An Information Entropy Method to Quantify the Degrees of Decentralization for Blockchain Systems

Keke Wu
Shenzhen Institute of Information Technology, Shenzhen, 518172, China
wukk@sziit.edu.cn

Bo Peng, Hua Xie, Zhen Huang
Shenzhen Silico Design Technology Co., LTD, Shenzhen, 518057, China
{pengbo & xiehua & huangzhen}@silico.design

***Abstract*—Decentralization is a key selling point of most public blockchain platforms. However, despite the widely acknowledged importance of this property, most researches on this topic lack quantification, and none of them performs a calculation on the degrees of decentralization they achieve in practice. In this paper, taking Bitcoin and Ethereum for instances, we propose an entropy method in information theory to quantify the decentralization for them. Using the information entropy, we calculate the discrete degrees of blocks mined and address balances to quantify the degrees of decentralization for Bitcoin and Ethereum systems, and the results of calculations indicate that Bitcoin's mining is more approximately 12% decentralized than Ethereum with full samples, and Bitcoin's wealth is more approximately 9.2% decentralized than Ethereum with 10,000 samples. Our method can be used to evaluate the degree of decentralization for any blockchain system.**

Keywords—decentralization; blockchain; Bitcoin; Ethereum; information entropy

I. INTRODUCTION

Bitcoin is a digital currency implementation based on blockchain technology that was invented by Satoshi Nakamoto in 2008 [1]. Bitcoin is the first digital currency system that has been tested in large scale and long time in history. As a public blockchain platform, for adapting to more complex and flexible application scenarios, Ethereum [2] has further extended the functions of Bitcoin for digital currency transactions, supporting the important feature of smart contract. The common advantage of blockchain systems (such as Bitcoin and Ethereum) is widely understood to be decentralization that does not have any central authority or server and their networks are peer-to-peer. By storing data across its decentralized network, the blockchain eliminates a number of risks that come with data being held centrally.

As we all know, Bitcoin uses the consensus mechanism of Proof of Work (POW). In POW algorithm, how many bitcoins you get depends on the effective work of your mining. The better the computer is, the more power it has, and the longer it takes to mine, the more bitcoins you get. The disadvantage of POW algorithm is that because the number of bitcoins is constant, the less bitcoins are mined, the higher the requirement of computing power is. People begin to speed up the acquisition of bitcoins through joint mining, and the pool mining is gradually rising. However, miners are scattered individuals, and individuals have no influence. Under such circumstances, the mine pool, which gathers the miners' calculation power, becomes the representative of the miners and gains the miners' voting rights. In this way, the problem of centralization of arithmetic becomes more and more prominent. The original intention of Bitcoin is to decentralize, but once the power of Bitcoin reaches 51%, it is a terrible thing, which means that users who control the power of 51% can forge transactions and control the network of Bitcoin to launch 51% attacks. For example, BitContinent has two pools, Antpool and BTC.com, which accounted for 41% of Bitcoin's online computing power in June 2018, threatening its de-centralization. With the rising price of Bitcoin and the considerable profits, more than 90% of Bitcoin's calculation power is concentrated in less than 20 enterprises, and more than 90% of Ethereum's calculation power is concentrated in less than 10 enterprises. Bitcoin and Ethereum's calculation power are monopolized, which is not a good phenomenon for the long-term development of block chain. Therefore, it is necessary to introduce quantitative research on the degree of decentralization to quantify the fairness of block chain network.

Since decentralization is one of the core technologies in blockchain, many researches about the decentralization were proposed. Croman and Gencer et al. proposed the technical evaluation of the blockchain decentralization system, mainly focusing on the network congestion or delay to evaluate the performance of the blockchain distributed network [3, 4]. Gervais et al. revealed that there are many important operations and decisions in Bitcoin system which is not decentralized, and they revealed that some nodes control services, decision-making, transactions and mining in Bitcoin system, finally they gave a way to optimize the decentralization of Bitcoin network [5]. Ron and Shamir analyzed the transaction data of Bitcoin and revealed the occurrence of large transactions in the Bitcoin system at a certain point in time [6]. The upper two papers are based on data analysis of Bitcoin transaction behavior to illustrate the drawbacks of the low extent of decentralization in Bitcoin system. Although these researches are mainly concerned about the decentralization of blockchain, none of them performs a measurement on the extent of decentralization they achieve in practice. The closest research work to ours is [7] that only focuses on evaluating a critical value of the number of nodes needed to control over 51% of the network by using a Nakamoto coefficient, rather than quantifying the discrete of a set of data of a blockchain target, such as blocks mined and address balance and so on.

978-1-7281-1190-2/19/$31.00©2019 IEEE

Herein, we must be able to measure the data discrete degrees of the targets of nodes in blockchain systems before we improve the decentralization. In this paper, we propose an information entropy method to measure and quantify the degrees of decentralization for blockchain systems. Using the information entropy, we measure the discrete degrees of blocks mined and address balances to quantify the degrees of decentralization for blockchain systems. The reminder of this paper is organized as follows. In section 2, we introduce the theory of the concept of information entropy, and the meanings of decentralization in blockchain systems. In section 3, we propose a quantitative model to measure the data discrete degree based on the information entropy, and in section 4, we calculate the discrete degrees of blocks mined and address balances by using the quantitative model. In section 5, we compare and analyze the results of the information entropy between Bitcoin and Ethereum systems. Finally, we conclude the paper in section 6.

II. BACKGROUND

In this section, we first introduce the theory of the information entropy, and then we illustrate the meanings of blockchain's decentralization.

A. Information Entropy

In 1948, Shannon, the founder of information theory, first proposed systematically the measurement of information. Influenced by Blotzmann's research method, he used probability and statistics method to take entropy as a measure of uncertainty and information quantity of random events, thus laying the scientific theoretical foundation of modern information theory.

Considering a randomized trial *X* with *n* possible results, uncertainty can be considered as a situation, that is, which of the *n* possible results will happen? The starting point of measuring information in information theory is to regard the amount of information obtained as the amount of uncertainty eliminated. The probability of random events can be described by probability distribution function. The probability distribution $p=(p_1,p_2,\cdots,p_n)$ of possible results is satisfied as below.

$$\sum_{i=1}^{n} p_i = 1 \text{, where } 0 \le p_i \le 1 . \tag{1}$$

Shannon introduced function:

$$H(X) = H(p_1, p_2, \cdots, p_n) = -k\sum_{i=1}^{n} p_i \log p_i . \tag{2}$$

This function is used as a priori uncertainty in the randomised test *X*. In formulua (2), $k \ge 0$ is a constant, and *H* is called as information entropy, or probability entropy or Shannon entropy, that is a measure of uncertainty in terms of probability distribution function. If *X* is interpreted as a set of *n* measures and p_i is the probability that the system is in the *i*-th micro-state, Shannon entropy and statistical mechanics entropy are the same.

Entropy is a measure of insufficient information or chaotic disorder of a system, or a measure of ignorance of a system. Let the complete information can be obtained, the entropy is equal to 0, otherwise, it will be greater than 0. For example, in random test *X*, let any $p_i = 1$, then $H = 0$. In this situation, we can make a decisive prediction of the test results without any uncertainty.

In the above formula (2), when the bottom of logarithm is 2, *e* and 10, the units of information are bit (Binary Digit), NAT (Natural Digit) and dit (Decimal Digit), respectively. In most cases, logarithms are used in the form of 2, and constant $k = 1$, i.e.,

$$H(X) = H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i . \tag{3}$$

The formula (3) is in exactly the same form as the microscopic interpretation formula of entropy given by Blotzmann. Information entropy can not only be used in the calculation of information quantity, but is essentially a universal mathematical measure of uncertainty. More theories about the information entropy, please refer to the entropy textbooks [8, 9].

B. Decentralization

In a general sense, decentralization means that not one single entity has control over all the processing. The so-called decentralization is essentially to reconstruct a new commercial social credit system. By creating more centers to remove the previous single center, it can promote the security and democracy of the commercial social credit system at the same time. In fact, it can also depress the credit cost of the commercial society. In blackchain systems, the decentralization means that no single individual can destroy transactions in the network, and any transaction request requires the consensus of most participants.

By definition, blockchains provide a basic level of decentralization because transactions are recorded by all users on the blockchain network. Any changes to the transaction record must be confirmed by the vast majority of blockchain users in order to be recognized as legitimate. If a single user or small group of users were to try to manipulate blockchain data in a way with which a majority of the network disagreed, their efforts would be thwarted by the rest of the network. This is one of the features that make blockchain technology so powerful.

It is also a feature that distinguishes blockchains in a crucial way from traditional databases, in which all records are stored in a central location. That location is usually controlled by a single party, who has the power to modify records without consulting anyone else. For privacy, this single party must be blindly trusted by everyone to protect the private information they store and not use it for gain without consent. Databases can and have been the single point of failure by which malicious actors attack, compromise, steal and profit off the private information they contain.

Bitcoin and Ethereum also have a peer-to-peer network for disseminating block and transaction information. Both Bitcoin and Ethereum also contain full nodes, which serve two critical roles: (1) to relay blocks and transactions to miners (2) and to answer queries for end users about the state of the blockchain.

In the Bitcoin and Ethereum protocols, users submit transactions for miners to sequence into blocks. Better decentralization of miners means higher resistance against censorship of individual transactions. Specifically, a decentralized system (such as Bitcoin or Ethereum) is composed of a set of decentralized subsystems (such as mining, exchanges, nodes, developers, clients, and so on). Srinivasan et al. used these six subsystems to calculate a critical value with a Nakamoto coefficient, and to illustrate how many nodes needed to control over 51% of the network in Bitcoin or Ethereum [7].

In this paper, we will calculate the discrete degrees by two targets (blocks mined and address balance) to measure the degrees of decentralization for Bitcoin and Ethereum systems. Please note: you may decide to use different subsystems or targets based on which ones you consider essential to decentralization of the system as a whole.

III. QUANTITATIVE MODEL

In information theory, as presented above, entropy is a measure of uncertainty. The larger the amount of information, the smaller the uncertainty and the smaller the entropy is. The smaller the amount of information, the greater the uncertainty and the greater the entropy is. According to the characteristics of entropy, we can judge the randomness or disorder degree of an event by calculating the entropy value. Therefore, we can also use the information entropy value to quantify the discrete degree of a data set. The greater the discrete degree of the data set, the smaller the randomness and disorder degree of the data set, then the larger the amount of information, the smaller the uncertainty and the smaller the entropy is. On the contrary, the smaller the discrete degree of the data set, the greater the randomness and disorder degree of the data set, then the smaller the amount of information, the greater the uncertainty and the greater the entropy is.

In blockchain systems, herein, we choose the blocks mined or address balance as a data set to calculate the degree of decentralization, respectively. The data set of blocks mined or address balance can be written as $X=(x_1,x_2,\cdots,x_n)$. For every element x_i in the data set, its probability event p_i is equivalent to the value of this element is divided by the sum of the data set, i.e.,

$$p_i=\frac{x_i}{s}\text{, where } s=\sum_{i=1}^{n}x_i\,. \tag{4}$$

Hence, we can establish equivalent model as follows,

$$H(X)=H(p_1,p_2,\cdots,p_n)\Leftrightarrow H(X)=H(\frac{x_1}{s},\frac{x_2}{s},\cdots,\frac{x_n}{s})\,, \tag{5}$$

According to the information entropy formula (3) above, we propose the information entropy model for a data set $X=(x_1,x_2,\cdots,x_n)$ as follows,

$$H(X)=-\sum_{i=1}^{n}\frac{x_i}{s}\log_2\frac{x_i}{s}=\sum_{i=1}^{n}\frac{x_i}{s}\log_2\frac{s}{x_i}\,. \tag{6}$$

The greater the entropy value, the smaller the discrete degree of the data set, the more randomness or disorder the degree of the data set, the greater the degree of decentralization is. On the contrary, the smaller the entropy value, the greater the discrete of the data set, the less randomness or the more order the degree of the data set, the smaller the degree of decentralization is. To be better decentralization, therefore, we need greater value of information entropy for any blockchain system.

In this paper, we use the information entropy model to calculate and quantify the degrees of decentralization for Bitcoin and Ethereum systems. We consider the decentralized degrees with the two targets: blocks mined, and address balance.

IV. CALCULATIONS

According to the quantitative model proposed above, let's now calculate information entropy values for the blocks mined and address balance in Bitcoin and Ethereum systems. We can see how decentralized each of them according to the quantitative model of information entropy.

A. Blocks Mined

For one mining pool, the quantity of blocks mined during a certain time period reflects the priority to account in blockchain systems. The data is more discrete (or more polarized), the ability of the pool controlling the entire blockchain system is more powerful, and the degree of decentralization of the blockchain system is lower. On the contrary, the data is more average, the ability of the pool controlling the entire blockchain system is weaker, and the degree of decentralization of the blockchain system is higher.

Information Entropy of Bitcoin Blocks Mined

We catch the data of Bitcoin blocks mined over the last 7 days from the website btc.com on Nov. 10, 2018, where Bitcoin's data will be updated in real time, as the following Figure 1. (Data from https://btc.com/ on Nov. 10, 2018 [10])

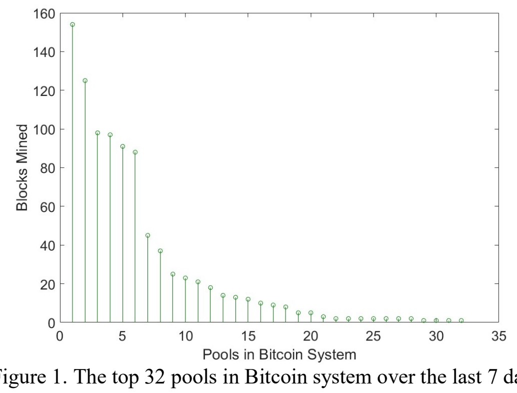

Figure 1. The top 32 pools in Bitcoin system over the last 7 days

There are all 32 pools at least 1 block mined in the entire Bitcoin network, and there are all 919 blocks mined in the entire Bitcoin network over the last 7 days on Nov. 10, 2018. We can see that the top 6 pools mined most blocks, and they can influence the decentralized degree of entire Bitcoin system.

According to the quantitative method proposed above, we use the data set X = {154, 125, 98, 97, 91, 88, 45, 37, 25, 23, 21, 18, 14, 13, 12, 10, 9, 8, 5, 5, 3, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1} with the sample number n = 32 to calculate the information entropy that is introduced above as follows.

$$s=\sum_{i=1}^{n}x_i=919\text{, where } x_i\in X. \tag{7}$$

$$\begin{aligned}H(X)&=-\sum_{i=1}^{n}\frac{x_i}{s}\log_2\frac{x_i}{s}=\sum_{i=1}^{n}\frac{x_i}{s}\log_2\frac{s}{x_i}\\&=\sum_{i=1}^{32}\frac{x_i}{919}\log_2\frac{919}{x_i}\\&\approx 3.82\end{aligned} \tag{8}$$

We obtain the value of information entropy of Bitcoin blocks mined is approximately equal to 3.82.

Information Entropy of Ethereum Blocks Mined

In the same way, we obtain the Ethereum block data over the last 7 days on Nov. 10, 2018 from the website etherscan.io, where Ethereum's data will be updated in real time, as the following Figure 2. (Data from https://etherscan.io/ on Nov. 10, 2018 [11])

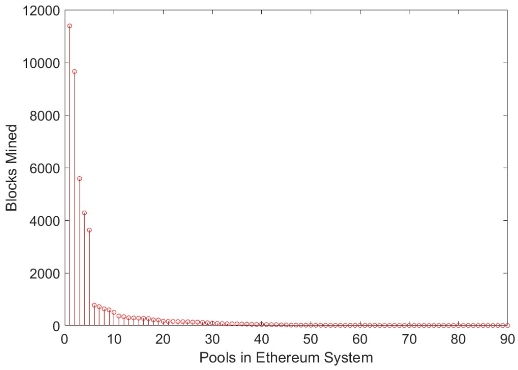

Figure 2. The top 90 pools in Ethereum system over the past 7 days

There are all 90 pools at least 1 block mined in the entire Ethereum network, and there are all 42682 blocks mined in the entire Ethereum network over the last 7 days on Nov. 10, 2018. We can see that the top 5 pools mined most blocks, and they can influence the decentralized degree of entire Ethereum system.

According to the quantitative method proposed above, in the same way, we use the data set X = {11386, 9651, 5585, 4280, 3627, 772, 715, 634, 595, 500, 360, 336, 295, 288, 280, 275, 265, 215, 212, 162, 157, 152, 148, 145, 142, 132, 130, 121, 104, 94, 78, 72, 67, 64, 63, 57, 53, 45, 41, 41, 35, 30, 30, 24, 19, 17, 16, 15, 13, 12, 11, 11, 8, 8, 8, 8, 7, 7, 7, 6, 6, 5, 5, 3, 3, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1} with the sample number n = 90 to calculate the information entropy that is introduced above as follows.

$$s=\sum_{i=1}^{n}x_i=42682\text{, where } x_i\in X. \tag{9}$$

$$\begin{aligned}H(X)&=-\sum_{i=1}^{n}\frac{x_i}{s}\log_2\frac{x_i}{s}=\sum_{i=1}^{n}\frac{x_i}{s}\log_2\frac{s}{x_i}\\&=\sum_{i=1}^{90}\frac{x_i}{42682}\log_2\frac{42682}{x_i}\\&\approx 3.41\end{aligned} \tag{10}$$

We obtain the value of information entropy of Ethereum blocks mined is approximately equal to 3.41.

B. Address Balance

This target lists the address of the top 10,000 sorted by balance. We believe that the decentralized blockchain should also decentralize wealth, and on the contrary, the more centralized tokens means that institutions or individuals with a large number of tokens are more likely to manipulate token prices.

Information Entropy of Bitcoin Address Balance

We catch the data of Bitcoin address balance (token) from the website bitinfocharts.com on Nov. 10, 2018 as the following Figure 3. (Data from https://bitinfocharts.com/top-100-richest-bitcoin-addresses.html on Nov. 10, 2018 [12])

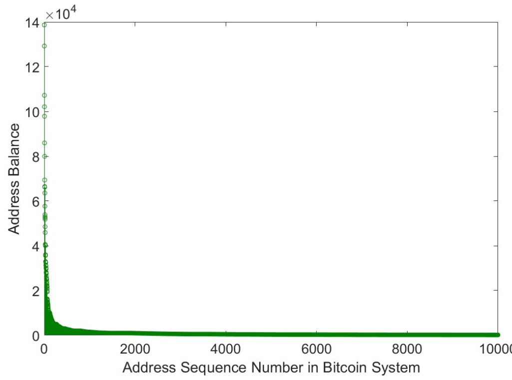

Figure 3. The top 10,000 address balance in Bitcoin system on Nov. 10, 2018

Herein we can ignore the decimal digits since the values of address balances are very huge. Therefore, we use the data set X = {138661, 129234, 107203, 102103, 97848, 85947, 79957, 69370, 66452, 66379, 66236, 66234, 63600, 57679, 53880, 53000, 52431, 51830, 48500, 45899, 40593, 40474, 40454, 40438, 40414, 40000, 36000, 35612, 32957, 32841, 32796, 32500, 32490, 31270, 31085, 31010, 31000, 30115, 30108, 29772, 29683, 28151, 27833, 27683, 27496, 26215, 25489, 25409, 25403, … …, 152, 152, 152, 152, 152, 152, 152, 152, 152, 152, 152} with the sample number n = 10,000 to calculate the information entropy as follows.

$$s=\sum_{i=1}^{n}x_i=10028525\text{, where } x_i\in X. \tag{11}$$

$$\begin{aligned}H(X)&=-\sum_{i=1}^{n}\frac{x_i}{s}\log_2\frac{x_i}{s}=\sum_{i=1}^{n}\frac{x_i}{s}\log_2\frac{s}{x_i}\\&=\sum_{i=1}^{10000}\frac{x_i}{10028525}\log_2\frac{10028525}{x_i}\\&\approx 11.33\end{aligned} \tag{12}$$

We obtain the value of information entropy of Bitcoin top 10,000 address balances is approximately equal to 11.33.

Information Entropy of Ethereum Address Balance

We also catch the data of Ethereum address balance from the website etherscan.io on Nov. 10, 2018 as the following Figure 4. (Data from https://etherscan.io/accounts on Nov. 10, 2018 [13])

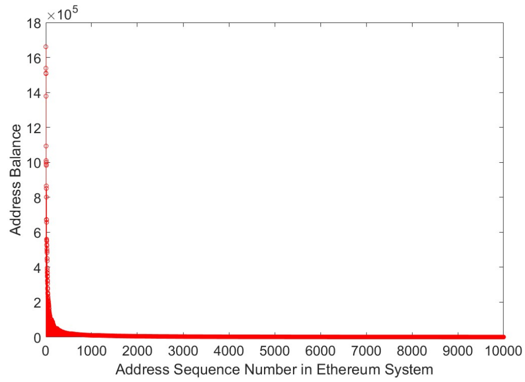

Figure 4. The top 10,000 address balances in Ethereum system on Nov. 10, 2018

Herein we still can ignore the decimal digits since the values of address balances are very huge. Therefore, we use the data set X = {1661213, 1538423, 1510066, 1507810, 1378754, 1093160, 1009999, 1000000, 988888, 982791, 865000, 850861, 801053, 672785, 672524, 670941, 657443, 560000, 558117, 552124, 549774, 530000, 523828, 505000, 493015, 483000, 450000, 450000, 450000, 436000, 395433, 382074, 369023, 365003, 350001, 345741, 325000, 319500, 306276, 281380, … …, 767, 767, 767, 767, 767, 767, 766, 766, 766, 766} and sample number n = 10,000 to calculate the information entropy as follows.

$$s=\sum_{i=1}^{n} x_i = 85780256\text{, where } x_i \in X\,. \tag{15}$$

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n}\frac{x_i}{s}\log_2\frac{x_i}{s} = \sum_{i=1}^{n}\frac{x_i}{s}\log_2\frac{s}{x_i} \\ &= \sum_{i=1}^{10000}\frac{x_i}{85780256}\log_2\frac{85780256}{x_i} \\ &\approx 10.38 \end{aligned} \tag{16}$$

We also obtain the value of information entropy of Ethereum address balances is approximately equal to 10.38.

V. Comparison and Analysis

We give the comparison of decentralization degrees between Bitcoin and Ethereum, and we analyze the centralized and decentralized influences in Bitcoin and Ethereum systems.

A. Blocks Mined Target

This target examines how many individual or organizational unions are needed to control more than 50% account power. For example, how many pools in PoW will add up to 50% of the total net power. This target intuitively reflects the difficulty of controlling a digital currency through 51% attacks. We believe that the more decentralized the blockchain, the less likely it is to control the entire blockchain by controlling a few individuals or organizations.

According to the results of calculations of information entropy above, the value of Bitcoin blocks mined entropy is larger than the value of Ethereum blocks mined entropy, as show in the following Figure 5. We can see that Bitcoin mining is more approximately 12% decentralized than Ethereum as measured by blocks mined over the past 7 days. Ethereum's mining is somewhat more centralized.

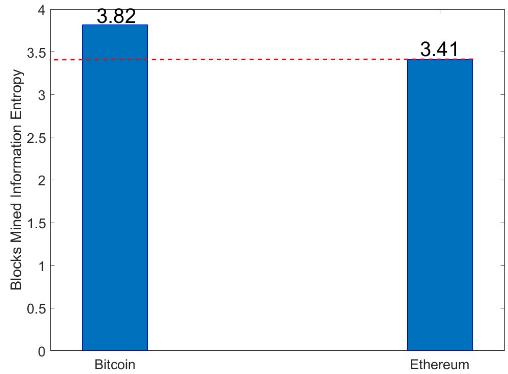

Figure 5. Information entropy of blocks mined comparison between Bitcoin and Ethereum

B. Address Balance Target

This target is a more controversial indicator, because many would say that addresses that appear to have a large number of tokens may be exchanges. Actually, those tokens are not exchanges, but are temporarily deposited on exchanges. However, Binance exchange was attacked recently, and which led to a sharp drop in the price of Bitcoin. Followed by MT. Gox bankruptcy, many trusts selling Bitcoin led to a sharp drop in prices and other incidents. Herein, we still believe that address balance decentralization is an important factor in the real decentralization of digital money.

According to the results of calculations of information entropy above, the value of Ethereum address balance entropy is also larger than the value of Bitcoin address balance entropy, as show in the following Figure 6. Hence, Bitcoin's wealth is more decentralized than Ethereum as measured by address balances. Ethereum's wealth is somewhat more centralized. We can see that Bitcoin address balance is more approximate 9.2% decentralized than Ethereum with 10,000 samples.

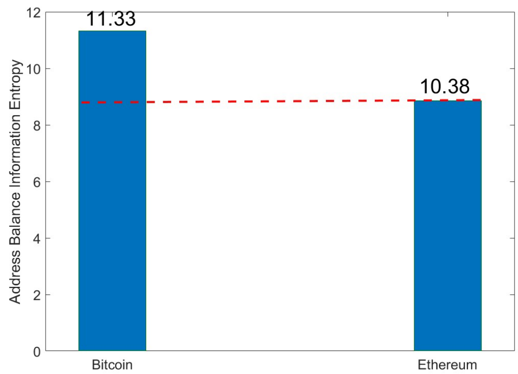

Figure 6. Information Entropy of address balances comparison between Bitcoin and Ethereum

VI. Conclusions

Many have said that decentralization is the most important property of systems like Bitcoin and Ethereum. If this is true, then it is critical to be able to quantify the degree of decentralization. More importantly, the information entropy method is one such general measurement method adapting to quantify the discrete degree for any data set. Therefore, our proposed information entropy quantitative model can be adapted to any blockchain system, and we can now calculate its information entropy for kinds of targets that you think they are important, and analyze whether this is plausibly a decentralization bottleneck for the system as a whole.

ACKNOWLEDGMENT

This work was supported by the Guangdong Natural Science Foundation (Grant No. 2018A030313746), and the Basic Research Project of Shenzhen (Grant No. JCYJ20170817114239348).

REFERENCES

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

[2] V. Buterin, Ethereum: a next generation smart contract and decentralized application platform, 2013, https://github.com/ethereum/wiki/wiki/White-Paper

[3] K. Croman, C. Decker, and I. Eyal et al., On Scaling Decentralized Blockchains (A Position Paper), 3rd Workshop on Bitcoin and Blockchain Research, February, 2016.

[4] A.E. Gencer, S. Basu, and I. Eyal et al., Decentralization in Bitcoin and Ethereum Networks, Financial Cryptography and Data Security (FC), 2018.

[5] A. Gervais, G. O. Karame, and S. Capkun et al., Is Bitcoin a decentralized currency, IEEE Security & Privacy, November, 2014.

[6] D. Ron and A. Shamir, Quantitative Analysis of the Full Bitcoin Transaction Graph, International Conference on Financial Cryptography and Data Security, FC 2013, LNCS 7859, pp. 6-24, 2013.

[7] B.S. Srinivasan and L. Lee, Quantifying Decentralization, https://news.earn.com/quantifying-decentralization-e39db233c28e

[8] Thomas M. Cover, Joy A. Thomas, Elements of Information Theory (Second Edition), A John Wiley & SONS, INC., Publicaiton, 2006.

[9] Ji G. Zhang, Vijay P. Singh, Information Entropy, China Water Publication, 2012.

[10] https://btc.com/stats/pool

[11] https://etherscan.io/stat/miner?range=7&blocktype=blocks

[12] https://btc.com/stats/rich-list

[13] https://etherscan.io/accounts