Uncovering the Bitcoin Blockchain: An Analysis of the Full Users Graph

Conference Paper · October 2016

DOI: 10.1109/DSAA.2016.52

CITATIONS
34

READS
1,084

3 authors, including:

Damiano Di Francesco Maesa
University of Cambridge
15 PUBLICATIONS **275** CITATIONS

SEE PROFILE

Laura Ricci
Università di Pisa
50 PUBLICATIONS **407** CITATIONS

SEE PROFILE

Uncovering the Bitcoin blockchain: an analysis of the full users graph

Damiano Di Francesco Maesa
Department of Computer Science
University of Pisa, Italy
Email: damiano.difrancescomaesa@for.unipi.it

Andrea Marino
Department of Computer Science
University of Pisa, Italy
Email: marino@di.unipi.it

Laura Ricci
Department of Computer Science
University of Pisa, Italy
Email: laura.ricci@unipi.it

***Abstract*—BITCOIN is a novel decentralized cryptocurrency system which has recently received a great attention from a wider audience. An interesting and unique feature of this system is that the complete list of all the transactions occurred from its inception is publicly available. This enables the investigation of funds movements to uncover interesting properties of the BITCOIN economy. In this paper we present a set of analyses of the user graph, i.e. the graph obtained by an heuristic clustering of the graph of BITCOIN transactions. Our analyses consider an up-to-date BITCOIN blockchain, as in December 2015, after the exponential explosion of the number of transactions occurred in the last two years. The set of analyses we defined includes, among others, the analysis of the time evolution of BITCOIN network, the verification of the "rich get richer" conjecture and the detection of the nodes which are critical for the network connectivity.**

I. INTRODUCTION

Internet changed our society giving rise to the so called digital revolution. Almost every individual can now directly connect to any other one on the globe allowing for a point-to-point information exchange. Even if several type of commercial interactions are possible on the network, a direct point-to-point value exchange (payment) is not yet possible, for most users, and a third party financial intermediary is required. To overcome this limitation, some novel proposals of digital currencies have been recently presented.

BITCOIN [1], the first true digital currency, came out the 3rd of January 2009, gaining wide mass media coverage and widespread popularity among the broad public of non specialists, so resulting in the first example of cryptocurrency economy worthy of analysis. After more than six years since the inception of BITCOIN, an economic community has risen around this cryptocurrency. Even if it still represents a niche and peculiar economical community, its importance in the real world has grown enough so that it no longer represents an experimental currency exploited only by computer science specialists. Several events of the BITCOIN economy, like the wild speculation, the value fluctuation and a major exchange failure witness that a true economic system has born around it.

A distinctive characteristic of BITCOIN is that it puts together economic and technological aspects. This produces interesting interrelations between the BITCOIN real world economy and the technological aspects of the distributed protocol. These interrelations have influenced its development as much as its original design principles. Several interesting aspects of the protocol are currently under investigation. These range from the improvement of the basic protocol with the definition of new anonymity mechanisms and consensus algorithms, to the analysis of the information mined from the blockchain. This paper focuses on the second aspect, i.e. the definition and analysis of the transaction graph built from the blockchain. BITCOIN keeps the entire transactions history public by design, so it represents one of the few economic communities that can be studied and analyzed in depth. Several recent works [2], [3], [4], [5], [6], [7] present analyses of the BITCOIN transaction graph whose goal is to find out some interesting property of its economy. Almost all of these works consider the state of the blockchain until the end of 2013. As shown in [8], the BITCOIN economy has experienced a huge explosion in the last few years, for instance the number of transactions has raised from roughly 10 millions in January 2013 to more than 100 millions of transactions in January 2016. Since the size of the blockchain fits together the number of transactions, its size exponentially increased as well. We believe that the exponential increase of the available data may enable a deeper understanding of the BITCOIN network and may also highlight novel characteristics of the network arisen in the last few years. On the other hand, the definition of classical and more complex analyses calls for scalable algorithms and tools able to support the huge size of the data.

This paper presents the framework we have developed to support the analyses of this huge amount of data, describes the analyses we have defined and discusses the results we have obtained. The main contributions of the paper may be summarized as follows:

- the definition of a scalable clustering algorithm to support the addresses graph reduction heuristics
- the study of the time evolution of the BITCOIN network which highlights several properties: a densification of the network (the giant component and average degree increase), and its characterization as a small-world and scale-free network. We also observed a high diameter compared to other complex networks.
- the detection of the most central nodes in the network: this highlights the most connected nodes and the nodes which are critical for the connectivity of the network.

- the verification of the *rich get richer* conjecture, both from the point of view of the balance of each node and from that of the connectivity point of view.
- the measurement of the evolution of the richness concentration.
- the analysis of the subgraphs of the BITCOIN network obtained by filtering the transactions with an amount larger than a threshold, for increasing values of the threshold.

The paper is organized as follows: Section II presents the related works, while Section III gives a brief overview of the BITCOIN protocol and presents the clustering algorithm. Section IV presents the results of the analyses. Finally, Section V discusses the conclusions and presents the future work.

II. Related Works

Several analyses of the BITCOIN network have been recently proposed. Most of them take in input the "user graph" that is extracted from the transactions graph through a well established heuristic rule. This rule, already introduced in the seminal paper [1], and extensively described in [9], establishes that all the input addresses of a multi-input transaction belong to the same user. The rule is based on the observation that every input of a multi-input transaction must be signed with the right private key and this implies that the signer knows all the private keys of the transaction and so it is the owner of all the input addresses. The resulting graph approximates the real users graph, because the heuristics may underestimate or overestimate the common ownership of some addresses. While underestimation occurs because addresses of the same owner have not been used in the same transaction, overestimation may occur because a set of users may collectively sign the same transaction [10]. The heuristic rule has been subsequently used in most analysis, like in [2], [3], [4], [5], [6], [7]. An exception are [3], [5], that also introduce a more sophisticated heuristic based on change addresses, i.e. the mechanism used to give money back to the input user in a transaction.

Let us now briefly review the most important analyses recently proposed. [2] considers only BITCOIN transactions carried out until May 2012. They discovered that the network contains a huge number of small transactions, but also a subset of transactions moving a large amount of money. The analyses are then focused on the large transactions in order to detect the ways amounts are accumulated and dispersed.

[6] does not apply any heuristic and directly analyses the transaction graph, extracted from the blockchain, whose state is considered as in May 2013. The authors identify an initial phase of growth of the BITCOIN network, characterized by a large fluctuation in the network characteristics and a trading phase characterized by more stable network measures. They find out that preferential attachment drives the growth of the network.

The main focus of the analyses presented in [3] is to highlight the gap between the potential and the actual anonymity of the BITCOIN protocol. The authors apply to the blockchain, as in April 2013, the two aforementioned heuristics to contract the transaction graph.

As most previous works, [7] considers the blockchain state at April 2013 and exploits only the first heuristic previously described for the contraction of the transaction graph. The authors categorize the transactions according to business categories by extracting the business tags of each address. Furthermore, they present an analysis of the geographic distribution of BITCOIN transactions.

III. Data: the BITCOIN network

A. *The* BITCOIN *protocol: preliminaries*

Users take part in the BITCOIN economy through addresses. An address is a double hash (firstly SHA-256 is applied and then Ripemd-160) of a public key derived form a ECDSA key pair. The address (and hence the public key) will be used by the user to send and receive payments, while the private key will be used by the user to provide proofs of ownership. Creating new ECDSA pairs (and so addresses) is not expensive at all and so each user can create and use multiple addresses. This leads to the use of pseudonyms, which means that each address is an user alias without any kind of information linking it to the user or to other addresses created by the same user. Pseudonymity is the only (weak) anonymity protection in BITCOIN. So to improve transactions privacy is recommended to create a new address to receive each new payment. While this is not computationally expensive, it can lead to an address management problem if the number of addresses keeps increasing.

To exchange funds between addresses, transactions are created. Transactions are multi input, multi output, it means that a transaction may have more than one input (address from which funds are withdrawn) and more than one output (address where funds are stored), and each transaction completely transfers funds from the inputs to the outputs (no change is left in the input addresses). Transactions are the only mean to manage funds, so funds can be divided or aggregated only by being spent. That is possible because a transaction involves addresses and not users and every user can have different addresses, so the user can use a transaction to split, merge or move funds between its own addresses. A transaction can also specify a voluntary fee to cover the expenses of the validation process (that we will explain briefly later). If the sum of input values exceeds the sum of output values then the exceeding value is considered a voluntary fee paid to the validator. In a transaction each output can be seen as a couple (amount, receiver address). Each input specifies, instead, where to withdraw the funds, so it does indicate an address only on abstraction, but in fact it indicates the previous transaction (through its hash) where the funds were created. Funds are represented by a transaction chain showing the passage of value (split and merge) between addresses, validated at each step by the previous owner signature. In BITCOIN transactions alone specify the entire state of the system. There is no coin exchanged between users, the coins are implicitly represented by the flow of value through transactions. New transactions

are created by any user and notified to the community with a gossip style broadcast message on the P2P BITCOIN network. We also note that a special kind of transaction called *coinbase* exists to allow for new value creation (distributed minting of new coins as part of the validation process). These special transactions have no inputs, only output addresses to whom newly minted value is credited.

In a transaction each input is signed by the owner with the private key corresponding to the address spending the funds. This digital signature guaranties that only the rightful owner can spend its funds, but it does not prevent it from spending them more than once in different transactions. This is the so called double spending problem. BITCOIN solution is to remember the history of all the past transactions to determine the actual owner of a fund, at each given time. The history is maintained in a distributed database called blockchain because transactions are grouped in blocks linked in a chain and the linking between blocks is achieved by saving the hash of the header of the previous block in the next block header. To make each block header (and so its hash) dependent from all transactions contained in that block, the root of the (implicit) Merkle tree [11], built from the block transactions hashes is included in the header. It is necessary to reach a distributed consensus to choose which block (and so which transactions) to add to the chain, because there could be incompatible transactions caused by a double spending attempt. The distributed consensus protocol introduced and used by BITCOIN is called Nakamoto consensus and relies on HashCash Proof-of-Works [12]. This Nakamoto consensus protocol is one of the most interesting aspects of BITCOIN but we will not discuss it further since it's beyond the scope of this paper.

B. Building the Graph

The BITCOIN dataset can be formally modelled by a weighted directed hypergraph $H = (A, T)$ where: A is the set of all addresses; T is the set of transactions, which can be modeled as a set of ordered pairs (A_1, A_2) with $A_1, A_2 \subseteq A$, meaning that the addresses in A_1 are paying the addresses in A_2 (see for instance [6]).

Moreover, to each transaction $s = (A_1, A_2) \in T$, we associate:

- a timestamp telling when the transaction took place.
- a distribution of amounts among the nodes in A_2 denoted as b_s. More formally, b_s is a function associating to each $a \in A_2$ a multiset of values in $\mathbb{R}$. Indeed, notice that there can be transactions associating to the same $a \in A_2$ more than one single amount.
- a fee ϕ_s (eventually 0) that associates to A_1 the voluntary taxes payed.

As seen in the previous section III-A, in BITCOIN each user controls different pseudonymous addresses. In order to infer the users of the network, we want to cluster all the addresses managed by the same user so that each cluster will ideally correspond to a single user. In particular, we group addresses according to the following desired property.

Algorithm 1: THE GRAPH BUILDING PROCESS

Input : A weighted directed hypergraph $H = (A, T)$, b_s for each $s \in T$
Output: A directed multigraph $G = (V, E, w)$

Procedure CLUSTER(H)
  $G'_H = (A, E') \leftarrow$ undirected graph with A as set of vertices and E' empty set of edges
  foreach $s = (A_1, A_2) \in T$ **do**
    Let $A_1 = \{a_1, a_2, \ldots, a_h\}$
    for $i \in \{1, \ldots, h-1\}$ **do** add $\{a_i, a_{i+1}\}$ to E'
  Let $C_1, \ldots, C_k$ be the connected components of G'_H
  return $C_1, \ldots, C_k$

$C_1, \ldots, C_k \leftarrow$ CLUSTER(H)
Let $c(a)$ be the vector associating to each $a \in A$ the cluster C_j such that $a \in C_j$
$\phi(C_i) \leftarrow 0$ for each C_i.
$G = (V, E, w) \leftarrow$ graph where $V = \{C_1, \ldots, C_k\}$ is the set of nodes, E is an empty set of arcs, $w : E \to \mathbb{R}$
foreach $s = (A_1, A_2) \in T$ **do**
  Let C_i be the unique cluster $c(a_1)$ for any $a_1 \in A_1$
  $\phi(C_i) \leftarrow \phi(C_i) + \phi_s$
  foreach $a_2 \in A_2$ **do**
    Let C_j be $c(a_2)$
    foreach $x \in b_s(a_2)$ **do**
      Add an arc e from C_i to C_j in E with weight $w(e)$ equal to x

Property 1. *For every two addresses x and y, if there exists a transaction (A_1, A_2) where $x, y \in A_1$, then x and y belong to a same cluster.*

Note that this is a sufficient but not necessary condition for x and y to be in the same cluster, since we merge clusters transitively: if for instance x and z belong to A_1 for some transaction $(A_1, A_2) \in T$ and z, y belong to A_3 for some transaction $(A_3, A_4) \in T$, then x, z, y will be assigned to the same cluster.

The clustering algorithm works as shown in procedure CLUSTER in Algorithm 1. Given the hypergraph $H = (A, T)$ above, the clustering algorithm produces a partition of A, i.e. the clusters $C_1, \ldots, C_k \subseteq A$ for some k, with $C_i \cap C_j = \emptyset$ $(1 \leq i, j \leq k\ ,\ i \neq j)$ and $C_1 \cup \ldots \cup C_k = A$. We define the undirected graph G_H whose nodes are all the addresses in A and two nodes $x, y \in A$ are linked whether there exists a transaction $(A_1, A_2) \in T$ such that $x, y \in A_1$. Then the k connected components of G_A are our clusters $C_1, \ldots, C_k$.

The following result holds.

Lemma 1. *Given $H = (A, T)$, the clustering corresponding to the connected components $C_1, \ldots, C_k$ of G_H satisfies Property 1.*

It is worth observing that building G_H as described above

can be costly: for each transaction $(A_1, A_2) \in T$, we have to create a clique among all the nodes in A_1, adding a quadratic number of edges, i.e. $|A_1| \cdot (|A_1| - 1)/2$. Instead of creating a clique, procedure CLUSTER in Algorithm 1 adds a simple path between the addresses in A_1, adding each time a linear number of edges. In other words, CLUSTER creates a graph G'_H whose set of nodes is A and whose set of edges is given by the following process: for each transaction $(A_1, A_2) \in T$, create a path among the nodes in A_1. The following property trivially holds.

Lemma 2. *G'_H and G_H have the same connected components.*

Once the clusters $V = \{C_1, \ldots, C_k\}$ have been identified in $H = (A, T)$ using G'_H, we create the weighted multigraph G whose set of nodes is V and there is an arc e from $C_i \in V$ to $C_j \in V$ whether there exists a transaction $(A_1, A_2) \in T$ such that $A_1 \cap C_i \neq \emptyset$ and $A_2 \cap C_j \neq \emptyset$. Roughly speaking, there is an arc from a cluster to another whether there exists a transaction from an address of the former to an address of the latter. Note that this is a multigraph since there can be several transactions from a cluster to another, possibly with different (or equal) amount. Moreover, for each value $x \in b_s(a_2)$ with $a_2 \in A_2$, we create an arc with weight x, since, as explained before, a same transaction s can assign more than one amount to a vertex $a_2 \in A_2$. Finally, we define ϕ for each C_i as the sum of ϕ_s for each transaction s payed by C_i.

We will refer to G as BITCOIN users graph. The building method is summarized by Algorithm 1. It is worth noting that the whole process is linear in the size of H, i.e. $O(|A| + \sum_{(A_1,A_2)\in T}(|A_1| + |A_2|))$.

C. Clustering Statistics

For the sake of completeness, in Figure 1 we show the distribution of clusters size. It is worth observing, that this distribution follows a power law.

We report in Table I some basic statistics of our dataset, like the total number of addresses and the total number of transactions, and some statistics about the result of our clustering process.

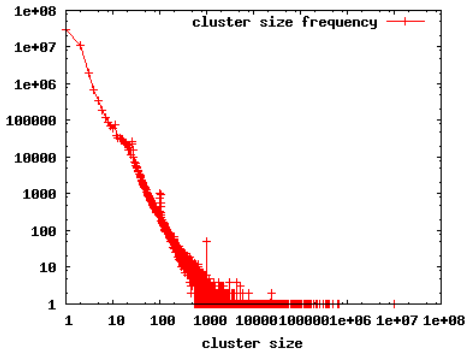


Fig. 1. Distribution of Cluster Sizes.

In the lower part of Table I we report the list of the top ten biggest clusters we obtained. We observe that the size of these clusters is several order of magnitude bigger than the average size of the clusters, making the distribution of the clustering size heavy tailed (see Figure 1). We will see that several clusters in this top ten list are in the top ten list for other topological centrality measures.

NUMBER OF ADDRESSES, I.E. $\|A\|$	113 221 083
NUMBER OF TRANSACTIONS, I.E. $\|T\|$	99 602 440
NUMBER OF CLUSTERS, I.E. NODES OF G	46 144 246
NUMBER OF ARCS OF G	294 705 549

THE 10 BIGGEST CLUSTERS		
CLUSTER ID	IDENTITY	SIZE
66 482	Mt. Gox	10 216 380
2 899 325	LocalBitcoins.com	676 402
26 784 111	GoCoin.com	611 885
11 032 019	AgoraMarket	497 995
12 388 597	EvolutionMarket	420 632
2 477 299	N/A	392 589
2 547 597	SilkRoadMarketplace	372 753
10 072 646	SilkRoad2Market	349 874
1 175 285	BTC-e.com1	348 438
11 828 673	999Dice.com	301 990

TABLE I
SOME CLUSTERING STATISTICS.

D. Data acquisition

As we explained in the previous section III-A all the BITCOIN transaction history is publicly available in the blockchain as a countermeasure against double spending. To obtain the blockchain is sufficient to set up a BITCOIN node in the P2P network and start requesting blocks to the other nodes. This process can take a lot of time so, to avoid it, we used a blockchain already downloaded and stored in Protocol Buffers format [13]. The blockchain used contains all the first 389 800 blocks, from the Genesis block until block height 389 799, hence containing all the BITCOIN transactions from 2009-01-03 18:15:05 GMT to 2015-12-23 09:40:52 GMT .

In section III-A we have given an high level description of BITCOIN transactions, but in practice the transactions stored in the blockchain contain scripts. The BITCOIN protocol uses a not Turing complete stack based scripting language, and scripts are (mostly) used in a transaction to specify conditions needed to redeem the funds of that transaction. The most common example of such condition is a signature. When a transaction is tested for validity, the input scripts are concatenated with the output scripts, evaluated, and all transaction scripts must evaluate to true for the transaction to be validated. Scripts can potentially be arbitrarily complex but in practice only few types of standardized scripts are used in transactions, those scripts are called *standard*. What's more important is that non-*standard* scripts are accepted but not relayed by compliant nodes, so they have less chances of actually ending up in the blockchain. The most used *standard* script types are called Pay to PubKey Hash (`p2pkh`), Pay to PubKey (`p2pk`), Pay to Script Hash (`p2sh`) and Pay to Multisig (`p2ms`). We have decided to parse the blockchain only interpreting the `p2pkh`, `p2pk` and `p2sh` types of scripts, dropping the transaction outputs containing other kinds of scripts. We did this to not

over-complicate our blockchain parser and because we thought that this kind of scripts where the ones used in transactions more suitable to apply our clustering heuristic, hence hoping to reduce the number of false positives returned by the heuristic clustering. In the end we successfully interpreted 295 144 677 scripts and failed to interpret 1 489 903 scripts, resulting in a coverage of 99.4977% of all transaction outputs. So we deem the information loss acceptable.

From the parsing of the raw blockchain with the script interpretation described before, we obtain our transactions dataset and on this dataset we apply our clustering algorithm and perform the analysis. We do not try to infer ourselves addresses identities and instead we rely on the public address tags datasets provided by [14], [15]. In the rest of this paper to suppose a cluster identity we look for identity tags (provided by those services) associated to the addresses belonging to that cluster. It's beyond the scope of this paper to evaluate the correctness of those tags.

IV. Analysis and Results

In this section we study the topological properties of the BITCOIN users graph G built in Section III-B. Recall that G is a weighted directed multigraph. We refer to U as the symmetric version of G, i.e. the graph where all the arcs become undirected.

In Section IV-A, we study the time evolution of G and U. For increasing values of time t we have considered just transactions that took place before t. We indicate with G^t (and U^t), with $1 \leq t \leq 20$, the graph induced by transactions whose time stamp is smaller than t, where t refers to the left part of Table II. Analogously, we indicate with $\phi^t(u)$ the fees payed by u until time t, i.e. $\phi(u)$ induced by transactions with timestamp smaller than t. The timestamps chosen are at constant intervals in time but with an high initial offset. We chose to start the timestamp snapshots from the beginning of 2013 because we considered it the time when the BITCOIN economy started to rise significantly and was mature enough for a systematical analysis. Moreover, for each graph snapshot at each timestamp considered during the connectivity analysis phase in Section IV-A we (non recursively) pruned the graph from the nodes with only one incoming arc. We choose to do so because otherwise the graph was biased from more recent nodes artificially isolated by the timestamp cutoff: indeed, we noticed that most of those nodes corresponded to nodes that had just received a payment and didn't have enough time to use that value in a subsequent transaction. This pruning does not skew our analysis since the nodes pruned were nodes reached by only one incoming arc.

In Section IV-B we report some centrality analysis based on the connectivity of the last snapshot of the network, considering harmonic centrality [16] and the degrees of the vertices.

In Section IV-C, we define the *richness* of a node according to its number of incoming transactions and its balance. We study how the sets of richest nodes change over time, proving that richness tends to concentrate.

TIME t	SNAPSHOT
1	Tue Jan 01 00:00:00 GMT 2013
2	Sun Feb 24 07:41:02 GMT 2013
3	Fri Apr 19 15:22:04 GMT 2013
4	Wed Jun 12 23:03:06 GMT 2013
5	Tue Aug 06 06:44:08 GMT 2013
6	Sun Sep 29 14:25:10 GMT 2013
7	Fri Nov 22 22:06:12 GMT 2013
8	Thu Jan 16 05:47:14 GMT 2014
9	Tue Mar 11 13:28:16 GMT 2014
10	Sun May 04 21:09:18 GMT 2014
11	Sat Jun 28 04:50:20 GMT 2014
12	Thu Aug 21 12:31:22 GMT 2014
13	Tue Oct 14 20:12:24 GMT 2014
14	Mon Dec 08 03:53:26 GMT 2014
15	Sat Jan 31 11:34:28 GMT 2015
16	Thu Mar 26 19:15:30 GMT 2015
17	Wed May 20 02:56:32 GMT 2015
18	Mon Jul 13 10:37:34 GMT 2015
19	Sat Sep 05 18:18:36 GMT 2015
20	Wed Dec 23 9:40:52 GMT 2015

AMOUNT a	THRESHOLD
1	0.000 001 BTC
2	0.000 01 BTC
3	0.000 1 BTC
4	0.001 BTC
5	0.01 BTC
6	0.1 BTC
7	1 BTC
8	10 BTC
9	100 BTC
10	1 000 BTC
11	10 000 BTC
12	100 000 BTC

TABLE II
THE TIME SERIES WE CONSIDERED (UPPER PART) FOR G^t AND U^t AND THE DIFFERENT AMOUNTS FOR G_a AND U_a (LOWER PART).

In the remaining part of the section, i.e. Section IV-D, we study the graph G for different transaction amounts. In particular, we call G_a (and U_a) the graph induced by transactions whose amount is smaller than a, with $1 \leq a \leq 12$, where the corresponding values of a are listed in the right part of Table II.

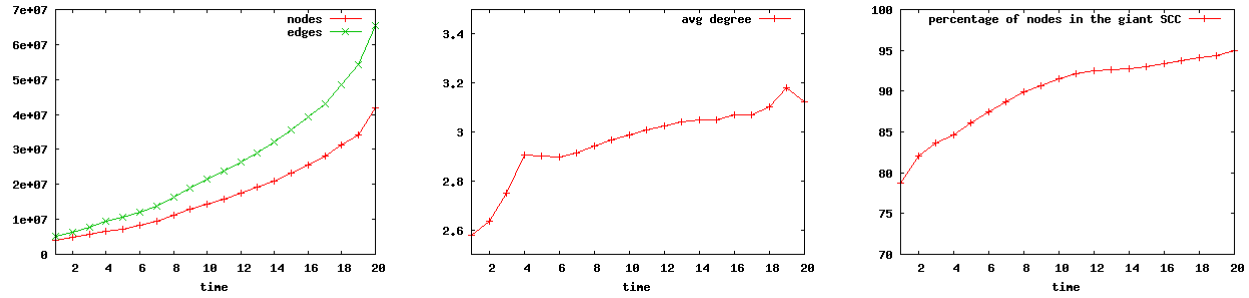
Even though the nodes of G and U correspond to clusters of addresses as seen in Section III-B, for the sake of simplicity, we will simply refer to them as vertices or nodes. Moreover, we will call the links in G as arcs, which are directed, and the ones of U as edges, which are instead undirected.

A. Connectivity Analysis Over Time

Since the interest of this section relies on the connectivity of the network, we consider G^t and U^t as simple graphs ignoring multiple arcs (or edges). Our analysis framework have been implemented using WEBGRAPH [17].

1) Densification: This section aims to observe the densification process taking place in BITCOIN. This phenomenon is described by Figure 2.

- Figure 2(a) shows the increase of number of nodes and arcs over time in G^t. Since the time slots are equally spaced, the plot highlights that the increase of nodes and arcs is more than linear.



(a) Number of nodes and arcs over time.

(b) Average out-degree over time.

(c) Percentage of nodes in the giant SCC over time.

Fig. 2. Densification of G^t

- Figure 2(b) shows the increase of the average out-degree of the nodes in G^t for increasing values of t (note that the average out-degree and the average in-degree are the same). This plot highlights that the number of arcs in G^t increases faster than the number of nodes.
- Figure 2(c) shows the behaviour of the percentage of nodes in the giant strongly connected component of G^t with respect to the total number of vertices in G^t, i.e. $|V^t|$. This value quickly increases, meaning that, even though $|V^t|$ grows quite fast (Figure 2(a)), the number of nodes in the giant strongly connected component grows much faster making the network much more robust.

2) Small-World: In order to test the small-word phenomenon [18], [19], we have computed diameter and average distance of the BITCOIN network. The distance $d(u, v)$ from a node u to a node v in a graph is the length of the shortest path from u to v. The diameter is defined as the $\max_{(u,v)\in V^t\times V^t} d(u, v)$, while the average distance is simply $\frac{1}{|V^t|^2}\sum_{(u,v)\in V^t\times V^t} d(u, v)$. In order to get more robust analysis, we have done these measurements considering the giant connected component of U^t for increasing values of t, where two users are connected whether they exchanged BITCOINs, ignoring the direction of the link. The average distance has been approximated using [20], while the diameter has been computed exactly using [21]. Note that, for a graph of n nodes and m edges, computing the average distance and the diameter requires $O(n\cdot m)$. The algorithm in [20] allows to approximate the average distance in $O(m)$ and the algorithm in [21] allows to compute exactly the diameter in $O(m)$ in practice in real world graphs.

As observed for many other real world networks [22], we have seen that the diameter is not increasing. Surprisingly, the diameter is constant and very long (i.e. 2050) if compared to the diameter of many other real-world networks, like Facebook [23], where the number of vertices is much higher and the diameter is 41, and others (see lasagne-unifi.sourceforge.net). Our preliminary observations suggest that this peculiarity of the BITCOIN users graph is caused by the fact that transactions are also used to merge and split user funds and not just for payments, as explained in section III-A. This is consistent with rare user cases observed in the past, obfuscating funds ownership using long fund splits chains (as noted for example in [2]). We plan to investigate extensively this topic in our future works.

Figure 3 shows the slow decrease over time of the low average distance, confirming, as expected, the small-world hypothesis. Note that this small average value compared to the high value of the diameter highlights that the nodes connected by long paths are present but few.

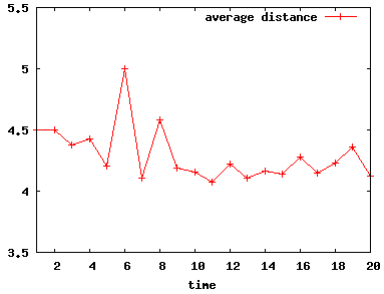


Fig. 3. Small-world: average distance of U^t for increasing values of t.

3) Degree Distribution: In Figure 4(a) and Figure 4(b) we show respectively the in-degree and the out-degree distributions of G^t with $t = 20$. As a further remark, we have seen that also the degree distribution of U^t, which for brevity is not shown here, follows a similar behaviour. It can be noticed that in both the plots (a) and (b) there are some outliers: there are some spikes close to $x = 1000$ in Figure 4(a), and to $x = 100$ in Figure 4(b). These spikes will be object of future investigations.

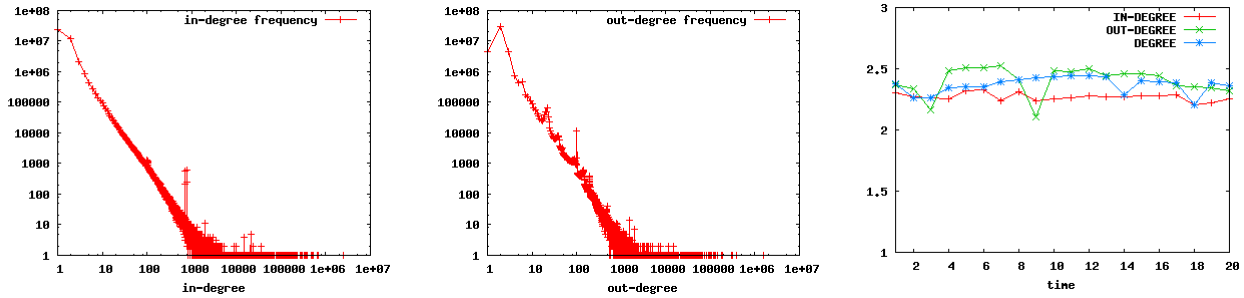
All the distributions above follow a power law. In Figure 4(c) we show the power law exponent for increasing values of t for the in-degree distribution of G^t (red line), the out-degree distribution of G^t (green line), and the degree distribution of U^t (blue line). The power law exponent seems to be constant over time confirming the estimations done in [6].

B. Centrality Analysis

In this section we report the most central vertices in the BITCOIN network. These vertices correspond to the most active vertices in G^t or the ones that play a crucial role for the connectivity of the network (see [16], for more details about

	HARMONIC		DEGREE		IN-DEGREE		OUT-DEGREE	
	IDENTITY	VALUE	IDENTITY	VALUE	IDENTITY	VALUE	IDENTITY	VALUE
1	Mt. Gox	11798171	Mt. Gox	3 386 581	Mt. Gox	2 452 049	Mt. Gox	1 591 319
2	2477299	10447302	LocalBitcoins.com	902 151	BTC-e.com1	683 875	2477299	381 426
3	LocalBitcoins.com	10320862	2477299	848 176	LocalBitcoins.com	650 269	FaucetBOX.com	317 742
4	Cex.io	10144968	BTC-e.com1	740 402	AgoraMarket	636 969	LocalBitcoins.com	301 692
5	FaucetBOX.com	10136604	AgoraMarket	722 331	SilkRoadMarketplace	527 718	14782788	191 867
6	26638073	10071881	SilkRoadMarketplace	577 124	2477299	511 239	MoonBit.co.in	180 161
7	MoonBit.co.in	10065853	BitPay.com1	500 990	BitPay.com1	493 067	3454364	178 349
8	19860816	10025701	BTC-e.com2	492 219	BTC-e.com2	479 452	26638073	176 508
9	Poloniex.com	9976766	Cryptsy.com	461 111	BitPay.com2	394 447	Cryptsy.com	148 015
10	Bittrex.com	9926321	BitPay.com2	401 254	Cryptsy.com	361 298	23144512	146 624

TABLE III
THE TOP 10 CENTRAL NODES ACCORDING TO HARMONIC CENTRALITY IN U^t, DEGREE IN U^t, IN-DEGREE AND OUT-DEGREE IN G^t FOR THE LAST SNAPSHOT, I.E. $t = 20$. FOR THE UNKNOWN IDENTITY WE REPORT THE NUMBER OF THE IDENTIFIER IN OUR DATASET.



(a) In-degree distribution of G^t with $t = 20$.

(b) Out-degree distribution of G^t with $t = 20$

(c) Power law exponent as the time goes by.

Fig. 4. Behaviour of Degree Distributions

centrality measures). Given $G^t = (V^t, E^t, w^t)$, the centrality of u can be defined as follows.

- HARMONIC: That is $\sum_{v \in V^t} \frac{1}{d(u,v)}$, where $d(u, v)$ is the distance between u and v in G^t. A large value means high centrality [16].
- DEGREE: That is the degree of u in U^t, i.e. the number of nodes paying or payed by u. Central nodes are supposed to have many connections.
- IN-DEGREE: That is the in-degree of u in G^t, i.e. the number of nodes that payed u.
- OUT-DEGREE: That is the out-degree of u in G^t, that corresponds to the number of nodes which have been payed by u.

In Table III, we report the top-k users according to the above measures, with $k = 10$. The corresponding computations have been done discarding disconnected parts from the graph, and without considering multiple arcs or edges. The HARMONIC centrality has been computed using [24] and has been shown to be an effective distance-based centrality measure [16].

As expected, the selected central vertices are almost all very popular. Mt. Gox (the most famous BITCOIN exchange before its failure at the beginning of 2014) is the most central according to all the measures not only considering its local connectivity, i.e. the degrees, but also for the connectivity of the whole network. The same applies to LocalBitcoins.com (the largest P2P BITCOIN trading platform). Indeed, the central nodes in the HARMONIC column have often high in- or out-degree, i.e. they are central in their big local community. The HARMONIC rank highlights which nodes are closer to all the others from a global point of view.

Interestingly, it seems difficult to identify some users, as for instance the one corresponding to vertex 2477299, that seems to be very central.

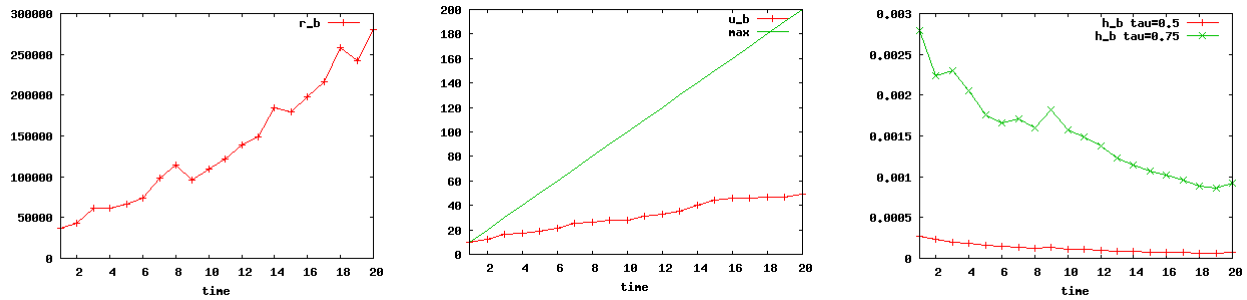
C. Rich get Richer and Concentration of Richness

This section is devoted to verify the *rich get richer* hypothesis and measure the concentration of richness, both on the balance and the connectivity point of view. Indeed, we consider two different definitions of richness: we say that a user is rich whether its balance or its number of incoming transactions is high with respect to the other users in the network. We aim to verify the following properties for both the definitions.

Property 2.

1) *The richest users at time t are richer than the richest users at time $t' < t$.*
2) *The richest users at a certain time t tend to remain the richest at time $t' > t$.*
3) *The richness gets more concentrated with the progression of time.*

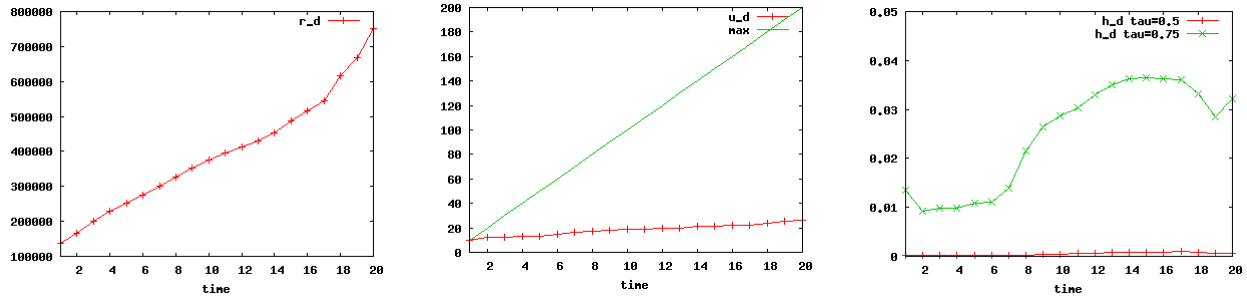
Given the weighted multigraph $G^t = (V^t, E^t, w^t)$ and ϕ^t for each $u \in V^t$, we formally define the richness of a node $u \in V^t$ as its balance $b^t(u)$ or its number of incoming transactions $d_t(u)$ as follows.



(a) r_b^t over time

(b) u_b^t over time and the maximum u_b^t possible

(c) h_b^t over time, fixing $\tau = 0.5$ and $\tau = 0.75$.



(d) r_d^t over time

(e) u_d^t over time and the maximum u_d^t possible

(f) h_d^t over time, fixing $\tau = 0.5$ and $\tau = 0.75$.

Fig. 5. Verifying Property 2.1, 2.2, and 2.3 for the definitions of richness given by Equation 1 (upper part, respectively (a), (b), and (c)) and Equation 2 (lower part, respectively (d), (e), and (f)).

	IN-TRANS		BALANCE	
RANK	IDENTITY	VALUE	IDENTITY	VALUE
1	Mt. Gox	22 399 043	39912924	169 731
2	SatoshiDice.com	12 879 343	23638585	157 997
3	LuckyB.it	3 620 428	542746	87 111
4	BitZillions.com	1 651 456	Mt. Gox	81 492
5	BTC-e.com1	1 430 992	177808	79 957
6	LocalBitcoins.com	1 356 029	6128144	69 370
7	Xapo.com	1 264 648	10597666	66 650
8	AgoraMarket	1 184 011	10467915	66 612
9	BetcoinDice.tm	1 159 024	10484095	66 583
10	2477299	1 101 024	10475912	66 452

TABLE IV
THE TOP 10 RICHEST NODES IN G^t WITH $t = 20$.

$$b^t(u) = \sum_{(v,u)\in E^t} w(v,u) - \sum_{(u,v)\in E^t} w(u,v) - \phi^t(u) + \beta^t(u) \tag{1}$$

$$d^t(u) = |\{(v,u) : (v,u) \in E^t\}| \tag{2}$$

Observe that the measure $b^t(u)$ is taking into account also the fees payed by user u as $\phi^t(u)$. Moreover, $\beta^t(u)$ is the increase of the balance of u up to time t (eventually 0) coming from special transactions called *coinbase* whose aim is minting new coins (as explained in Section III-A).

Given an integer k, we denote respectively as B_k^t and D_k^t the k nodes having maximum b^t and d^t. Table IV shows the sets D_k^t (IN-TRANS columns) and B_k^t (BALANCE columns) for $t = 20$ and $k = 10$. The identity of these nodes is the name or the identifier of the node in our dataset in the case we do not know its name. As far as we know, many of the BALANCE column identifiers correspond to BITCOIN accumulator addresses (single addresses that received huge amounts of BTCs over time without spending them).

Note that, the measure $d^t(u)$ corresponds to the in-degree of u in the multigraph, i.e. the number of transaction outputs paying u. This is different from the in-degree considered in Section IV-B which corresponds to the degree in the simple graph, i.e. the number of users paying u.

A definition similar to $d^t(u)$ can be done considering the transactions outgoing from u, obtaining similar results. However, the number of transactions outgoing from a user can be arbitrarily increased by the user (performing transactions with small amounts) to increase its connectivity. We argue that our definition of $d^t(u)$ is more robust for modelling economic importance.

1) Verifying Property 2.1: To verify that the richest users in G^t are richer than the richest in $G^{t'}$ with $t' < t$, we study the following quantities over time.

$$r_b^t = \frac{\sum_{u\in B_k^t} b^t(u)/k}{\sum_{u\in V^t} b^t(u)/|V^t|}, \qquad r_d^t = \frac{\sum_{u\in D_k^t} d^t(u)/k}{\sum_{u\in V^t} d^t(u)/|V^t|}$$

Basically, r_b^t (resp. r_d^t) is the ratio between the average balance (resp. incoming transactions count) of the top-k richest users with respect to the average balance (resp. incoming

transactions count) of all the users in the network. As this ratio gets higher, the disparity of the richest nodes with respect to all the others gets bigger.

Figure 5(a) clearly shows that r_b^t increases over time, meaning that this disparity increases. On the other hand, Figure 5(d) shows that the same applies to r_d^t.

2) Verifying Property 2.2: In order to test the diversity of the richest node sets, i.e. B_k^t (resp. D_K^t), varying t, we study the following quantities.

$$u_b^t = |\bigcup_{i=1}^{t} B_k^i| \qquad u_d^t = |\bigcup_{i=1}^{t} D_k^i|$$

Since $|B_k^t| = k$ (resp. $|D_k^t| = k$), in the case the richest node sets does not change for each time i with $1 \leq i \leq t$, we have $u_b^t = k$ (resp. $u_d^t = k$). On the other hand, if the sets B_k^i (resp. D_k^i) change completely for each i then we have $u_b^t = t \cdot k$ (resp. $u_d^t = t \cdot k$).

Figure 5(b) and (e) show the behavior of u_b^t and u_d^t over time with respect to the expected behaviour if sets would change (green line). Both the sets B_k^t and D_k^t are very stable. Fixing $t = 20$ and $k = 10$, u_b^t, the set of all the k-richest nodes in the *history* of BITCOIN, is less than 50 instead of 200. This stability seems to be even more evident in the case of u_d^t, where $|\bigcup_{i=1}^{t} D_k^i|$ is smaller than 30 instead of 200.

3) Verifying Property 2.3: To measure the concentration of richness in $G^t = (V^t, E^t, w^t)$, fixing a threshold τ, we considered the following measures.

$$h_b^t = min\left\{k \,:\, \frac{\sum_{u \in B_k^t} b^t(u)}{\sum_{u \in V^t} b^t(u)} > \tau\right\} / |V^t|$$

$$h_d^t = min\left\{k \,:\, \frac{\sum_{u \in D_k^t} d^t(u)}{\sum_{u \in V^t} d^t(u)} > \tau\right\} / |V^t|$$

As an example, consider $\tau = 0.75$:

- h_b^t is the minimum (normalized) k such that B_k^t owns the 75% of the richness, in terms of balance, of the whole network;
- h_d^t is the minimum (normalized) k such that D_k^t owns the 75% of incoming connections of the network.

A small value for h_b^t (or h_d^t) means that the richness is concentrated in few users. The increase of concentration of richness can be witnessed checking whether h_b^t (and h_d^t) decreases over time.

In Figure 5(c) and (f), we report the behaviour of both h_b^t and h_d^t for increasing values of time t in our time series setting $\tau = 0.5$ and $\tau = 0.75$. Figure 5(c) refers to h_b^t and clearly decreases over time showing that the balance becomes more concentrated as the time passes; this is more evident especially considering an higher value of τ, i.e. $\tau = 0.75$.

On the other hand, the results for h_d^t, shown in Figure 5(f), do not satisfy Property 2.3. Indeed, contrarily to h_b^t, h_d^t seems to increase over time. The densification process shown in Section IV-A1 and this fact suggests that the increase of arcs is too big to be suitably absorbed from a same percentage of nodes.

D. Further Analysis for Different Transaction Amounts

In this section we show the growth of graph G_a for increasing values of a. Recall that G_a is the graph G built in Section III-B induced by transactions whose amount is smaller than a, where the correspondence between a and real BTCs is provided by Table II.

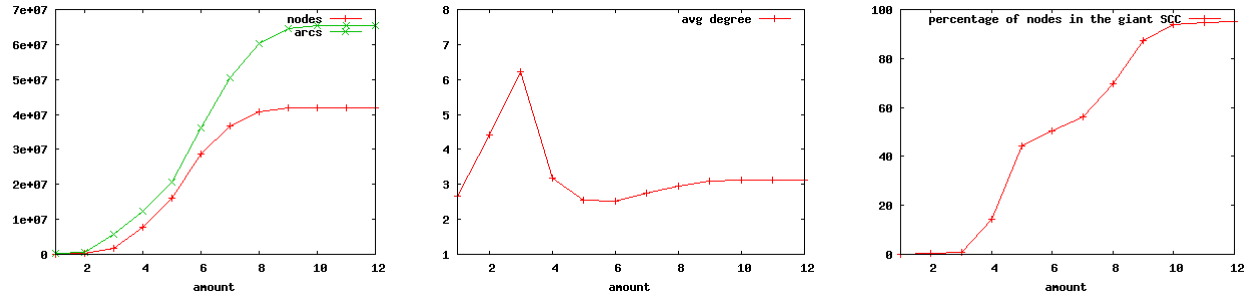
Figure 6(a) shows the increase of number of nodes and arcs in G_a for increasing a. We can see that the bigger increase of nodes is when transactions of amount between 3 and 7 are introduced (see the angles of red and green lines). From $a \geq 8$, the number of nodes and arcs is stable, meaning that not many transactions have an amount greater than 1 BTC. Figure 6(b) shows the average degree for increasing a. There is a spike for $a = 3$ meaning that the maximum relative increase of edges with respect to the nodes is when introducing transaction with amount between 0.00001 BTC and 0.0001 BTC. This is due to the fact that nodes doing this kind of transactions often perform even smaller transactions. Looking at Figure 6(c), we can see that these smaller transactions are not well connected meaning that these take place in independent parts of the network: some parts of the network are giving, some other parts are receiving, but rarely they are exchanging. This suggests that no well connected micro-economy is present, but all the transactions up to $a = 8$ are needed to make the great majority of the network strongly connected. Interestingly, we can see a similar shape between Figure 6(c) and the green line in Figure 6(a).

Figure 7 shows the average distance in U_a, the undirected version of G_a. The starting increase for $a < 4$ is due to the fact that we are merging independent parts of the network, creating new paths (this is consistent with Figure 6(c)). The decrease for $a > 4$ is due to the fact that bigger transactions are creating shortcuts for relationships already existing.

For the sake of completeness, we mention that we observed a stable power law exponent for increasing values of a for all the degree distributions, similarly to Figure 4. In particular, the exponent is constantly about 2.3 for the out-degree distribution in G_a and the degree distribution in U_a; it is 2.2 for the in-degree distribution in G_a.

V. Conclusions

This paper presents a set of analyses of the users graph of BITCOIN. The blockchain we have considered, is that of December 2015 and includes about 100 millions of multi input, multi output transactions. To support the construction of the user graph from such a huge amount of data, we have defined a scalable clustering algorithm. The analyses reveal a set of interesting properties of the BITCOIN network, like the "rich get richer" property and the existence of central nodes acting as privileged bridges between different parts of the network. We have also observed that the diameter of the BITCOIN network is much larger than the one of social networks and that the degree distributions have some spikes for some specific values. We plan to investigate the actual users behaviors leading to these peculiar properties in a future work. We also plan to extend our analysis to highlight the



(a) Number of nodes and arcs of G_a for increasing amounts a.

(b) Average degree in G_a for increasing amounts a.

(c) Percentage of nodes in the giant SCC of G_a for increasing amounts a.

Fig. 6. Densification of G_a for increasing amounts a (see also Table II)

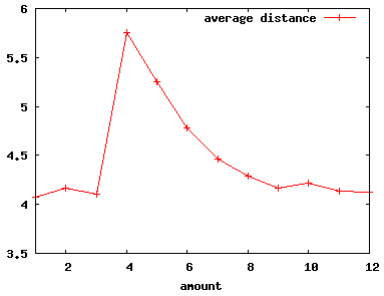


Fig. 7. Average distance of G_a for increasing values of a.

relation between the structure of the users graph and other interesting properties of the BITCOIN economy, for instance how speculators can be detected by an analysis of the users graph.

ACKNOWLEDGMENT

The authors would like to thank Dr. Christian Decker and Prof. Roger Wattenhofer of the Distributed Computing Group, ETH Zurich for providing us the blockchain in Protocol Buffers format.

This work was supported by PRA, Progetto di Ricerca di Ateneo, "Big Data, Social Mining and Risk Management", University of Pisa.

REFERENCES

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[2] D. Ron and A. Shamir, "Quantitative analysis of the full bitcoin transaction graph," in *Financial Cryptography and Data Security - 17th International Conference, FC 2013, Okinawa, Japan, April 1-5, 2013, Revised Selected Papers*, 2013, pp. 6–24.

[3] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage, "A fistful of bitcoins: characterizing payments among men with no names," in *Proceedings of the 2013 Internet Measurement Conference, IMC 2013, Barcelona, Spain, October 23-25, 2013*, 2013, pp. 127–140.

[4] M. Ober, S. Katzenbeisser, and K. Hamacher, "Structure and anonymity of the bitcoin transaction graph," *Future Internet*, vol. 5, no. 2, pp. 237–250, 2013.

[5] E. Androulaki, G. Karame, M. Roeschlin, T. Scherer, and S. Capkun, "Evaluating user privacy in bitcoin," in *Financial Cryptography and Data Security - 17th International Conference, FC 2013, Okinawa, Japan, April 1-5, 2013, Revised Selected Papers*, 2013, pp. 34–51.

[6] D. Kondor, M. Pósfai, I. Csabai, and G. Vattay, "Do the rich get richer? an empirical analysis of the bitcoin transaction network," *PloS one*, vol. 9, no. 2, p. e86197, 2014.

[7] M. Lischke and B. Fabian, "Analyzing the bitcoin network: The first four years," *Future Internet*, vol. 8, no. 1, 2016.

[8] "Block chain info charts." [Online]. Available: https://blockchain.info/charts/

[9] R. Fergal and M. Harrigan, "An analysis of anonymity in the bitcoin system," in *Proceeding of 2011 PASSAT/SocialCom 2011*. IEEE, 2011, pp. 1318–1326.

[10] T. Ruffing, P. Moreno-Sanchez, and A. Kate, "Coinshuffle: Practical decentralized coin mixing for bitcoin," in *Computer Security-ESORICS 2014*. Springer, 2014, pp. 345–364.

[11] R. C. Merkle, "A digital signature based on a conventional encryption function," in *Advances in Cryptology - CRYPTO '87, Santa Barbara, California, USA, August 16-20, 1987, Proceedings*, 1987, pp. 369–378.

[12] C. Dwork and M. Naor, "Pricing via processing or combatting junk mail," in *Advances in CryptologyCRYPTO92*. Springer, 1992, pp. 139–147.

[13] "Protocolbuffers." [Online]. Available: https://developers.google.com/protocol-buffers/

[14] "Block chain info tags." [Online]. Available: https://blockchain.info/tags

[15] "Wallet explorer." [Online]. Available: https://www.walletexplorer.com/

[16] P. Boldi and S. Vigna, "Axioms for centrality," *Internet Mathematics*, vol. 10, no. 3-4, pp. 222–262, 2014. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/15427951.2013.865686

[17] ——, "The webgraph framework i: Compression techniques," in *Proceedings of the 13th International Conference on World Wide Web*, ser. WWW '04. ACM, 2004, pp. 595–602.

[18] A. Bavelas, "Communication patterns in task-oriented groups," *Journal of the Acoustical Society of America*, vol. 22, pp. 725–730, 1950.

[19] D. J. Watts, *Small worlds : the dynamics of networks between order and randomness*, 1999.

[20] P. Boldi, M. Rosa, and S. Vigna, "Hyperanf: Approximating the neighbourhood function of very large graphs on a budget," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 625–634.

[21] M. Borassi, P. Crescenzi, M. Habib, W. A. Kosters, A. Marino, and F. W. Takes, "On the solvability of the six degrees of kevin bacon game - A faster graph diameter and radius computation method," in *Fun with Algorithms - 7th International Conference, FUN 2014, Lipari Island, Sicily, Italy, July 1-3, 2014. Proceedings*, 2014, pp. 52–63.

[22] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 177–187.

[23] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, "Four degrees of separation," in *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 2012, pp. 33–42.

[24] P. Boldi and S. Vigna, "In-core computation of geometric centralities with hyperball: A hundred billion nodes and beyond," in *Proceedings of the 13th IEEE International Conference on Data Mining Workshops (ICDM)*, 2013, pp. 621–628.