training and 20% validation, and further tested on the test set.

Considering that the number of gaming service nodes are far smaller than other types of key agents, cost-sensitive learning is used to solve the class imbalance problem. Specifically, we calculate the weight of each node as $w_i = k/p_i$, where $k = 1/\text{no. of types}$, and p_i is the proportion of the number of samples in the type to which sample i belongs in the entire sample set used to train the model, e.g., $15/315 = 1/21$ for gaming services.

4 Results

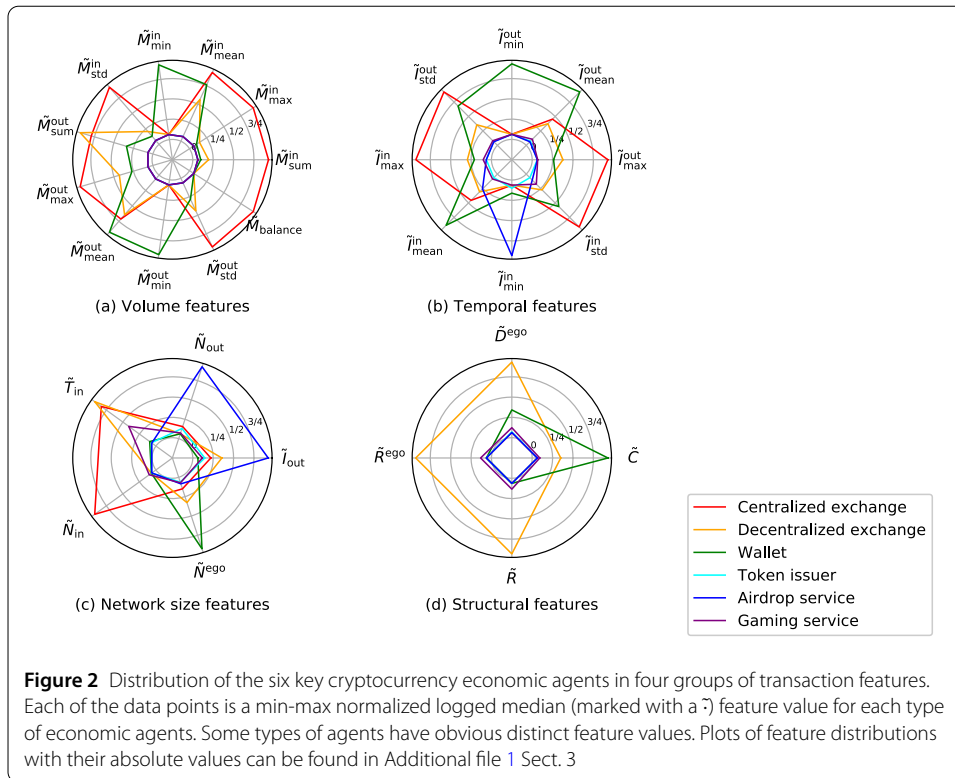
In this section, we first explore the signature features of different types of key economic agents and then use machine learning models to test the identifiability of the agents using their transactional behavior. Finally, we explore the importance of different features in identifying the agents.

4.1 Key agents' transaction patterns

The comparison of feature differentiability is shown in Fig. 2. Each of the data points is a min-max normalized logged median (marked with a $\tilde{\cdot}$) feature value for each type of economic agents. We have also provided the plots of feature value distributions using their original scales in Additional file 1 Sect. 3. Some obvious differentiation between different types of agents are discussed in the following paragraphs, though more subtle differences can be inspected by naked eyes from the figure.

Centralized exchanges (red lines) show distinctions in the total volume of tokens transferred into the centralized exchanges, e.g., $\tilde{M}_{\max}^{\text{in}}$, $\tilde{M}_{\text{std}}^{\text{in}}$, and $\tilde{M}_{\text{sum}}^{\text{in}}$, their balances $\tilde{M}_{\text{balance}}$, the maximum time interval between transactions $\tilde{T}_{\max}^{\text{in}}$ and $\tilde{T}_{\max}^{\text{out}}$, and the number of incoming edges \tilde{N}_{in} , i.e., the number of received transactions. These patterns indicate that centralized exchanges accept many incoming transactions from many users, and the received tokens tend to stay in the exchanges' addresses. However, the deposits to and withdrawals from the centralized exchanges distribute unevenly over time, implying that exchange users' activities might be driven by rare market events.

Decentralized exchanges (orange lines) have large incoming transactions \tilde{T}_{in} and the total volume of withdrawals $\tilde{M}_{\text{sum}}^{\text{out}}$. These features indicate that the users of decentralized exchanges tend to sell their tokens in many small transactions and buy in large bulks. Decentralized exchanges are particularly distinguishable in terms of their network structural features: the reciprocity \tilde{R} , density \tilde{D}^{ego} , and reciprocity \tilde{R}^{ego} of their 2-depth ego networks are significantly larger than other types of agents. These features indicate that decentralized exchanges are more popular among sophisticated users who are likely to store tokens in their own blockchain addresses and regularly transfer tokens among themselves.



Wallets (green lines) have also shown distinguishable features, such as large incoming and outgoing transaction time intervals, e.g., \tilde{t}_{mean}^{in} , \tilde{t}_{min}^{out} , and \tilde{t}_{mean}^{out} , and large minimum values \tilde{M}_{min}^{in} and \tilde{M}_{min}^{out} of tokens transferred in and out of the wallet addresses, respectively. These features indicate that cryptocurrency wallets are used as traditional banks: they do not have a high transaction frequency, but on average have larger transaction volumes. Like banks, wallet services usually serve as proxies for users and deal with other key economic agents directly and, therefore, have a large clustering coefficient \tilde{C} and larger 2-depth ego networks \tilde{N}^{ego} .

Token issuers (cyan lines) and airdrop services (blue lines) are both economy agents that disseminate tokens to investors. However, they exhibit different characteristics in their transaction behaviors. For token issuers, the level of activities, even in their most active period, are low. But for airdrop services, since they give out tokens for free to a larger user group rather than sell tokens to investors, the standard deviation of received transaction intervals \tilde{t}_{std}^{in} , initiated transactions \tilde{T}_{out} , and the out-degree \tilde{N}_{out} , i.e., the number of transaction recipients, are significantly larger than other key agents. Notably, the median values of volume features of both token issuers and airdrops addresses collected in our dataset are close to 0, which is largely because most of the tokens that had been disseminated did not reach cryptocurrency exchanges and hence were never priced.

Gaming addresses (purple lines) do not show any distinctive features from other key agents. Their network size features are similar to decentralized exchanges, while the network structural features and the volume features are close to airdrop services and token issuers.

Table 3 The prediction of five classifiers to the sample dataset

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
LR	0.771	0.756	0.740	0.738
SVM	0.750	0.636	0.657	0.636
MLP	0.840	0.804	0.780	0.785
LightGBM	0.890	0.879	0.853	0.858
Random Forest	0.893	0.888	0.862	0.865

Table 4 The identifiability of each type of key agent using the random forest classifier

Type	Precision	Recall	F1
Centralized exchange	0.938	0.910	0.924
Decentralized exchange	0.990	0.980	0.985
Wallet	1.000	0.950	0.974
Token issuer	0.750	0.780	0.765
Airdrop service	0.843	0.910	0.875
Gaming service	0.727	0.640	0.681

4.2 Model classification

The five models yield considerably high prediction power to the collected dataset; see Table 3. The random forest classifier achieved the highest scores in accuracy (89.3%), macro precision (88.8%), macro recall (86.2%) and macro F1 (86.5%).

For each type of key agent, Table 4 shows the precision, recall, and F1 from the random forest classifier. Centralized exchange, decentralized exchange, and wallet addresses can all be accurately distinguished from other types of key agents with precisions >90%. Airdrop services, preserving a certain extent of particular transactional features, can be identified with >80% probability. However, token issuers and gaming services can only be identified with 70% precision due to the lack of distinguishable transactional features.

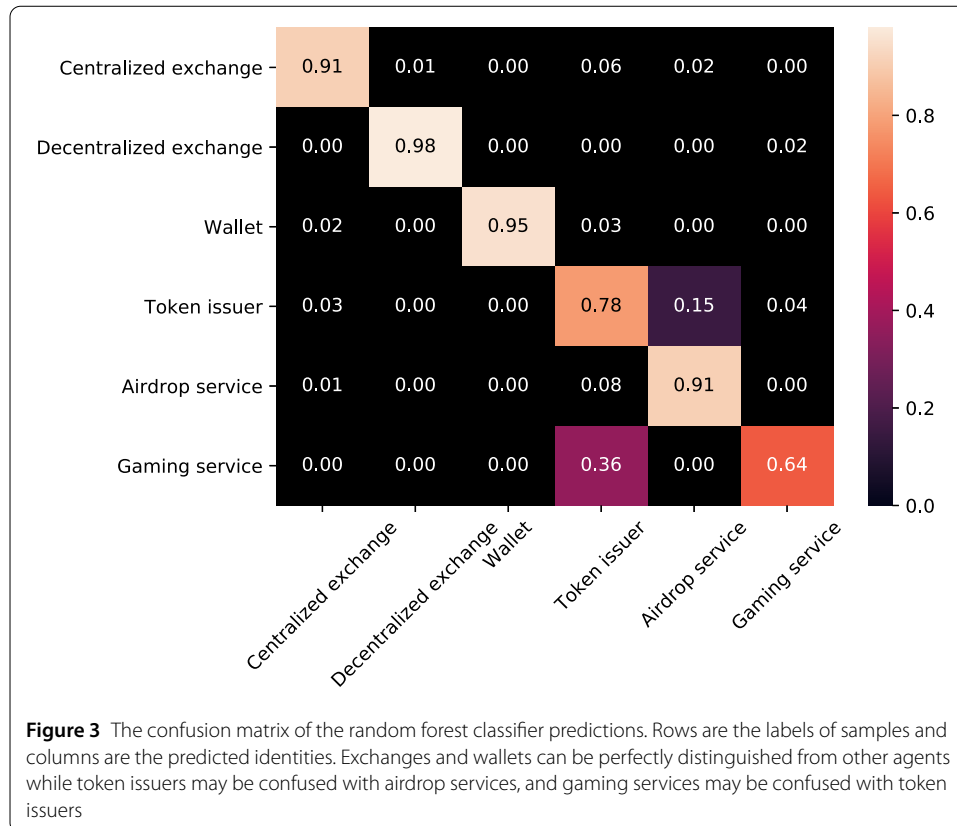
More specifically, Fig. 3 shows the confusion matrix of the random forest classifier predictions. Token issuer addresses are confused with airdrop services with 15% probability, while gaming services are misinterpreted into token issuers 36% of the time.

4.3 Analysis of informing features

Interpretable models such as RF provide quantitative descriptions of the importance of features (e.g., predictive power) in classification tasks. Figure 4 shows the feature importance in the random forest classifier based on the permutation importance. The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled [24]. The larger the decrease, the higher the predictive power a feature can hold.

Network size features, especially the out-degree N_{out} and the size of the ego network N^{ego} , are ranked highest. Temporal feature average interval of sent transactions I_{mean}^{out} is also ranked high. Moreover, transaction network structural features, such as the density D^{ego} and reciprocity R^{ego} of the 2-depth ego networks and the reciprocity R of the target addresses, are also ranked high. Volume features did not show high feature importance in the model.

Following the similar logic of permutation feature importance, we adapt a forward and backward feature selection-like process to investigate the importance of groups of features. Table 5 shows the prediction results of the random forest classifier using different combinations of feature groups, with the same hyper parameter settings. It can be seen



that using all groups of features achieved the highest macro F1 score. When using a single group of transaction features, network sizes have the highest predictive power. When using two groups, the combination of network size and temporal features achieves the highest identifiability. When using three groups, that is, emitting one group of features, the combination that leaves out temporal features provides the lowest prediction, which indicates that the temporal features provide the most uncorrelated information to other groups of features in identifying key agents in the cryptocurrency economy.

5 Conclusion and discussion

Key agents are the most significant parties in the cryptocurrency economy. These very few addresses deal with most of the transactions stored in the blockchains. A full understanding of these entry points could lay a solid ground for future exploration of the behavior of other economic agents, such as marketplaces, merchandisers, and various illicit activities.

Cryptocurrency transactions that are publicly stored in blockchains offer a unique data source to the study of cryptocurrency economy user behaviors. In this article, we have extracted transaction patterns, e.g., transaction volumes, transactions time interval, and transaction network structural features, e.g., the connectivity among blockchain addresses, to characterize and identify six types of key economic agents, namely, centralized exchanges, decentralized exchanges, cryptocurrency wallets, token issuers, airdrop services, and gaming services, in the cryptocurrency economy.

Centralized exchanges, decentralized exchanges, and online cryptocurrency wallets all show distinctive features. Centralized exchanges act as escrow between the buyers and sellers, and hence receive large amounts of deposit and hold large balances. Decentralized

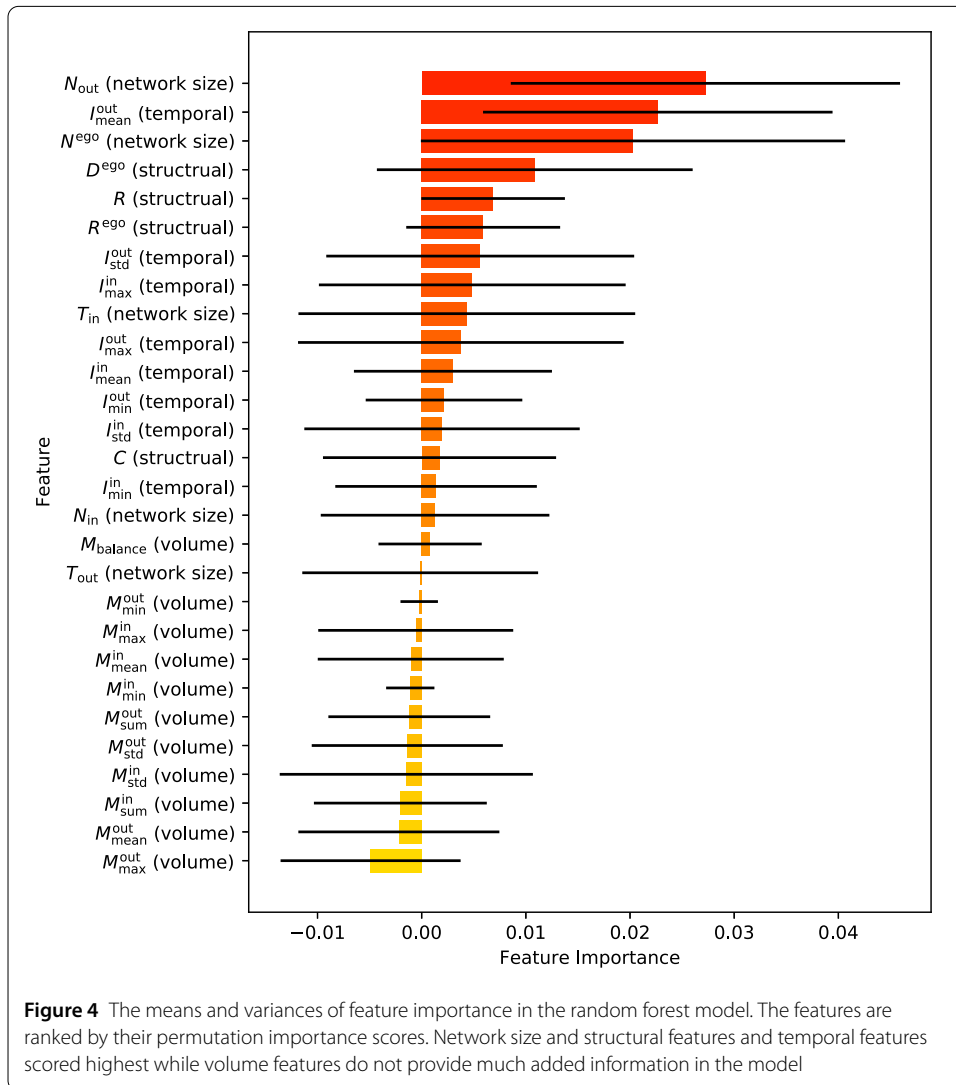


Table 5 Prediction results of the random forest classifier using different combinations of features

Groups of features	Accuracy	Macro Precision	Macro Recall	Macro F1
Volume	0.488	0.675	0.552	0.469
Temporal	0.804	0.767	0.728	0.730
Network size	0.813	0.752	0.737	0.737
Network structure	0.665	0.610	0.612	0.605
Temporal + Volume	0.844	0.827	0.813	0.814
Size + Volume	0.851	0.820	0.820	0.811
Size + Temporal	0.857	0.854	0.820	0.827
Size + Structure	0.848	0.816	0.787	0.789
Structure + Volume	0.790	0.759	0.752	0.749
Structure + Temporal	0.834	0.839	0.765	0.775
Size + Temporal + Volume	0.867	0.853	0.848	0.845
Structure + Temporal + Volume	0.872	0.867	0.833	0.841
Size + Structure + Volume	0.861	0.828	0.813	0.814
Size + Structure + Temporal	0.882	0.889	0.847	0.850
All groups of features	0.893	0.888	0.862	0.865

exchanges trade with users automatically and, therefore, show significantly higher reciprocity in their transaction network structure. Online wallet services can be considered cryptocurrency banks and, therefore, have a higher minimum value of withdrawal transactions. Token issuers and airdrop services both disseminate tokens to investors. However, since airdrop services give out tokens to a larger user group, they have a much larger number of outgoing transactions than token issuers. Gaming services typically receive many incoming transactions but did not show distinctive features compared to the other types of key agents.

Machine learning algorithms trained on the extracted features showed strong predictive power for the six types of key agents, e.g., macro $F1 = 0.865$. The prediction results are robust to different sampling criteria and model hyper parameter settings. However, even though the exchanges and wallet services can be differentiated accurately from other types of key agents, token issuers, airdrop services, and gaming services can sometimes be confused with each other. Feature importance analysis has indicated that network size and structural features possess the highest predictive power for the key agents, while transaction temporal features provide the most independent information from all other groups of features.

However, the categorization of key cryptocurrency economic agents into six types can be further discussed. For example, decentralized exchanges and online wallets can be easily divided further among themselves, probably corresponding to different business models that the services adopt (see Sect. 4 in Additional file 1 for an exploratory plot).

The significance of blockchain technology is that all user activities are faithfully stored and accessible to the public, enabling any illicit activities, such as market manipulation in cryptocurrency exchanges, the hacking of online wallets, and cheating in games, to be immediately exposed to the public. Though many newly developed cryptocurrencies, e.g., Zcash and Monero, see this nature as a weak link in the original Bitcoin design and try to conceal the traceability of transactions by cryptography designs, we argue that the identifiability of cryptocurrency economy agent roles does not jeopardize the privacy and security feature of cryptocurrency but rather reinforces the trustworthiness of the entire cryptocurrency economy. Understanding the economic roles associated with each blockchain address promotes confidence in their transaction counterparts and is thus the first step toward creating a fully transparent and self-regulated decentralized economy.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1140/epjds/s13688-021-00276-9>.

[Additional file 1](#). Supplementary information (DOCX 2.0 MB)

Funding

This work is supported by City University of Hong Kong (grant no. 7200649 and 6354050).

Availability of data and materials

Blockchain data are publicly available in Ethereum database. Label data are publicly available on Etherscan.io and our Github repository https://github.com/abcdefg3381/cryptocurrency_analysis. Price data are publicly available on coinmarketcap.com

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

XFL designed the study, HHR, SHL, and XJJ conducted data analysis and all authors wrote the paper. All authors read and approved the final manuscript.

Authors' information

Dr. Xiao Fan Liu is an assistant professor in City University of Hong Kong, Hong Kong SAR, China. Mr. Huan-Huan Ren, Ms. Si-Hao Liu, and Mr. Xin-Jian Jiang are masters degree students in Southeast University, Nanjing, China.

Author details

¹Web Mining Laboratory, Department of Media and Communication, City University of Hong Kong, 18 Tat Hong Avenue, Kowloon, Hong Kong SAR, China. ²School of Computer Science and Engineering, Southeast University, 2 Dongnandaxue Road, 211189 Nanjing, China.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 5 July 2020 Accepted: 21 April 2021 Published online: 01 May 2021

References

1. Ethereum Blockchain Explorer. <https://etherscan.io>
2. Dinh TTA, Liu R, Zhang M, Chen G, Ooi BC, Wang J (2018) Untangling blockchain: a data processing view of blockchain systems. *IEEE Trans Knowl Data Eng* 30(7):1366–1385. <https://doi.org/10.1109/Tkde.2017.2781227>
3. Blockchain Explorer – Search the Blockchain. <https://www.blockchain.com>
4. Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. <https://bitcoin.org/bitcoin.pdf>
5. Ron D, Shamir A (2013) Quantitative analysis of the full bitcoin transaction graph. In: Proceedings of the 17th international conference on financial cryptography and data security. FC '13, pp 6–24. https://doi.org/10.1007/978-3-642-39884-1_2
6. Guo DC, Dong JQ, Wang K (2019) Graph structure and statistical properties of Ethereum transaction relationships. *Inf Sci* 492:58–71
7. Meiklejohn S, Pomarole M, Jordan G, Levchenko K, McCoy D, Voelker GM, Savage S (2016) A fistful of bitcoins: characterizing payments among men with no names. *Commun ACM* 59(4):86–93
8. Nick JD (2015) Data-driven de-anonymization in bitcoin. Master thesis, ETH-Zürich. <https://doi.org/10.3929/ethz-a-010541254>
9. Toyoda K, Ohtsuki T, Mathiopoulos PT (2018) Multi-class bitcoin-enabled service identification based on transaction history summarization. In: IEEE int. Congr. Cybermatics: IEEE conf. Internet things, green comput. Commun., cyber, phys. Soc. comput., smart data, blockchain, comput. Inf. technol. IEEE, Halifax, pp 1153–1160
10. Jourdan M, Blandin S, Wynter L, Deshpande P (2018) Characterizing entities in the bitcoin blockchain. In: 18th IEEE int. conf. Data min. workshops. (ICDMW 2018), Sentosa, Singapore, pp 55–62
11. Lin Y-J, Wu P-W, Hsu C-H, Tu I-P, Liao S-W (2019) An evaluation of bitcoin address classification based on transaction history summarization. In: 1st IEEE int. conf. Blockchain cryptocurrency. (ICBC 2019), Seoul, Republic of Korea, pp 302–310
12. Akcora CG, Dey AK, Gel YR, Kantarcioglu M (2018) Forecasting bitcoin price with graph chainlets. In: Phung D, Tseng VS, Webb GI, Ho B, Ganji M, Rashidi L (eds) 22nd Pacific-Asia conf. Adv. knowl. Discov. Data min. (PAKDD 2018), Melbourne, VIC, Australia, vol 10939, pp 765–776
13. Liang J, Li L, Chen W, Zeng D (2019) Targeted addresses identification for bitcoin with network representation learning. In: 17th IEEE int. conf. Intell. Secur. Inf. (ISI 2019), Shenzhen, China, pp 158–160
14. Wu J, Yuan Q, Lin D, You W, Chen W, Chen C, Zheng Z (2019) Who Are the Phishers? Phishing Scam Detection on Ethereum via Network Embedding. [arXiv:1911.09259](https://arxiv.org/abs/1911.09259)
15. Chen W, Zheng Z, Ngai ECH, Zheng P, Zhou Y (2019) Exploiting blockchain data to detect smart ponzi schemes on Ethereum. *IEEE Access* 7:37575–37586. <https://doi.org/10.1109/ACCESS.2019.2905769>
16. Torres CF, Steichen M, State R (2019) The art of the scam: demystifying honeypots in Ethereum smart contracts. In: 28th USENIX secur. Symp., Santa Clara, USA, pp 1591–1607
17. Bartoletti M, Pes B, Serusi S (2018) Data mining for detecting bitcoin ponzi schemes. In: 2018 crypto valley conf. Blockchain technol. (CVCBT 2018), Zug, Switzerland, pp 75–84
18. Toyoda K, Mathiopoulos PT, Ohtsuki T (2019) A novel methodology for hyip operators' bitcoin addresses identification. *IEEE Access* 7:74835–74848
19. Ostapowicz M, Żbikowski K (2019) Detecting fraudulent accounts on blockchain: a supervised approach. In: Cheng R, Mamoulis N, Sun Y, Huang X (eds) 20th int. conf. Web inf. syst. Eng. (WISE 2019), vol 11881. Springer, Hong Kong, pp 18–31
20. Weber M, Domeniconi G, Chen J, Weidele DK, Bellei C, Robinson T, Leiserson CE (2019) Anti-money laundering in bitcoin: experimenting with graph convolutional networks for financial forensics. [arXiv:1908.02591](https://arxiv.org/abs/1908.02591)
21. Sun Yin HH, Langenheldt K, Harlev M, Mukkamala RR, Vatraru R (2019) Regulating cryptocurrencies: a supervised machine learning approach to de-anonymizing the bitcoin blockchain. *J Manag Inf Syst* 36(1):37–73
22. WalletExplorer.com Smart Bitcoin Block Explorer. <https://www.walletexplorer.com/>
23. Liu XF, Jiang X-J, Liu S-H, Tse CK (2020) Knowledge discovery in cryptocurrency transactions: a survey. [arXiv:2010.01031](https://arxiv.org/abs/2010.01031)
24. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32