Information Sciences 492 (2019) 58–71

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Graph structure and statistical properties of Ethereum transaction relationships

Dongchao Guo[a], Jiaqing Dong[b], Kai Wang[c,*]

[a] *School of Computer Science, Beijing Information Science and Technology University, Beijing 100101, China*
[b] *Research Institute of Information Technology, Tsinghua University, Beijing 100084, China*
[c] *School of Computer Science and Technology, Harbin Institute of Technology, Weihai, Shandong 264209, China*

ARTICLE INFO

Article history:
Received 23 August 2018
Revised 5 April 2019
Accepted 6 April 2019
Available online 8 April 2019

MSC:
05C82
62P35
91G80
62F03

Keywords:
Blockchain
Ethereum
Transaction relationships
Heavy tail
Power law

ABSTRACT

In recent years, the rapid development of blockchain technologies has attracted considerable attention. However, little effort has been devoted toward investigating the large amount of trade data recorded in blockchains. This paper focuses on transaction data in Ethereum, which is a prominent public blockchain platform supporting not only secure cryptocurrency transfer but also various decentralized applications. By means of the framework of network science theory, we find that several transaction features, such as transaction volume, transaction relation, and component structure, exhibit a heavy-tailed property and can be approximated by the power law function. In particular, we find that the transaction relations follow a bow-tie structure with negative assortativity if they are regarded as a directed graph. The popular hubs tend to connect to a large number of common users. We believe that the aforementioned statistics can be ascribed to the vast diversity of transactions and the existence of a number of cryptocurrency exchanges. To the best of our knowledge, this study is the first to not only carry out a relatively comprehensive investigation of the transaction data recorded in Ethereum but also probe the statistical laws underlying the transaction relationships from the perspective of network science.

1. Introduction

In recent years, with the emergence of an increasing number of well-established blockchain platforms along with blockchain-based high-quality decentralized applications, blockchain technology has attracted considerable attention and research interest [1,12,20,21,29,30,35,37]. Moreover, the significance of blockchain technology may be comparable to that of the Internet in daily life, owing to its potential to facilitate secure transfer of valuable digital assets and trusted cooperation in untrusted cyberspace without resorting to any third-party authority [11,37].

The term "blockchain" may refer to, but is not limited to, the following: (a) a specific data structure composed of a chain of ordered data blocks secured by cryptography technologies, (b) a blockchain platform (e.g., Bitcoin [28] and Ethereum [43]) supporting trusted cryptocurrency transfer and smart contracts, or (c) a bundle of technologies constituting a blockchain platform, such as a ledger and distributed consensus [29]. Blockchain was originally designed to securely transfer cryptocurrency, i.e., a type of digital currency, between anonymous users in a trusted decentralized manner without resorting to any trusted third-party authority. Bitcoin is the first practical platform constructed on the basis of blockchain technology. In this

[*] Corresponding author.
E-mail address: dr.wangkai@ieee.org (K. Wang).

https://doi.org/10.1016/j.ins.2019.04.013

sense, Bitcoin is widely believed to have initiated the era of blockchain 1.0, exemplifying the first-generation blockchain technology at its best. With the rapid development of Bitcoin, the underlying blockchain technology has gained rapid acceptance globally. The onset of the second-generation blockchain technology was marked by the introduction of Ethereum. While Bitcoin can support only built-in cryptocurrency and value transfer, Ethereum, as the first blockchain platform with a built-in Turing-complete language, extends the functionalities of the blockchain platform to the maximum extent and makes it possible to support nearly all types of decentralized applications.

As a promising technology with the potential for substantially changing the current Internet architecture, blockchain can facilitate not only secure value transfer but also trusted cooperation between untrusted entities in a cyber-physical-social (CPS) system [11,37]. In recent years, considerable effort has been devoted toward understanding blockchain technology, improving it, and applying it to various scenarios. Previous studies have mainly focused on blockchain-based cryptocurrency [29], performance and security analysis [1,20], improvement of the blockchain platform [21], and various applications [26,35]. As blockchain-based cryptocurrency is an attractive digital asset, many researchers have been focusing on it considering its rapid growth in terms of market capitalization. Narayanan et al. reviewed the history of blockchain and cryptocurrencies and provided a comprehensive introduction to Bitcoin [29]. Valfells and Egilsson studied how one could profit from mining Bitcoin given limited electric power [41]. Blockchain is an enabling technology for future information infrastructure. By means of blockchain technology, Pentland et al. launched the MIT Connection Science Research Initiative, which aims to create an Internet of trusted data that provides secure access for everyone in an integrated CPS system [37]. In addition, some effort has been devoted toward improving the current information infrastructure. For example, an information system based on a decentralized DNS was established by Blockstack [6]. Castro et al. proposed a system based on a decentralized public ledger to provide autonomous systems (ASs) with automatic methods to establish and verify connectivity agreements [9]. Considering blockchain as an enabling technology of the digital society, Davidson et al. noted that blockchain technologies make trusted cooperation possible, which indeed boosts innovations [11]. Considering the potential of blockchain technology as discussed above, it is of great importance to study the performance and security of the blockchain platform itself [1,7,12,20]. In addition, blockchain technology has applications in data protection and data sovereignty. Cao et al. [8] proposed a cloud-assisted e-Health system to prevent electronic health records (EHRs) from being illegally modified by using blockchain technology, and they showed that the proposed system can provide a strong security guarantee with high efficiency. Zhang et al. [44] proposed a blockchain-based fair online payment framework for outsourcing services in cloud computing, and they showed the soundness, robust fairness, and efficiency of the framework. Li et al. [25] proposed a blockchain-based architecture for secure distributed cloud storage, which outperforms traditional cloud storage architectures in terms of the file loss rate and network transmission delay.

The above mentioned studies regarded the blockchain platform as a distributed system and investigated the system performance and security. Note that the basic function of any blockchain platform is to securely conduct transactions that are stored in the distributed ledger (i.e., the blockchain as a data structure). From the data science perspective, analysis of the massive volume of big data in a blockchain can provide significant insights that could inspire researchers in many areas. By combining the information from an advertising website called Backpage and the transaction data of Bitcoin, Portno et al. proposed a method to identify human traffickers [38]. Tasca et al. tried to categorize Bitcoin addresses into some clusters corresponding to real business entities and investigated the transaction behavior pattern in each business category on the basis of observed transaction data for a period of seven years [40] by using the cluster detection technique proposed by Doll et al. [13]. Garcia et al. ascribed the bubble of Bitcoin market capitalization to the complex feedback and interactions of a socio-economic system [19]. Ron and Shamir investigated user behaviors of the Bitcoin transaction graph by using some statistical methods [39].

Meanwhile, the transaction data of other blockchain platforms, of which Ethereum is perhaps the most popular example, has not attracted much attention. As mentioned above, the Ethereum platform is considered an example of the second-generation blockchain technology. Thus, it is of great importance to study the Ethereum platform from multiple perspectives. In this paper, we focus on the transaction data recorded in the Ethereum blockchain. In particular, we investigate the transaction relations between Ethereum users from the network science perspective. Network science, sometimes referred to as the complex network theory, is regarded as a general modeling and analysis framework for studying the structural and dynamical characteristics of many real complex systems [3,5,24,32,36].

The main contributions of this paper can be summarized as follows.

- The data of transactions recorded in the Ethereum blockchain are investigated from the network science perspective. Some ubiquitous laws that govern the Ethereum transaction relationships are uncovered.
- Several critical transaction features, such as the transaction volume, outgoing or incoming transaction relation number, and component size of the transaction network, exhibit a heavy-tailed property and can be approximated by the power law distribution.
- The structure of the giant weakly connected component of the directed transaction network can be described by a sophisticated bow-tie schematic model. The topology shows negative assortativity, which means that the hubs tend to connect with nodes with low degrees and that the topology does not exhibit the so-called “rich club” phenomenon with respect to the connectivity pattern.
- The aforementioned statistical results are ubiquitous and irrelevant to the data sets.

- The results of experiments and analyses conducted in this study provide implications for researchers whose interests lie in cryptocurrency, cybersecurity, and other related areas.

The remainder of this paper is organized as follows. Section 2 provides a brief primer on the complex network theory and network measurements used in this paper. Section 3 describes how we obtain up-to-date raw data of blocks from the Ethereum blockchain and extract transactions from the raw data. In addition, the transaction relations are modeled as a network, followed by statistical analyses of the Ethereum transaction network. Specifically, we obtain several network measurements to gain insights into the statistical properties of the transaction network in the remainder of Section 3. Finally, Section 4 concludes the paper.

2. Network measurements

In this section, we introduce several critical concepts and measurements of network science [32,42], which will be used in this paper. A complex system composed of interconnected components can be modeled by means of the framework of network science. With these measurements, we can statistically measure the structural characteristics of any network in a quantitative manner. The terms "network" and "graph" are interchangeable in this paper. We introduce the definitions of graphs and graph structures using linear algebra and graph theory. We also introduce the concepts and definitions of network measurements such as degree, degree distributions, and assortativity. In addition, we specify how these measurements are obtained. Finally, we present two common graph models and discuss how the models can properly fit real observational data sets.

A network or a graph G can be described by $G = (\mathcal{N}, \mathcal{L})$, where the sets $\mathcal{N}$ and $\mathcal{L}$ denote the set of nodes (vertices) and the set of links (edges), respectively. The graph G is composed of N nodes interconnected by L links. An example of a graph with three nodes and two links can be specified as follows: $\mathcal{N} = \{1, 2, 3\}$ and $\mathcal{L} = \{(1, 2), (1, 3)\}$. A node can be referred to by its rank i in the set $\mathcal{N}$ of nodes. Two nodes i and j are said to be adjacent or called neighbors if they are directly connected by a link. Further, let $N = |\mathcal{N}|$ and $L = |\mathcal{L}|$ denote the cardinalities of the sets of nodes and links, respectively.

With linear algebra, it is more convenient to represent a network by an $N \times N$ matrix A with elements $a_{i,j}$, which is the so-called adjacency matrix. An undirected network can be represented by a symmetric adjacency matrix where all the elements are either 1 or 0. Conversely, a directed network can be represented by an asymmetric adjacency matrix. For an undirected network, the degree D_i of node i is defined as the number of its directly connected neighbors. Specifically, $D_i = \sum_j a_{i,j}$. The degree distribution $P[D = k]$ (or $P(k)$ in short) is defined as the probability distribution of the nodal degree. In other words, the distribution $P[D = k]$ denotes the probability of a randomly selected node with k neighbors. For a graph with finite size, the degree distribution $P(k)$ can be calculated by

$$P(k) = \frac{\sum_{j=1}^{N} 1_{D_j = k}}{N} = \frac{\text{number of nodes with degree } k}{\text{number of all nodes}}, \tag{1}$$

where the indicator function $1_X = 1$ if the event X is true and 0 otherwise. For an undirected network with finite size, one can calculate the average degree $E[D]$ by

$$E[D] = \langle k \rangle = \sum_k kP(k) = \frac{2L}{N}. \tag{2}$$

One may want to determine how one type of node with degree k is connected to another type of node with degree k', which is called the degree correlation. Introducing the joint degree distribution (JDD) $P(k, k')$ as a measurement is useful to some extent. However, the measurement JDD is merely an indirect way to provide insights into the degree correlation. A direct way to measure the degree correlation is by calculating the assortativity coefficient r of a graph [33], which summarizes the JDD as a single scalar:

$$r = \frac{\frac{1}{L}\sum_{e_{i,j}\in\mathcal{L}} D_i D_j - \left(\sum_{e_{i,j}\in\mathcal{L}} \frac{1}{2L}(D_i + D_j)\right)^2}{\frac{1}{L}\sum_{e_{i,j}\in\mathcal{L}} \frac{1}{2}(D_i^2 + D_j^2) - \left(\sum_{e_{i,j}\in\mathcal{L}} \frac{1}{2L}(D_i + D_j)\right)^2}. \tag{3}$$

The assortativity coefficient r can be considered as a special type of the Randić index, which is the most studied and most frequently applied graph measure defined for quantifying the topological properties of graphs. Further details on the Randić index and its applications can be found elsewhere [27].

For a directed network, we define the in-degree and out-degree of node i as $D_{\text{out},i}$ and $D_{\text{in},i}$, respectively. The in-degree and out-degree of the node represent the number of incoming links from its directed neighbors and the number of outgoing links to its directed neighbors, respectively. Similarly to the degree distribution $P[D = k]$, one can define the out-degree distribution and the in-degree distribution as $P_{\text{out}}(k)$ and $P_{\text{in}}(k)$, respectively. For a specific network, one can compute the in-degree and out-degree of node i as follows:

$$D_{\text{out},i} = \sum_{j=1}^{N} a_{i,j}, \tag{4}$$

$$D_{\mathrm{in},i} = \sum_{j=1}^{N} a_{j,i}. \tag{5}$$

The corresponding degree distributions are defined as follows:

$$P_{\mathrm{out}}(k) = \frac{\sum_{j=1}^{N} 1_{D_{\mathrm{out},j}=k}}{N}, \tag{6}$$

$$P_{\mathrm{in}}(k) = \frac{\sum_{j=1}^{N} 1_{D_{\mathrm{in},j}=k}}{N}. \tag{7}$$

One can define a map $W : \mathcal{L} \to \mathbb{R}$ such that each link is assigned a so-called weight. The weight as a measurement typically denotes the strength of the connection between nodes. A typical directed weighted graph can be represented in matrix form as follows:

$$A = \begin{bmatrix} 0 & 0.5 & 0 \\ 1.25 & 0 & 0.01 \\ 0 & 0.75 & 0 \end{bmatrix}.$$

A path $P_{i\to j}$ between two nodes is defined as the ordered sequence of links between them, where two consecutive links are connected to the same node and no repeated links exist. Node j is said to be reachable from node i if there exists any path $P_{i\to j}$. The shortest path is defined as the path with the minimum sum of weights of the ordered link sequence between the nodal pair concerned. The hopcount $H_{i\to j}$ is defined as the cardinality of the link set of the shortest path. In other words, the hopcount is the number of links between the nodal pair concerned. The average hopcount $\bar{H}$ is defined as the mean value of the hopcount $H_{i\to j}$. The diameter of the graph is defined as the hopcount of the longest shortest path. As graph is said to be connected only if there is at least one path for each nodal pair. A directed graph is said to be strongly connected if and only if each node is reachable from the other nodes. Further, it is said to be weakly connected only if its undirected counterpart is connected.

Erdős and Rényi [15] proposed a description of the topological structure of complex networks with the random graph theory. Their ER random graph with Poisson degree distribution is the most attractive example. The connection between any pair of nodes in the ER random graph is established with a constant probability factor p, and the corresponding degree distribution is

$$P(k) = \lim_{k\to\infty} \binom{N-1}{k} p^k (1-p)^{N-1-k} \simeq e^{-c}\frac{c^k}{k!}, \tag{8}$$

where the notation $c = \langle k \rangle = pN$ is the average degree.

Intriguingly, it was later shown that the assumption that the nodes of all real-world networks are purely randomly wired is not true. Many studies have shown that the degree distributions of a large number of real-world networks tend to hold a heavy tail and can be approximated by the power law function as follows:

$$P(k) \sim k^{-\alpha}. \tag{9}$$

Owing to the lack of a characteristic degree in the power law case when $N \to \infty$, this type of network is called a scale-free network. Many technology networks, such as the AS-level Internet [18] and the World Wide Web (WWW) [4] have been reported to follow the scale-free connectivity pattern. As shown by previous studies, the emergence of the power law degree distribution can be ascribed to two simple principles of generating a network, i.e., the growth of a network and the preferential attachment of a newborn link. The scale-free property is widely believed to reflect the vast diversity in complex networks and can thus be regarded as a critical and ubiquitous law in nature [2].

We consider the subgraph structure of unweighted graphs. A subgraph of G is a graph formed from a subset of nodes and links of G. In the random graph theory, an undirected graph is composed of a giant component and a number of finite components, where each component is a subgraph of the entire network. The giant component is defined as the connected subgraph with size scaling with $O(N)$, while the finite components can be considered as small fragments not comparable to the entire network with respect to the graph size. The distribution $P(s)$ of component size as a macro-level measurement is introduced to measure the topology of the entire network.

The graph structure of a directed graph is more complicated. It is necessary to introduce the concepts of weakly and strongly connected components. A weakly connected component (WCC) is a connected subgraph such that its corresponding undirected graph is connected. The giant weakly connected component (GWCC) is the maximum WCC (i.e., the one with max cardinality). Hence, the GWCC of a directed graph is equivalent to the giant component of the corresponding undirected graph if its size is of the order $O(N)$. A strongly connected component (SCC) is defined as a subgraph such that there exist both a path $P_{i\to j}$ and a path $P_{j\to i}$ for each pair of nodes in the SCC.

In this study, we aim to determine which distribution is the best for approximating the sampling data. Statistical fluctuations are unavoidable. In statistics, a simple way to deal with the adverse effect of sampling fluctuations on the estimated distribution is to use the complementary cumulative distribution function (CCDF) defined as follows:

$$P_c(k) = \int_{k_0=k}^{\infty} P(k_0)dk_0. \tag{10}$$

Considering the power law distribution as an example, the relation between the distribution and its corresponding cumulative distribution is

$$P_c(k) = \int_{k_0=k}^{\infty} c_0 k_0^{-\alpha} dk_0 = c_1 k^{-\alpha+1}, \tag{11}$$

where c_0 and c_1 are normalized constants. Thus, the cumulative distribution of a power law distribution still follows a power law form. The difference between them is no more than a constant exponential factor 1. The parameter estimation approach described above, which is based on the cumulative distribution, is straightforward and easy to employ. Nevertheless, the accuracy of this method is not satisfactory. By means of the maximum likelihood (ML) method and the Kolmogorov–Smirnov (KS) statistic, Clauset et al. [10] proposed a novel technique to obtain accurate power law parameter estimates for the observed data. In other words, the observations are consistent with the power law distribution model in which the best-fit value for the power law exponent can be extracted using their technique. The maximum likelihood estimator (MLE) for the continuous case and the discrete case are

$$\hat{\alpha} = 1 + m\left[\sum_{i=1}^{m} \ln \frac{x_i}{x_{\min}}\right]^{-1}, \tag{12}$$

$$\hat{\alpha} \simeq 1 + m\left[\sum_{i=1}^{m} \ln \frac{x_i}{x_{\min} - 0.5}\right]^{-1}, \tag{13}$$

where $\hat{\alpha}$ is the estimated power law estimator, m denotes the number of observed data, x_i denotes the observed value such that $x_i \geq x_{\min}$, and $x_{\min}$ is the lower limit on the power law behavior. The best choice for $x_{\min}$ is such that

$$\min_{x_{\min}} \max_{x \geq x_{\min}} |P_{c1}(x) - P_{c2}(x)|, \tag{14}$$

where P_{c1} and P_{c2} denote the cumulative distribution functions (CDF) for the estimated distribution and the observed data with a value of at least $x_{\min}$, respectively. Clauset et al. used the likelihood ratio test (LR test) for comparing different fit models to the power law model with respect to the goodness of fit.

3. Data analysis

In this section, we first elaborate how we acquire the data, conduct preprocessing, and extract the relevant information. In particular, we explain how to model the transaction relations between a wide variety of Ethereum accounts under the framework of the complex network theory. Then, we conduct sets of experiments to study the statistical and structural properties of the transaction network. Specifically, we obtain the aforementioned measurements of the transaction network. The results enable us to gain more physical insights into the dynamics of Ethereum and understand how Ethereum works as a blockchain-based system. Our work can inspire related researchers to further improve Ethereum in terms of its security, performance, robustness, and many other aspects.

3.1. Data acquisition and preprocessing

Ethereum is widely known as a blockchain 2.0 platform that supports not only trusted cryptocurrency transfer but also smart contracts. It comes with a built-in Turing-complete programming language, enabling users to create, deploy, and run a wide variety of smart contracts. A transaction in Ethereum can indicate not only cryptocurrency transfer but also the execution of decentralized smart contracts. An Ethereum transaction is a signed data package targeted at a recipient. As shown in Fig. 1, the transaction typically consists of the nonce field of the signer's account, the address of the recipient, the Ethereum coin units (termed as "ether") to be transferred, an extra data field for contract execution, the signature, and some other useful fields. The nonce field indicates the total count of transactions already executed by this account. One can determine the sender's address from the signature field. In this paper, we focus on three critical fields of the transaction, namely the address of the sender, the address of the recipient, and the units of transferred cryptocurrency. These three fields describe how many cryptocurrency units flow in the Ethereum network from the sender to the receiver.

As with other blockchain-based cryptocurrency, the data of Ethereum transactions are recorded in the blockchain, which is composed of a series of linked data blocks. Each data block includes a number of transactions and some other information. Thus far, the entire Ethereum blockchain contains more than 5 million blocks, equivalent to 300 Gigabytes. Further statistics can be found elsewhere https://www.etherchain.org. To get the details of transactions of the Ethereum network, we set up

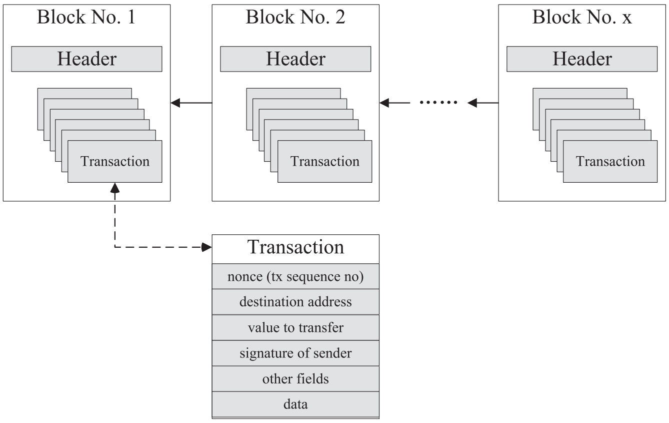

Fig. 1. Schematic illustration of the data structure of the Ethereum blockchain.

the Ethereum client "Geth" [16] in the full mode and synchronize it with the main network. We collect the raw data of all up-to-date blocks and then parse all transactions from these blocks with a script written in Python with the web3 interface provided by Ethereum [17].

For comparison, we select two data sets for analysis. Data set I contains 100 thousands blocks with serial numbers from 200000 to 300000, and more than 680,000 transactions. Data set II contains 610,000 transactions extracted from blocks having serial numbers from 3000000 to 3200000. The former data set represents the early stage of Ethereum, while the latter represents the up-to-date development. We describe each transaction by a vector $(i, j, a_{i,j})$, which means that $a_{i,j}$ units of Ethereum coins are transferred from address i to address j. Then, we can model the transaction relations by the directed weighted network, where node i represents address i. The link weight is the sum of the transferred units of Ethereum coins during the sampled time period. In the following, we start our investigation with an undirected or unweighted counterpart of a directed weighted network, as some measurements are well defined only for an undirected or unweighted graph.

Fig. 2 shows different types of graph representations of the transaction relations. The nodal label represents the address of an Ethereum account, while the weight of a link denotes the units of transferred Ethereum coins. The Ethereum address is usually denoted by a number in hexadecimal format. However, for readability, we denote the address in decimal format, as shown in Fig. 2. Here, we omit the temporal information of a transaction, i.e., when the transaction is carried out. Instead, we focus on the time-average topology. In other words, the topology concerned is considered static rather than temporal. Eventually, we get the network representation of the transaction relation over the sampling period. Fig. 3 shows the topology of the transaction network for data set I.

3.2. Component structure of undirected unweighted graph

We start with the simplest modeling method in order to quickly get an overview of the topology formed by the transactions during the sampling period. Specifically, we treat the transaction relations extracted from the sampling data as an undirected unweighted network and investigate the sizes of the undirected components. The undirected unweighted network is composed of $N = 5007$ nodes and $L = 8517$ links for data set I, while it is composed of $N = 300914$ and $L = 342012$ for data set II. The component structure analyses provide an overview of the entire network and show remarkable similarity for two data sets. Specifically, the component structure analyses demonstrate that the network as a whole is composed of a giant component and a number of finite components. As defined above, the giant component is the largest connected subgraph scaled with $O(N)$, while the finite components are small fragments. The component size distribution of the entire network is also calculated. The fact that more than 80% of the nodes are reachable from one another by following incoming or outgoing directed links implies that most nodes are weakly connected in the directed graph. If one tracks the transaction relation forward or backward, the transaction network is well connected. The history of all transactions is well recorded in the blockchain, which makes it possible to trace and audit the transaction features. The well auditable property of the Ethereum platform contributes toward establishing a public blockchain-based economic system with high-level security.

Remarkably, the distribution of the finite component size exhibits a heavy-tailed property and can be approximated by the power law model. The experimental results are shown in the $\log - \log$ plots in Fig. 4, where the inner graphs are the cumulative distributions. The scattered markers denote the experimental results while the straight dashed line denotes the fit model. The circle located in the bottom right corner represents the giant component, while the others denote the finite components. We investigate the finite components by omitting the giant component. Notably, the end segment of the component size distribution deviates from the "pure" power law, indicating that the finite components with large size have a tail heavier than the pure power law and may follow a different distribution. As the large-size finite components

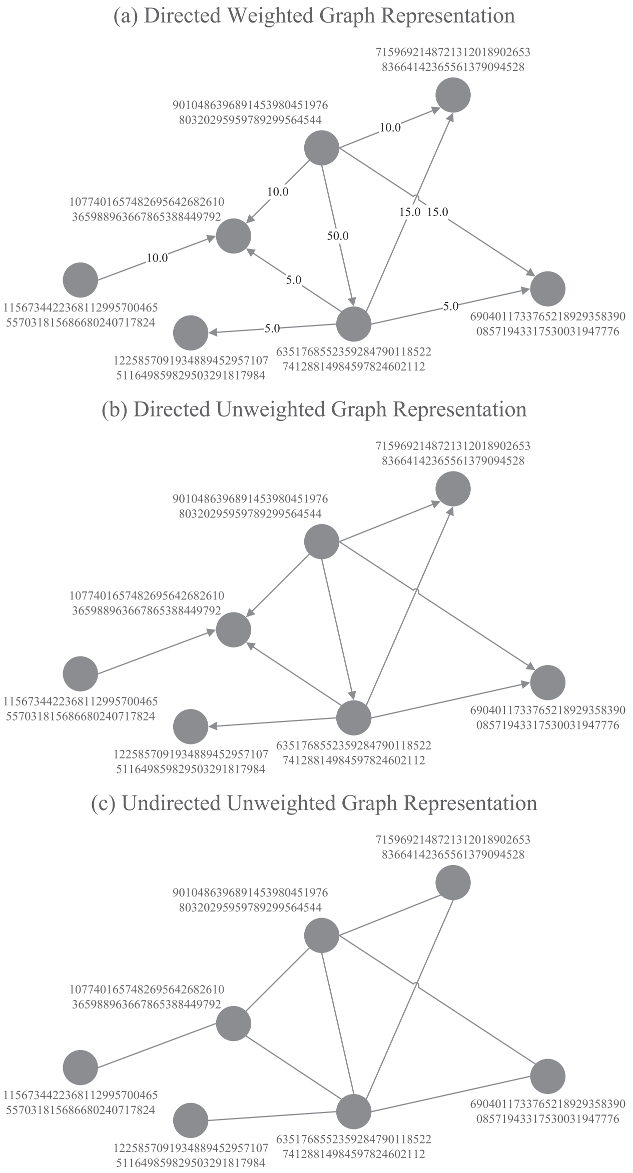

Fig. 2. Graph (network) representation of the transaction relation between different Ethereum addresses: (a) directed weighted graph, (b) directed unweighted graph, and (c) undirected unweighted graph.

constitute only a small portion of all finite components, the power law model is still a good fit to the entire observations. In summary, the finite component size exhibits the power law pattern, although the exponent is a time variant.

The random graph theory [31] can be adopted to theoretically interpret the finite component size distribution of the transaction network. However, the empirical evidence that we have obtained implies that the random graph framework might not be suitable for modeling the Ethereum transaction network. The reason is that the finite component size distributions of random graphs [34] tend not to follow the power law form and do not have a heavy tail as reported in [31]. By contrast, the finite component size distribution of the transaction network has a tail that is much heavier than the power law.

3.3. Bow-tie structure of directed unweighted graph

The analysis of the undirected graph provides some insights into the overall topological structure of the transaction network. Specifically, there is a giant connected component that consists of more than 80% of the nodal set. To obtain additional details of the topology, we further focus on the directed unweighted graph of the transaction relations and try to study the component structure. In the case of the directed graph, the component structure becomes complicated compared

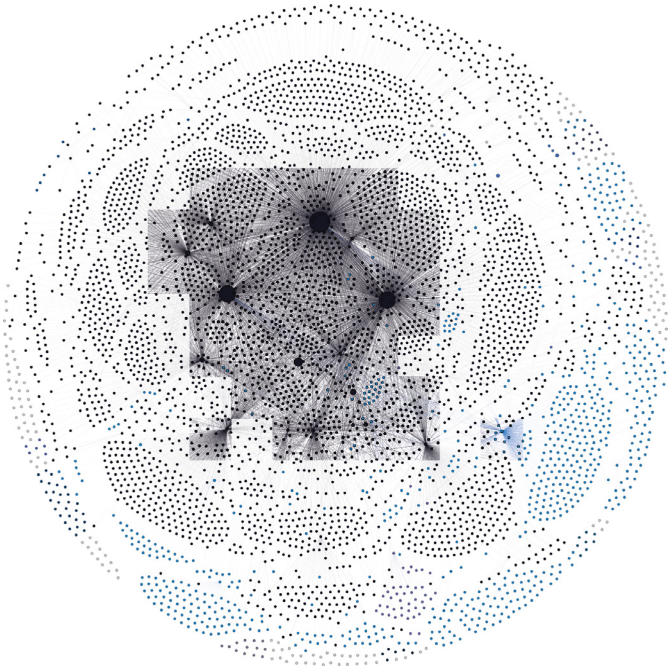

Fig. 3. Pictorial representation of the transaction network topology for data set I.

with its undirected counterpart. One may wonder whether there exists some pattern in the topology and whether this pattern is stable with respect to the data set.

We find that the component structure follows a bow-tie-like pattern. Fig. 5 shows a schematic illustration of the bow-tie structure of the directed unweighted transaction network concerned. The network is composed of a GWCC and some finite components. Actually, the GWCC of the directed network is indeed the giant component of its undirected counterpart. In the GWCC of the directed network, one can identify the SCC that can be considered as the core. As defined previously, the SCC is the maximal subgraph such that there exist both a path $P_{i\to j}$ and a path $P_{j\to i}$ for each pair of nodes in SCC. Furthermore, we find that there are some other types of components surrounding the core, i.e., the in-component, the out-component, the tube component (TUBE), and the tendrils. The in-component (IN) and the out-component (OUT) are the sets of nodes reachable to and from the SCC, respectively. Another intriguing component is the tube component, which can be regarded as bridges from the IN component to the OUT component. The remaining nodes belong to either the tendrils or the fragments disconnected from the GWCC.

There are several algorithms for calculating the SCC set. After identifying the SCC set, other sets can be found by graph search algorithms such as BFS (breath first search) along with some inferences. For example, we can start by determining the IN set. A node i belonging to the IN set is such that any node in SCC is reachable from i, as the nodes in SCC are defined to be reachable only from the nodes in the IN set. A node is in the IN + SCC set if it can reach any node in the OUT set. Thus, having already found the nodes in SCC, we can identify nodes in IN and OUT. Eventually, we can subtract the IN, OUT, and SCC sets from the GWCC set to get the tendrils, where the GWCC is solved by identifying the giant connected component in the transaction network regarded as an undirected graph.

For the data sets concerned, the measurements of the directed unweighted transaction network are calculated and stated in the following. For data set I, the network is composed of $N = 5007$ nodes and $L = 8710$ directed links. The GWCC consists of the SCC (898 nodes), the IN component (1766 nodes), the OUT component (1326 nodes), the TUBE (21 nodes), and a number of tendrils (725 nodes). For data set II, the network is composed of $N = 300914$ nodes and $L = 342012$ directed links. The GWCC consists of the SCC (6865 nodes), the IN component (47447 nodes), the OUT component (38218 nodes), the TUBE (16091 nodes), and a number of tendrils (78212 nodes). The remainder constitutes the other nodes in GWCC and fragments disconnected from GWCC. The experimental results indicate that the GWCC follows a bow-tie structure with a tightly large IN component and a narrow tunnel between the IN and OUT components.

3.4. Degree distribution and connectivity pattern

We investigate the distributions of the out-degree and the in-degree by running experiments on both the data sets concerned. The in-degree and the out-degree correspond to the numbers of incoming and outgoing transactions, respectively. The degree distributions give us a macro-scale view of the transaction relations. If most users tend to transfer the Ethereum coins to a specific number of "close friends", the observations over nodal degrees can be approximated by the Poisson model.

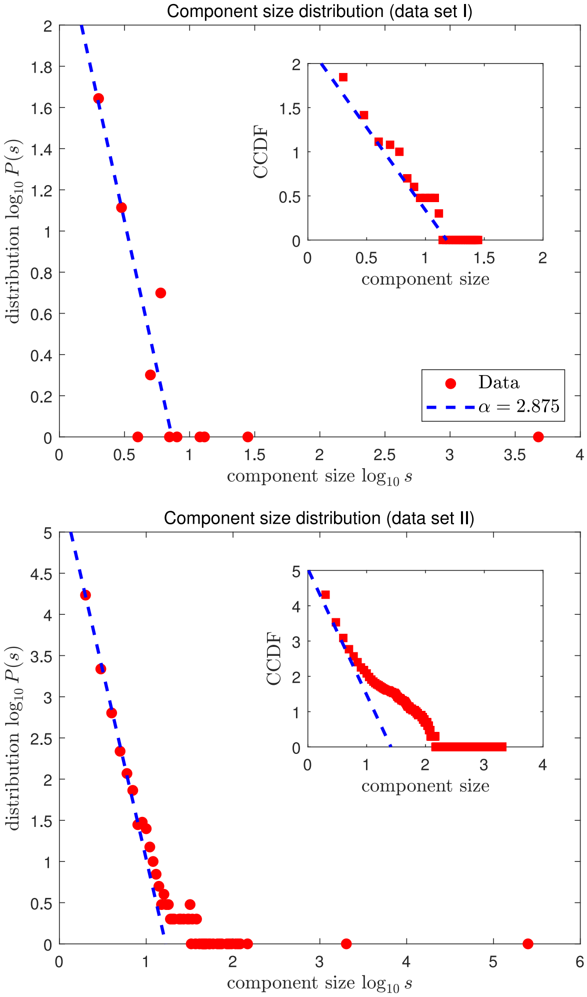

Fig. 4. Component size distributions $P(s)$ of the undirected graph for data set I and data set II. The inner graphs are the complementary cumulative distributions.

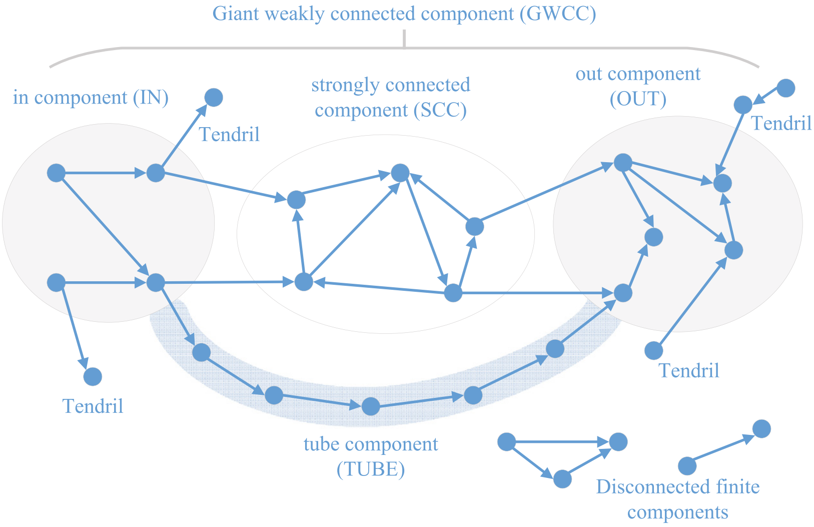
Giant weakly connected component (GWCC)
in component (IN)
Tendril
strongly connected component (SCC)
out component (OUT)
Tendril
Tendril
Tendril
tube component (TUBE)
Disconnected finite components

Fig. 5. Schematic illustration of the component structure of the directed network.

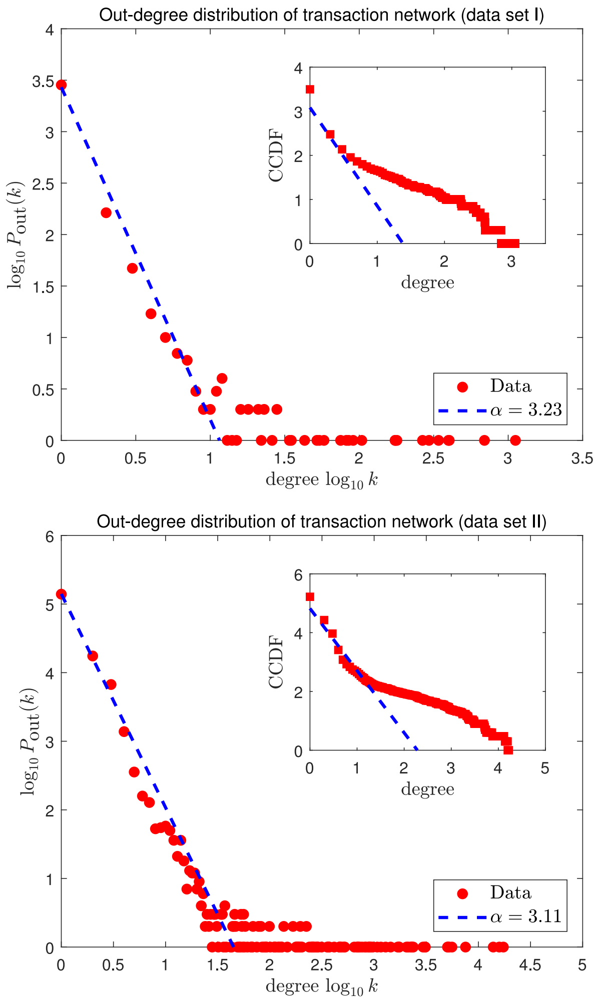

Fig. 6. Transaction networks in different time periods show similarity in out-degree distributions $P_{\text{out}}(k)$. The inner graphs represent the complementary cumulative degree distributions.

The experimental results are shown in Figs. 6 and 7. The observations, cumulative distributions, and corresponding fits are illustrated. In all the log-log plots, the power law model provides a reasonable fit to observations in the sense of statistics, although the ranges of the x-axis vary. In the following, we consider the in-degree distributions as an example. The in-degree distributions exhibit a power law, although the exponent is different over different data sets. Note that the end segment (i.e., the tail) of the observations deviates from the pure power law. This bias shows a tail that is heavier than the pure power law and is likely to follow a different distribution. Nevertheless, the power law model is a good fit to the observations, as the number of nodes with large degrees is relatively small and does not affect the fit significantly. Further investigation is required to better understand the in-degree distribution with complex patterns. The LR tests show that the pure power law is superior to the exponential, log-normal, and truncated power laws, suggesting an extremely heavy tail. It might be approximated by a piecewise power law or a combination of the power law and some other distribution, as suggested by the convexity of the deviation. We suspect that this anomalous turnover in the distribution may result from cryptocurrency exchanges. Specifically, we assume that the initial segment of the distribution denotes private cryptocurrency users while the end segment denotes influential cryptocurrency exchanges and investors.

In summary, we can say that the distributions of the in-degree and out-degree share the same pattern. Both these distributions can be approximated by the power law model, albeit with different exponents. Furthermore, the degree distributions show noticeable similarity for the two data sets, suggesting that the power law is a time-invariant connectivity pattern. It is reasonable to consider that there exists a type of structural similarity between networks extracted from distinct data sets with respect to the degree distribution. Actually, the degree distribution as a measurement is only one way to measure the similarity of two networks. Many other methods can be used to measure the similarity of two networks [14]. Moreover, the distributions exhibit a tail heavier than common heavy-tailed functions such as the log-normal and the power law functions.

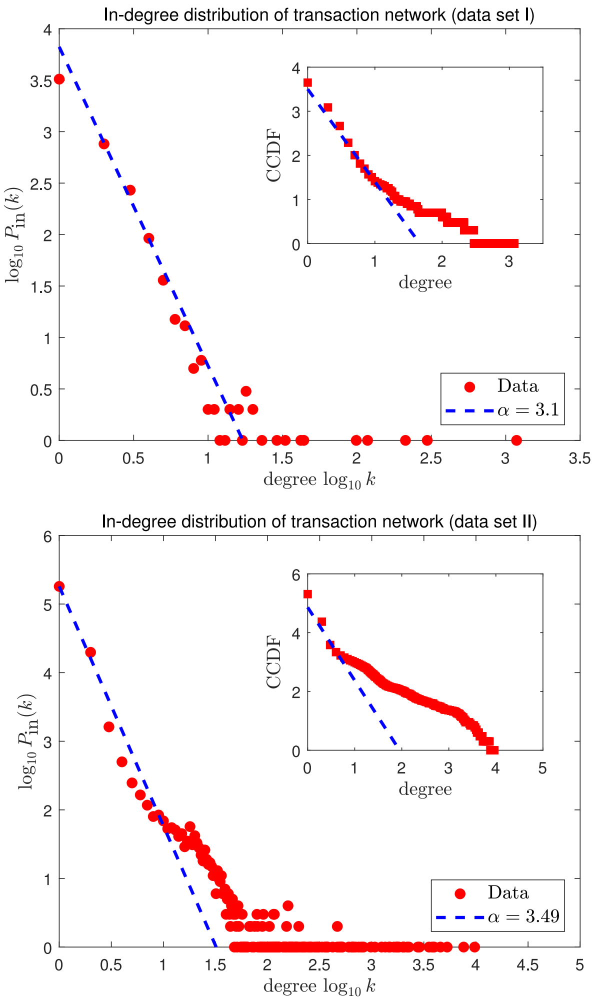

Fig. 7. In-degree distributions $P_{\text{in}}(k)$ show noticeable similarity for the two data sets. The inner graphs are the corresponding complementary cumulative degree distributions.

Meanwhile, the network with the power law degree distribution is reported to be extremely vulnerable to intentional attacks. Once the hubs, i.e., the nodes with high degrees, are controlled or damaged, the entire network will crash. The power law connectivity pattern is unsatisfactory in the sense that it ruins the design philosophy of the public blockchain platform. The design philosophy of public blockchains such as Ethereum and Bitcoin is decentralization. The endeavor to build a free, decentralized system is ruined by the appearance of trading centers and mining pools, which tend to have stronger connectivity than others.

The assortativity coefficient r is also calculated for the data sets concerned. The results are $r = -0.3525$ and $r = -0.2122$, respectively. A negative value of assortativity indicates negative degree correlation. Specifically, nodes with large degrees tend to trade with nodes with small degrees. Although there exist many cryptocurrency exchanges, indicating some kind of centralized tendency, the transaction network tends not to exhibit the “rich club” phenomenon.

3.5. Diameter and shortest path for undirected graphs

The average hopcount (i.e., average shortest path length for undirected graphs) and the diameter are (3.64, 12) and (5.41, 134) for the two data sets concerned. This indicates that the transaction network is the so called small-world network with $\bar{H} \sim \ln N$, although the diameter increases at a faster rate. The typical distance between any pair of anonymous users is extremely small and the ether paid by one node may return at a relatively high speed. In other words, the relatively low average shortest path indicates that the circulation period might be small. As a result, the Ethereum cryptocurrency has good liquidity. We suspect that this is because of the systematic design philosophy of restricted money supply of public blockchain platforms such as Bitcoin and Ethereum. The rate of supplying newborn “money” in these blockchain platforms is designed to decline. It is widely believed that this might result in the so-called deflation as well as the collapse of

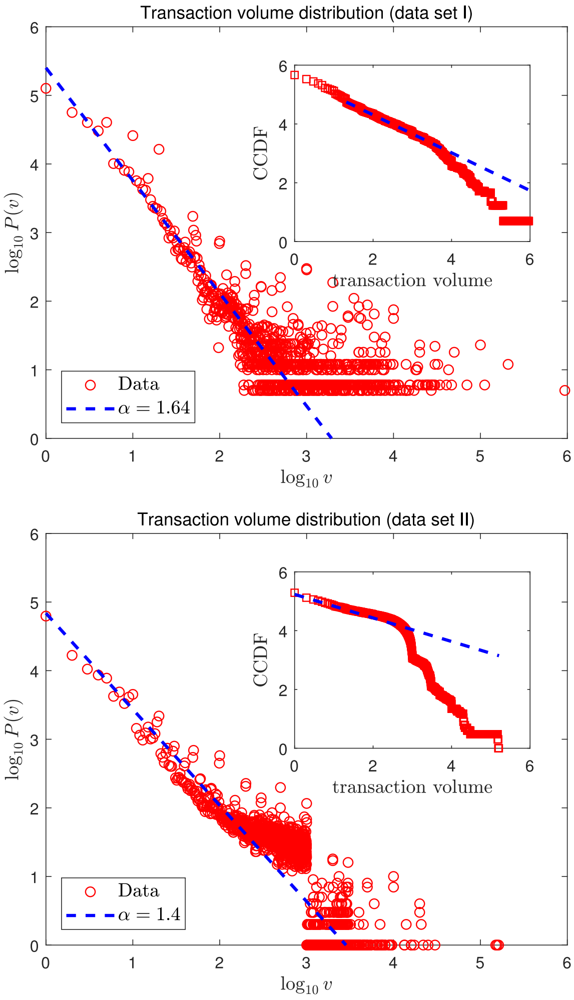

Fig. 8. Distributions $P(v)$ and complementary cumulative distributions of the transaction volume (i.e., link weight).

cryptocurrency as “money”. The existence of numerous exchanges also contributes to the liquidity, although this leads to centralization to some extent.

3.6. Transaction volume

The study of the transaction volume refers to the distribution of the transferred units of trading coins (i.e., the ethers). In other words, we focus on the link weight distribution of the weighted network and try to identify some time-invariant properties of the transactions. The transaction volume distributions as well as the fit models for both data sets concerned are shown in Fig. 8. The transaction volume distributions for the data sets concerned can be approximated by the power law distribution. The LR test results show that the power law model is superior to the exponential model but inferior to the log-normal and cut-off power law models. Thus, it seems reasonable to consider that there exists moderate support for the power law model with a cut-off as a good fit for the link weight distribution of the weighted network.

What we have found above visually shows the vast diversity of the transaction volume, which means that most of the users have limited incomings while a small population involves a large amount of digital money transferred. The power law distribution with a cut-off indicates a noticeable inequality between the users of the Ethereum coins to some extent. Note that small payments constitute the majority of the transactions. This may imply that Ethereum is a type of the so-called micro-payment system. Meanwhile, the energy cost for a small payment is equivalent to that for a large amount. As a result, the energy consumption for making a deal is relatively high for public blockchain systems such as Ethereum and Bitcoin. In this sense, Ethereum, as an international payment system, might not be comparable with centralized payment systems such as Visa and PayPal with respect to energy savings.

4. Conclusion and open issues

To the best of our knowledge, this study is the first to investigate the data of transactions recorded in the Ethereum blockchain and probe the statistical laws of the data from the perspective of network science. By means of the framework of network science, the transaction relations between different user accounts were regarded as a graph. Some critical network measurements were introduced to measure the statistical properties of the graph, which provided physical insights into the transaction relations. We found that several transaction features, such as the transaction volume, transaction relation, and component structure of the graph, exhibit the same heavy-tail property and can be approximated statistically by the power law function. Although the tail is heavier than the "pure" power law, it still provides a good fit to the data set as a whole. Notably, we found that the transaction relations follow a bow-tie structure if they are regarded as a directed graph. Moreover, there is no "rich club" phenomenon in the transaction relations, implying that the popular hubs tend to connect to a large number of common users. We believe that the above-mentioned statistics can be ascribed to the vast diversity in transactions and the existence of a number of cryptocurrency exchanges.

Directions for future work include inferring the statistical properties of the entire history data, investigating the temporal property of the transaction network [22], and gaining a better understanding of the complex interaction between the transaction network and the social network [23]. Specifically, our further research will focus on the temporal properties of transactions as well as applications for improving the security and performance of the blockchain. Understanding the working of the socio-economic network consisting of human beings and cryptocurrency platforms may provide us with a better understanding of blockchain economics. Moreover, we will provide the structural Ethereum data extracted using our developed program as well as the Python code to the public.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments and suggestions. We would also like to thank Clauset et al. for providing the MATLAB and Python code for testing the goodness of fit of the power law model. This work was funded in part by the National Natural Science Foundation of China (Grant No. 61702439), Shandong Provincial Natural Science Foundation, China (Grant No. ZR2017BF018), and China Postdoctoral Science Foundation (Grant No. 2017M620786).

References

[1] N. Atzei, M. Bartoletti, T. Cimoli, A survey of attacks on ethereum smart contracts sok, in: Proceedings of the 6th International Conference on Principles of Security and Trust - Volume 10204, Springer-Verlag New York, Inc., New York, NY, USA, 2017, pp. 164–186, doi:10.1007/978-3-662-54455-6_8.
[2] A.-L. Barabási, Scale-free networks: a decade and beyond, Science 325 (5939) (2009) 412–413, doi:10.1126/science.1173299. http://science.sciencemag.org/content/325/5939/412.full.pdf.
[3] A.-L. Barabási, Network science: luck or reason, Nature 489 (7417) (2012) 507, doi:10.1038/nature11486.
[4] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512, doi:10.1126/science.286.5439.509. http://science.sciencemag.org/content/286/5439/509.full.pdf.
[5] A.-L. Barabási, et al., Network Science, Cambridge university press, 2016.
[6] Blockstack, Blockstack, 2019. (https://blockstack.org/). Accessed: 2019-04-05.
[7] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J.A. Kroll, E.W. Felten, Sok: research perspectives and challenges for bitcoin and cryptocurrencies, in: 2015 IEEE Symposium on Security and Privacy, 2015, pp. 104–121, doi:10.1109/SP.2015.14.
[8] S. Cao, G. Zhang, P. Liu, X. Zhang, F. Neri, Cloud-assisted secure ehealth systems for tamper-proofing ehr via blockchain, Inf. Sci. 485 (2019) 427–440, doi:10.1016/j.ins.2019.02.038.
[9] I. Castro, A. Panda, B. Raghavan, S. Shenker, S. Gorinsky, Route bazaar: automatic interdomain contract negotiation, 15th Workshop on Hot Topics in Operating Systems (HotOS XV), USENIX Association, Kartause Ittingen, Switzerland, 2015.
[10] A. Clauset, C.R. Shalizi, M. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703, doi:10.1137/070710111.
[11] S. Davidson, P. De Filippi, J. Potts, Economics of blockchain, 2016, doi:10.2139/ssrn.2744751.
[12] T.T.A. Dinh, R. Liu, M. Zhang, G. Chen, B.C. Ooi, J. Wang, Untangling blockchain: a data processing view of blockchain systems, IEEE Trans. Knowl. Data Eng. 30 (7) (2018) 1366–1385, doi:10.1109/TKDE.2017.2781227.
[13] A. Doll, Btctrackr, 2014. (https://github.com/adoll/btctrackr). Accessed: 2019-04-05.
[14] F. Emmert-Streib, M. Dehmer, Y. Shi, Fifty years of graph matching, network alignment and network comparison, Inf. Sci. 346-347 (2016) 180–197, doi:10.1016/j.ins.2016.01.074.
[15] P. Erdos, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hungarian Acad. Sci. 5 (1) (1960) 17–60.
[16] Ethereum, Geth, 2019. (https://github.com/ethereum/go-ethereum). Accessed: 2019-04-05.
[17] Ethereum, Python web3, 2019, (https://github.com/ethereum/web3.py). Accessed: 2019-04-05.
[18] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, ACM SIGCOMM Comput. Commun. Rev. 29 (4) (1999) 251–262, doi:10.1145/316194.316229.
[19] D. Garcia, C.J. Tessone, P. Mavrodiev, N. Perony, The digital traces of bubbles: feedback cycles between socio-economic signals in the bitcoin economy, J. R. Soc. Interface 11 (99) (2014), doi:10.1098/rsif.2014.0623. http://rsif.royalsocietypublishing.org/content/11/99/20140623.full.pdf.
[20] A. Gervais, G.O. Karame, K. Wüst, V. Glykantzis, H. Ritzdorf, S. Capkun, On the security and performance of proof of work blockchains, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, in: CCS '16, ACM, New York, NY, USA, 2016, pp. 3–16, doi:10.1145/2976749.2978341.
[21] Y. Gilad, R. Hemo, S. Micali, G. Vlachos, N. Zeldovich, Algorand: scaling byzantine agreements for cryptocurrencies, in: Proceedings of the 26th Symposium on Operating Systems Principles, in: SOSP '17, ACM, New York, NY, USA, 2017, pp. 51–68, doi:10.1145/3132747.3132757.
[22] P. Holme, J. Saramäki, Temporal Networks as a Modeling Framework, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–14. doi:10.1007/978-3-642-36461-7_1
[23] D.Y. Kenett, J. Gao, X. Huang, S. Shao, I. Vodenska, S.V. Buldyrev, G. Paul, H.E. Stanley, S. Havlin, Network of Interdependent Networks: Overview of Theory and Applications, Springer International Publishing, Cham, pp. 3–36. doi:10.1007/978-3-319-03518-5_1.

[24] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne, Computational social science, Science 323 (5915) (2009) 721–723, doi:10.1126/science.1167742. http://science.sciencemag.org/content/323/5915/721.full.pdf.

[25] J. Li, J. Wu, L. Chen, Block-secure: blockchain based scheme for secure p2p cloud storage, Inf. Sci. 465 (2018) 219–231, doi:10.1016/j.ins.2018.06.071.

[26] C. Lin, D. He, X. Huang, X. Xie, K.-K.R. Choo, Blockchain-based system for secure outsourcing of bilinear pairings, Inf. Sci. (2018), doi:10.1016/j.ins.2018.12.043.

[27] Y. Ma, S. Cao, Y. Shi, I. Gutman, M. Dehmer, B. Furtula, From the connectivity index to various randić-type descriptors, MATCH Commun. Math. Comput. Chem. 80 (1) (2018) 85–106.

[28] S. Nakamoto, Bitcoin: apeer-to-peer electronic cash system, 2019. (http://bitcoin.org/bitcoin.pdf). Accessed: 2019-04-05.

[29] A. Narayanan, J. Bonneau, E. Felten, A. Miller, S. Goldfeder, Bitcoin and Cryptocurrency Technologies: a Comprehensive Introduction, Princeton University Press, 2016.

[30] N. Narula, W. Vasquez, M. Virza, zkledger: privacy-preserving auditing for distributed ledgers, in: 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18), USENIX Association, Renton, WA, 2018, pp. 65–80.

[31] M. Newman, Component sizes in networks with arbitrary degree distributions, Phys. Rev. E 76 (2007) 045101, doi:10.1103/PhysRevE.76.045101.

[32] M. Newman, Networks, Oxford university press, 2018.

[33] M.E. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (2002) 208701, doi:10.1103/PhysRevLett.89.208701.

[34] M.E. Newman, S.H. Strogatz, D.J. Watts, Random graphs with arbitrary degree distributions and their applications, Phys. Rev. E 64 (2001) 026118, doi:10.1103/PhysRevE.64.026118.

[35] O. Novo, Blockchain meets iot: an architecture for scalable access management in iot, IEEE Int. Things J. 5 (2) (2018) 1184–1195, doi:10.1109/JIOT.2018.2812239.

[36] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Modern Phys. 87 (2015) 925–979, doi:10.1103/RevModPhys.87.925.

[37] A. Pentland, D. Shrier, T.H.I. Wladawsky-Berger, Towards an internet of trusted data: a new framework for identity and data sharing, Massachusetts Institute of Technology. Input to he Commission on Enhancing National Cybersecurity, 2016.

[38] R.S. Portnoff, D.Y. Huang, P. Doerfler, S. Afroz, D. McCoy, Backpage and bitcoin: uncovering human traffickers, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '17, ACM, New York, NY, USA, 2017, pp. 1595–1604, doi:10.1145/3097983.3098082.

[39] D. Ron, A. Shamir, Quantitative analysis of the full bitcoin transaction graph, in: A.-R. Sadeghi (Ed.), International Conference on Financial Cryptography and Data Security, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 6–24, doi:10.1007/978-3-642-39884-1_2.

[40] P. Tasca, A. Hayes, S. Liu, The evolution of the bitcoin economy: extracting and analyzing the network of payment relationships, J. Risk Finance 19 (2) (2018) 94–126, doi:10.1108/JRF-03-2017-0059.

[41] S. Valfells, J.H. Egilsson, Minting money with megawatts [point of view], Proc. IEEE 104 (9) (2016) 1674–1678, doi:10.1109/JPROC.2016.2594558.

[42] P. Van Mieghem, Performance Analysis of Complex Networks and Systems, Cambridge University Press, 2014.

[43] G. Wood, Ethereum: a secure decentralised generalised transaction ledger, 2019. (http://gavwood.com/paper.pdf). Accessed: 2019-04-05.

[44] Y. Zhang, R.H. Deng, X. Liu, D. Zheng, Blockchain based efficient and robust fair payment for outsourcing services in cloud computing, Inf. Sci. 462 (2018) 262–277, doi:10.1016/j.ins.2018.06.018.