UDC 004.65:[336.743-021.131]-047.44
https://doi.org/10.33186/1027-3689-2022-11-153-174

Blockchain analysis of the Bitcoin market. (Part 3)

Igor Makarov[1], Antoinette Schoar[2]

[1]London School of Economics Houghton Street London WC2A 2AE UK, i.makarov@lse.ac
[2]MIT Sloan School of Management 100 Main Street, E62-638 Cambridge, MA 02142 and NBER, aschoar@mit.edu

Abstract. The detailed analysis of the Bitcoin network and its main participants. The expert authors (Igor Makarov, London School of Economics, Antoinette Schoar, MIT Sloan School of Management) completed the study authorized by the National Bureau of Economic Research (NBER), the US-based private agency. The Bitcoin network is defined as a new database comprising many of public and proprietary sources to link bitcoin address to real object, and an extensive set of algorithms to extract information on market key players behavior. Three major pieces of analysis of the Bitcoin eco-system were conducted. First, the authors analyze the transaction volume and network structure of the main participants on the blockchain. Second, they document the concentration and regional composition of the miners which are the backbone of the verification protocol and ensure the integrity of the blockchain ledger. Finally, they analyze the ownership concentration of the largest holders of Bitcoin. The researchers found that 1/3 of all bitcoins issued were owned by 10,000 individual investors. They conclude that the high concentration makes the first cryptocurrency market vulnerable to hypothetical hacker attack. The translator notes that paraphrasing English text in Russian was rather challenging due to the newness of the financial agenda and introduction of the term *entity* extensively used in the Western countries though new to Russia. Nevertheless, it is necessary to introduce readers to the bitcoin technology which will be also practical and useful for the library and information community.

Keywords: cryptocurrency, bitcoin, blockchain, transaction, miner, multiple ownership

Cite: Makarov I., Shoar A. Blockchain analysis of the Bitcoin market. (Part 3) / I. Makarov, A. Shoar // Scientific and technical libraries. 2022. No. 11. P. 153–174. https://doi.org/10.33186/1027-3689-2022-11-153-174

4. Miners

Miners are the backbone of the verification process of the Bitcoin blockchain. Their role is to process and verify Bitcoin transactions by solving a computationally difficult problem. For this service, miners are rewarded with newly created Bitcoins and transaction fees.

A proof of work protocol like Bitcoin requires a majority of decentralized miners to be honest for its record keeping function to work. If a single miner or a set of colluding miners were to command a majority of the mining power in the network, the ledger could become controlled by the colluding group and result in the infamous 51% attack, in which the group can alter the previously verified records.

It is therefore important to understand how distributed or reversely how concentrated the mining capacity is. The discussion of miner concentration in the existing literature so far has focused on mining pool concentration. By design, the probability of mining a block and obtaining a block reward in the Bitcoin blockchain is proportional to the hashing power spent on mining. This provides strong incentives for miners to pool their computing power and co-insure each other. As a consequence, mining in the Bitcoin blockchain is dominated by mining pools. Figure 9 (view original figure here: https://www.nber.org/system/files/working_papers/w29396/w29396.pdf) shows the evolution of mining pool shares over time.

Figure 9 shows that mining is dominated by just a few pools. Six out of the largest mining pools are registered in China and have strong ties to Bitmain Techonologies, which is the largest producer of Bitcoin mining hardware, Ferreira et al. (2019). The only non-Chinses pool among the largest pools is SlushPool, which is registered in the Czech Republic.

But while pools function like aggregators of hashing capacity and can therefore have substantial influence over the Bitcoin protocol, they do not necessarily control their miners. As Cong et al. (2020a) emphasize, the power that a pool operator has vis-a-vis the miners depends on the ease with which miners can shift capacity across pools, which in turn depends on the underlying size distribution of the miners. The latter also affects the systemic risk of Bitcoin. The higher is the concentration of mining capacity, the easier it becomes for a hostile party to disrupt or take over the existing mining capacity by (physically) attacking a few miners.

Unlike information about mining pools, which is commonly available, information about individual miners is not readily available[1].

To fill this gap, we use transactions data from the Bitcoin blockchain to trace mining rewards from different pools to the miners that work with them. Since each pool uses its own algorithm to distribute rewards, we build separate algorithms for each pool to map out the pool's distribution dynamic. In Table 1, the number of blocks and bitcoins mined by each pool in 2015-2021 is specified. This is a complex process since pools organize their distribution protocols differently from one another and often accumulate rewards in several layers of distribution addresses before sending them to the miners. The details of how we trace miners are explained in the Appendix.

We track the largest 20 pools except for four Chinese pools: BTCC Pool, BixIn, Huobi Pool, and OKExPool. These four pools are closely integrated with their corresponding exchanges. In particular, their redistribution addresses are held on these exchanges, which impedes the tracing of individual miners. Of the pools we trace, Bitfury and Lubian are private pools, which we treat as single entities. To the best of our knowledge, this is the first study that accurately links miners to their mining pools.

Some miners choose to collect their rewards using their private wallets and some send their rewards directly to their accounts with an exchange or on-line wallet services. We call the former type private-wallet miners and the later exchange-wallet miners. We differentiate between private-wallet and exchange-wallet miners because in the case of private-wallet miners we can more precisely identify the size of a miner since we can assign different mining addresses that belong to the same cluster to one miner. For exchange-wallet miners, we cannot group different addresses together so we treat each exchange mining address as a separate miner. As a result, we can only provide a lower bound for the size of these exchange-wallet miners since a given entity could control several addresses.

---

[1] Miners often use the scriptSig filed to include the name of their mining pool as part of the coinbase transaction, which makes it possible to assign the rewards to pools.

Table 1

Summary statistics for mining pools

Pool name	bitcoins mined	blocks mined
AntPool	876,845	53,535
F2Pool	840,083	51,701
BTC.com	425,200	35,095
BTCC	353,253	17,719
BitFury	351,880	18,185
SlushPool	320,982	21,657
ViaBTC	258,443	21,302
BWPool	250,044	12,733
BTC.TOP	222,190	17,039
Poolin	209,018	19,833
KnCMiner	109,923	4,466
Huobi Pool	86,571	9,044
Bixin	80,682	5,778
GHash.IO	47,644	1,912
1THash	42,711	4,780
Eligius	41,002	1,650
OKExPool	40,241	3,957
Binance Pool	32,395	4,683
BTC Guild	24,731	985
WAYI.CN	17,486	1,465
Lubian.com	13,279	1,783
BytePool	12,712	1,002
BATPOOL	6,266	441
SpiderPool	4,367	493
tigerpool.net	3,629	285
Sigmapool.com	2,204	217

To separate private-wallet miners from exchange-wallet miners we first check if a miner's address belongs to a known exchange or entity. Since our data can miss some exchanges or OTC desks, we treat all miner addresses that belong to suspiciously large clusters as exchange-wallet clusters. These are clusters that (1) consist of many addresses, (2) receive a large number of bitcoins that cannot be traced to mining activity, (3) have many mining addresses as their members. This means we err on the side of being conservative when defining miner size. In the next step, we screen out entities that receive irregular rewards and that received less than $1000 or fewer than 25 times of reward over their lifetime. Finally, we manually check the largest 150 largest independent-wallet miners by USD rewards to ensure that we are not mistaking re-distribution addresses for miners. After applying these filters, we end up with 105,494 private-wallet clusters and 137,656 exchange-wallet addresses. The exchange-wallet addresses belong to 305 known exchanges and on-line wallets and 284 unknown clusters. Since a miner's reward is proportional to its mining capacity we measure each miners' capacity as the bitcoins that are sent by pools through distribution transactions[2].

In Figure 10 (view original figure here: https://www.nber.org/system/files/working_papers/w29396/w29396.pdf) we plot how our algorithm captures the mining capacity in the Bitcoin blockchain from January 2015 till the beginning of 2021 as a proportion of all coinbase rewards that are available in a given week. The blue line shows the rewards that are captured by the pools that we can trace. This information is obtained from public information by the mining pools at an aggregate level. Early in the sample, our mining pools cover about 60% of the mining rewards, but by the end of the sample, this number is close to 90%. The red line shows the distributed mining rewards that we can trace on the blockchain from the pool's distribution address to the underlying miners, for our twenty mining pools. We can see that we are able to trace about 90% of the pool rewards. Finally, the green line in Figure 10 shows that rewards collected by exchange-wallet miners. It shows that exchange-wallet and private-wallet miners each command about 50% of total capacity.

---

[2] Pools differ in the amount they charge their miners and payout schemes, see Cong et al. (2020a). Because pools compete with each other we expect these differences to have a small impact on measuring miners' capacity.

4.1. Concentration of Mining Capacity

We now analyze the concentration of mining capacity across individual miners. Each month, we sort active miners by their size and calculate what percentage of total mining capacity is controlled by different quantiles. The results for the top 50%, 10%, 5%, 0.5%, and 0.1% miners are presented in Figure 11 left panel (view original figure here: https://www.nber.org/system/files/working_papers/w29396/w29396.pdf). The figure shows that Bitcoin mining is concentrated and the concentration of mining capacity has been relatively stable over time. The top 50% of miners control almost all mining capacity. Top 10% control 90% and just 0.1% control close to 50%.

Next, we calculate how many miners are necessary to cover 10%, 20%, 30%, 40%, or 50% of total mining capacity. Figure 11 right panel shows that for the 50% threshold, which is of particular interest because of the dangers of a 51% attack, between 2015 and 2017 it typically took less than 50 miners. At the beginning of 2018, the number was as high as 250 miners, but by the end of 2020 fell again under 50 miners. Assuming that missing pools have similar concentration and given that by the end of 2020 we trace about 90% of all mining pool capacity, our results suggest that by the end of 2020, the largest 55–60 miners controlled at least half of all Bitcoin mining capacity. Figure 11, right panel, also highlights that the concentration of mining capacity is counter-cyclical. It decreases following sharp increases in the Bitcoin price and increases in periods when the price drops such as in 2018.

Also, concentration increases after the Bitcoin halving dates – the dates when the block reward halves, July 2016 and May 2020 in our sample. These results suggest that the set of large miners is relatively stable, and it is small miners which enter and leave the mining business in response to price shocks. Thus, the risk of the 51% attack increases in times when the Bitcoin price drops precipitously or following the halving events.

4.2. Geographic Concentration of Miners

Next, we investigate the geographic distribution of miners, which has been another area of concern. Having control over a majority of mining capacity, de facto, means control over a cryptocurrency. As a result, geographic concentration increases the risk that a private or a state actor in one part of the world, could gain control over the network and inflict large losses on the general public and financial institutions if they are holding bitcoins.

Determining the geographical distribution of miners is not an easy task. So far, the main data has come from the analysis of miners' IP addresses[3].

When a miner connects to a pool server, the pool operator can see the IP address of the miner. Unless a miner uses a VPN address, the pool operator can use this IP address to determine the geographical location.

In this paper, we utilize a new approach, which takes advantage of our ability to trace miners on the blockchain. Since we can observe miners' addresses on the blockchain we can also see at which exchanges they cash out their rewards. We conjecture that miners in a particular region would most likely send their rewards to an exchange that is prevalent in this region. By studying to which exchanges miners send their rewards we can infer their location.

There are several advantages of our method over existing ones. First, we are able to cover the majority of the universe of miners and not only a few select pools. Second, our method may give a more accurate picture than using IP addresses, especially for miners that operate in countries where mining is restricted. In such countries, miners might deliberately hide their location or instruct pools not to reveal their location in fear of information being revealed to the local authorities or regulators.

One limitation of our approach is that some exchanges are not region-specific, but operate across many jurisdictions. Since miners can send bitcoins to such internationally accessible exchanges independent of the miner's location, observing flows to them does not necessarily tell us where the miner is located. To capture these exchanges, we create a separate category that we call International. As a result, we end up classifying exchanges into four large categories: Chinese, US-Europe, International, and Other. The International category includes exchanges that operate across many jurisdictions, and rely on stable coins like tether; examples are exchanges such as Binance and Gate.io. The Other category includes all identified exchanges in regions outside the above ones. Table 2 shows the regional distribution of exchanges.

---

[3] One of the best-known data providers based on this approach, Cambridge Center for Alternative Finance, has been collecting aggregated data from three pools: BTC.com, Poolin, ViaBTC, and recently from Foundry USA.

Table 2

Location of exchanges

Exchange name	Region
Binance US	US/Europe
Bitstamp	US/Europe
Coinbase	US/Europe
Coinsquare	US/Europe
Gemini	US/Europe
Kraken	US/Europe
Liquid	US/Europe
LocalBitcoins	US/Europe
Paxful	US/Europe
Uphold	US/Europe
BTCChina	China
Bitkan	China
BixIn	China
Bkex	China
EXX	China
Huobi	China
MXC.com	China
OkCoin	China
Allcoin	International
BCEX	International
Bibox	International
BigONE	International
Binance	International
Bit-Z	International
BitForex	International
Bitfinex	International
Bittrex	International
Cobinhood	International
CoinEgg	International
CoinEx	International
Gate.io	International
HitBTC	International
Kucoin	International
OKEx	International
Poloniex	International
Tidex	International
ZB.com	International

Using this proxy for miner location, Figure 12 Panels A and B (view original figure here: https://www.nber.org/system/files/working_papers/w29396/w29396.pdf) show how the mining capacity is distributed across regions. Panel A plots the monthly value of Bitcoin rewards that are cashed out by miners in different regions and Panel B the percentages across different regions[4].

Starting in 2015 we see that a majority of mining capacity is located in China, between 60% to 80% in the period between 2015 and the middle of 2017. After the second half of 2017 we see a slight drop in the mining capacity of miners that cash out on Chinese exchanges, the fraction falls to 50%. However, at the same time, we see a significant increase in the miners that cash out on International exchanges, in particular on Binance. Binance was founded in 2017 and quickly became one of the largest and liquid exchanges, which made it an attractive trading venue for miners to cash out their rewards. We show in the next section that it is the second most popular destination after Huobi among Chinese miners. Taken together the monthly bitcoins cashed out on Chinese and International exchanges suggest that since 2017, Chinese miners have dominated the mining landscape and accounted for about 70% of total mining capacity, which is in line with previous estimates.

4.3. Xinjiang Event

In order to verify the validity of our approach of identifying miner locations by looking at where miners cash out their Bitcoin rewards, we take advantage of a recent incidence in the Xinjiang province of China.

In April of 2021, a major coal mine was flooded and killed several miners.

In response to the event, the Chinese government shut down the mine for the weekend of April 17–18, 2021 and with it, the electricity supply for the whole region was shut down.

Typically this is a region that has heavily subsidized electricity prices due to the abundant energy from coal mining and thus has attracted a lot

---

[4] In this graph we focus on rewards cashed out by exchange-wallet miners and private pools. Many large private-wallet miners tend to accumulate their rewards over time, and some do not cash them out at all. The regional distribution of private-wallet miners that cash out their rewards is in line with that of the exchange-wallet miners.

of Bitcoin miners to locate there. During the time of the accident, worldwide Bitcoin mining capacity dropped by over 35%. Since only miners that were physically located in Xinjiang province were directly affected by the shutdown, by identifying miners for whom hashing capacity dropped significantly during the weekend of April 17-18 2021, we can precisely pinpoint miners that must be physically located in this region of China. Since most of the large miners in China are operating across multiple locations within the country, we do not necessarily expect that many miners have a 100% drop.

To identify affected miners with a high degree of accuracy, we focus on those that received rewards every day in the period before April 8. This approach allows us to identify a total of 5012 miners. We measure capacity based on the coinbase rewards that miners received. Figure 13 plots the time series of miners that lost more than 20% hashing capacity between April 8 and May 8. We see that there are 1,158 miners that lost 20%, 804 miners that lost more than 50% of their mining capacity, and 460 miners which lost 100% of income (view original figure here: https://www.nber.org/system/files/working_papers/w29396/w29396.pdf). After the coal mine was reopened and access to electricity was restored, we see a swift return to almost the same level of capacity as before the event. But some of the smallest miners seem to have dropped off.

If we take the 804 miners that lost more than 50% of their hashing capacity due to the event, 608 of them come back on-line by April 23. Out of these miners 403 are exchange miners. This set of miners uses the following exchanges to trade Bitcoin in the period before the mining accident: Huobi (42%), Binance (10%), OKEx (9%), BixIn 25 (6%), EXX (4%), Bit.com (4%), and 15% is cashed on unknown exchanges. We only use the period before the mining accident to abstract from any disruptions that might have happened due to the accident. For the 205 independent miners, 140 sent Bitcoin to named entities. The exchanges used by the majority of these independent miners are again: Huobi (40%), Binance (26%), OKEx (8%), and BixIn (4%). The results validate our assignment of Chinese exchanges since we see that this set of miners, for whom we know that they are located in China, are using predominantly China-origin exchanges and Binance. More generally the results provide support for our approach of using the region where miners cash out their Bitcoin rewards to determine their geographic location.

5. Ownership of Bitcoin

Since the inception of Bitcoin in 2009, there has been intense interest in the question of who are the largest owners of Bitcoin, and how much they actually own. There are websites dedicated to tracking the addresses with the largest Bitcoin holdings, the so-called "rich list", one of the most well-known and widely followed lists in the crypto community. But the question of ownership concentration is not only a matter of curiosity and intrigue. From a public policy perspective, it is important to understand the ownership and concentration of Bitcoin holdings since it determines who is positioned to benefit most from any price appreciation. Are these a select few investors or the general public? To shed light on these questions, we study the ownership and concentration of Bitcoin holdings as of the end of 2020.

Determining the concentration of ownership is more complicated than just tracking the holdings of the richest addresses since not all large addresses represent individuals. Many public entities, e. g., exchanges and on-line wallets, hold Bitcoin on behalf of other investors. Therefore, the first step in our analysis is to differentiate between addresses belonging to individual investors and those belonging to intermediaries.

When market participants deposit their bitcoins with exchanges or on-line and custodial wallets they forfeit their bitcoins to the exchange. Exchanges usually mix all deposits together and store them in the so-called cold wallets – Bitcoin addresses stored on special devices not connected to the Internet because of security concerns.

A given intermediary typically has only a few Bitcoin addresses that constitute its cold wallet but these addresses hold very large balances. For example, the cold wallet of Binance, which is one of the largest cold wallets, holds 300,000 bitcoins as of the end of June, 2021[5]. However, not all exchanges have a cold wallet that is as distinct as Binance's cold wallet. Because cold wallets typically consist of few addresses and send and receive funds only infrequently, the default clustering algorithm in many cases does not link them to the corresponding hot wallets of exchanges. Therefore, identifying cold wallets presents a significant challenge.

---

[5] https://bitinfocharts.com/bitcoin/wallet/Binance-coldwallet.

To address this challenge, we scrutinize the addresses in the "rich" list that have a balance of at least 1000 bitcoins as of Dec 31, 2020. There were 2258 such addresses, which controlled 7.9 million bitcoins – almost half of all bitcoins in circulation. Since cold wallets hold large balances, their addresses are very likely among these "rich" addresses. The fact that so few addresses control almost half of the bitcoins in circulation is often taken as prima facie evidence of the high concentration of Bitcoin holdings. This view, however, neglects the fact that some of these addresses belong to cold wallets and therefore, represent holdings of a large number of people.

We deal with the shortcomings of the default clustering algorithm by developing a suite of algorithms based on graph analysis to classify addresses into two groups: addresses that belong either to individual investors or those that belong to intermediaries. For each rich address, we first check if it belongs to a cluster identified in our database. If the address does not belong to any known entity we build a network of clusters that sends bitcoins to this address (or the cluster that contains this original address). This is a recursive process. First, we find clusters that send their balances directly to the address. In many cases, there is a unique such cluster. For example, 1GR9qNz7zgtaW5HwwVpEJWMnGWhsbsieCG receives all its balance from another address 1MzG9Gx5G3ZTXtEQT4FJg23Cb3g S6UF982 on May 17, 2018, which in turn gets all its balance from an unknown old large cluster that dates back to 2014.

The cases where there is a unique parent cluster at each step are particularly simple. Here we stop the process if (1) we reach a cluster that belongs to a known entity, or (2) we reach a large unknown cluster, or (3) we reach a sufficiently old cluster, which we know is not a cold wallet of any exchange or online wallet. In the first case, if a known entity is an active intermediary, e. g., exchanges or online wallet, we mark the rich address as linked to an intermediary entity. If the known entity is an individual entity, e. g., a miner, or defunct intermediary we mark it as belonging to an individual. In the second case, if a large unknown cluster is an active cluster, we classify the initial rich address as linked to an intermediary, or to an individual investor, otherwise. Finally, in the last case, we classify the initial rich address as belonging to an individual investor. In the case where a rich address receives its balance from several

clusters, we continue tracing flows to each parent cluster. The following outcomes are typically realized. First, the process can link the address to a network dominated by a single large cluster, in which case we follow the same classification rules as in the case of a 27 unique parent cluster. For example, Figure 20 (view original figure here: https://www.nber.org/system/files/working_papers/w29396/w29396.pdf) shows the network realized from tracing flows to 1P5ZEDWTKTFGxQjZphgWPQUpe554WKDfHQ (abbreviated as 1P5ZE, which has been the third richest address at the time of writing this paper. The picture shows that all its flows originate from a single cluster containing address 1FzWLkAahHooV3kzTgyx6qsswXJ6sCXkSR (abbreviated as 1FzWL). The latter cluster is an active large unidentified cluster, which mostly interacts with major exchanges. Therefore, we classify 1FzWL as an intermediary. Since 1P5ZE not only receives flows from 1FzWL but also sends them back we conclude that 1P5ZE is a cold wallet of 1FzWL.

The second common outcome is when the address' balance is traced to at least two known entities. Unless the address belongs to a large active cluster we mark the address as individual in this case. Finally, in a few cases where we are uncertain about whether an address belongs to an intermediary or an individual, we mark those addresses as ambiguous. Overall, out of the total 2258 rich addresses, we classify 1 013 as individual, 1 154 as linked to intermediaries, and 47 as ambiguous. Figure 21 shows the amount of Bitcoin held in the wallet of intermediaries over time (view original figure here: https://www.nber.org/system/files/working_papers/w29396/w29396.pdf). The balance held at intermediaries started accelerating in 2014 has been steadily increasing over time. By the end of 2020 it was equal to 5.5 million bitcoins, roughly one-third of Bitcoin in circulation at the time.

We now contrast the holdings of intermediaries with those of individuals, which we proxy for in two ways. First, we include rich addresses that we classified as individual in our analysis of "rich" addresses. Second, we include all unknown clusters that had a balance between 1 and 1000 bitcoins on Dec 31, 2020 and that have not been active in the entire year of 2020. We impose the inactivity constraint to separate individual wallets from wallets that might possibly belong to intermediaries. Some of these clusters might be old or even forgotten

addresses, and others are likely to belong to long-term investors. There are 400,000 of such clusters and they collectively control 8.5 million bitcoins by the end of 2020. This is 3 million bitcoins more than what is held in exchange wallets.

Figure 22 (view original figure here: https://www.nber.org/system/files/working_papers/w29396/w29396.pdf) shows the evolution of the individual bitcoin balances over time. In Panel A we calculate the date of the first transaction for each individual cluster and consider it as a proxy for the age of this cluster. We then assign the balance a cluster holds at the end of 2020, to the inception date of the cluster. This allows us to decompose the holdings of individual investors as of 2020 into the age of the owners. Panel B shows how the balances accumulated over time.

The results show there were a few time periods when substantial balances of bitcoins were established. First, there are more than 1 million bitcoins mined by the investor of Bitcoin, Satoshi Nakamoto, in the early days of Bitcoin blockchain. The true identity of Satoshi Nakamoto remains unknown to this date, and with it, the ownership of these early bitcoins. Other periods when substantial balances were accumulated coincide with times of very rapid Bitcoin price appreciation and subsequent crashes such as 2014, end of 2017, and beginning of 2018.

In a final step, we now look at the concentration of individual Bitcoin ownership. In Figure 23, we sort individual clusters according to their balance at the end of 2020 and plot their cumulative balance against the number of individual clusters that are holding these bitcoins. Figure 23 (view original figure here: https://www.nber.org/system/files/working_papers/w29396/w29396.pdf) shows that participation in Bitcoin is still very skewed toward a few top players even at the end of 2020. We see that only 1000 clusters control three million bitcoins and the top 10,000 own more than five million bitcoins which is about a quarter of all outstanding bitcoins.

It is also important to note that this measurement of concentration most likely is an understatement since we cannot rule out that some of the largest addresses are controlled by the same entity. In particular, in the above calculations, we do not assign the ownership of early bitcoins, which are held in about 20,000 addresses, to one person (Satoshi Nakamoto) but consider them as belonging to 20,000 different individuals.

6. Conclusions

We study the transaction behavior and ownership patterns of the main market participants in the Bitcoin eco-system using data from the Bitcoin blockchain. Our analysis highlights three major sets of findings. First, we show that exchanges play a central role in the Bitcoin system. They explain 75% of real Bitcoin volume, while other types of transactions, such as illegal transactions or mining rewards, explain only a minor part of total volume. Exchanges are also the most connected nodes on the blockchain. The strong interconnectedness of exchanges and the ease with which tainted bitcoins can be intermingled with clean volume, has important implications for the transparency and traceability of transactions, and the enforcement of Know-YourCustomer (KYC) norms across the network.

Second, we document the concentration and regional composition of Bitcoin miners, the entities providing the verification of transactions on the Bitcoin platform. Unlike 29 information about mining pools, information about individual miners was previously not available. We show not only is the Bitcoin mining capacity highly concentrated, but it varies counter-cyclically with the Bitcoin mining rewards. As a result, the risk of a 51% attack increases in times when the Bitcoin price drops precipitously or after the halving events.

Third, we study the ownership and concentration of Bitcoin holdings. We show that while the balances held at intermediaries have been steadily increasing since 2014, even by the end of 2020 it comprises only 5.5 million bitcoins, about one-third of Bitcoin in circulation. In contrast, individual investors collectively control 8.5 million bitcoins, almost half the bitcoins in circulation by the end of 2020. Within individual holdings, there is significant skewness in ownership.

Our results suggest that despite the significant attention that Bitcoin has received over the last few years, the Bitcoin eco-system is still dominated by large and concentrated players, be it large miners, Bitcoin holders or exchanges. This inherent concentration makes Bitcoin susceptible to systemic risk and also implies that the majority of the gains from further adoption are likely to fall disproportionately to a small set of participants.

Appendix

Pass-through volume

Many Bitcoin clusters have a very short lifespan and are therefore unlikely to represent stand-alone or economically independent entities. In what follows, we call these clusters short-term clusters. These types of pass-through addresses are often created by wallet programs or are part of a user's attempt to either consolidate their Bitcoin addresses or create possible divisions of their holdings. We reassign volume associated with short-term clusters to the clusters that directly interact with short-term clusters, and eliminate short-term clusters from further analysis. In doing so, we differentiate between two cases shown in Figure 14. In the first case, depicted in the left panel, a short-term cluster P has a single incoming transaction and a single outgoing transaction. In the second case, depicted in the right panel, a short-term cluster can have multiple incoming and outgoing transactions. We separate the two cases because the first case is much more prevalent and significantly easier to deal with. There are 256 million clusters of the first type and 34 million of the second type, correspondingly. These clusters account for 53% and 4% of the full blockchain volume, respectively. 99:7% of the first type of clusters consist of a single address.

Formally, we classify a cluster as a short-term cluster of the first type if the following four conditions are satisfied.

1. The cluster has only one incoming transaction and one outgoing transaction.

2. The cluster has no balance left after the two transactions.

3. The time difference between its two transactions is less than a week, or fewer than 1068 blocks on the blockchain.

4. The incoming transaction is not a CoinJoin transaction.

For a non-CoinJoin transaction, the first condition ensures (with the default clustering algorithm) that the short-term cluster receives its flows from a single cluster (cluster A in the picture). This makes it straightforward to eliminate the short-term cluster and reassign its volume: we simply record volume from P to Bi as volume from A to Bi, i = 1; ::; N:

The default BlockSci clustering algorithm treats CoinJoin transactions separately and does not automatically group sending addresses together. As a result, in this case, the short-term cluster receives its flows from several different clusters, and becomes a special case of the second type of cluster.

We classify a cluster as a short-term cluster of the second type if the following three conditions are satisfied.

1. The cluster's current balance is less than 0:001 BTC.

2. The time difference between the cluster's first transaction and its last transaction is less than one week, or fewer than 1068 blocks on the blockchain.

3. The cluster is created at least one week before the end of the database.

The main complication with factoring out short-term clusters of the second type arises from the fact some of them may form a cycle. For example, Figure 15 depicts a situation where two short-term clusters P_1 and P_2 send flows p_{12} and p_{21} to each other.

Elimination of short-term clusters of the second type, which are not part of any cycle, is straightforward: we record volume from Aj , j = 1; :::;M to Bi, i = 1; :::; N as

$$\frac{w_j}{\sum_{k=1}^{M} w_k} \times v_i, \tag{1}$$

see Figure 14. When short-term clusters form a cycle, e. g., as shown in Figure 15, this procedure leads to an infinite recursion. To avoid it, consider the map F defined as

$$F = \begin{pmatrix} \frac{w_1+p_{21}}{w_1} & -\frac{p_{12}}{w_2} \\ -\frac{p_{21}}{w_1} & \frac{w_2+p_{12}}{w_2} \end{pmatrix} \tag{2}$$

Note that

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} p_{21} + w_1 - p_{12} \\ p_{12} + w_2 - p_{21} \end{pmatrix} = F \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \quad (3)$$

where we used the fact the each short-term cluster Pi has to have zero balance. Therefore,

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = F^{-1} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$
$$= \frac{1}{w_1 p_{12} + w_2 p_{21} + w_1 w_2} \begin{pmatrix} w_1 p_{12} + w_1 w_2 & w_1 p_{12} \\ w_2 p_{21} & w_2 p_{21} + w_1 w_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}. \quad (4)$$

The matrix F^{-1} defines a map from.

$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \text{ to } \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$$

In a general case, where n short-term clusters from a cycle, the matrix F can be constructed as follows. First, for each short-term cluster P_k let w_k be the total inflows from all non-short-term clusters to P_k, v_k be the total outflows from P_k to all non-short-term clusters, and p_{kl} and p_{lk} be the flows from P_k to P_l and from P_l to P_k, respectively. Define matrix T as follows:

$$T_{ij} = \begin{cases} -p_{ij}, & \text{for } i \neq j \\ \sum_k p_{ki}, & \text{for } i = j \end{cases}$$

Let I(n) be the n-by-n identity matrix and W be a diagonal matrix with diagonal elements Wii = wi; i = 1; :::; n: Then $F = I + TW^{-1}$:

We partition all interconnected short-term clusters of the second type into disjoint components using Julia LightGraphs package and its strongly connected components routine[6]. For each strongly connected

[6] See Bondy and Murty (2008), 3.4 and https://github.com/JuliaGraphs/LightGraphs.jl for more details.

component, we construct matrix F, as described above, and compute its inverse. Finally, we use matrix F^{-1} to factor out volume of short-term clusters that belong to this component.

Identifying miners from mining pools

We use the data collected from BTC.com to find out which block was mined by which pool. Table 1 provides summary statistics of the mining pools. It reports the total number of blocks and Bitcoin mined by each pool. We trace the pools which are marked in bold font. Private pools are marked in italic.

In what follows, we document how we trace miners using one of the largest pools, AntPool, as an example. We start our analysis by identifying a pool's coinbase reward collection addresses. We collect these addresses by looking at the coinbase transactions of the blocks that are mined by this pool. Figure 16 shows an example of such a transaction in Block 684887 for AntPool. As a reward for its mining effort in this transaction, AntPool collected 6.25 BTC in block rewards and 0.56 BTC in transaction fees using address 12dRugNcdxK39288NjcDV4GX7rMsKCGn6B. The coinbase signature of AntPool is underlined in red.

Typically, pools use few addresses to collect their coinbase rewards. For example, AntPool over its history has used a total of 72 addresses, and in fact collected most of its rewards only in two addresses, 1Nh7u... and 12dRu... since 2018. Figure 17 shows a time-series of the decomposition of the rewards collected by each of these collection addresses.

Having collected mining rewards, pools then distribute them back to the miners that work with the pool. Each pool uses its own distribution algorithm. Typically, pools first pass on the rewards to a set of designated distribution addresses, which then distribute rewards to individual miners. Figure 18 shows the flow chart for AntPool. The coinbase collection addresses are marked in light green and designated distribution addresses in light blue. In the case of AntPool there are 13 designated distribution addresses, which distribute 97% of the total rewards. We create similar flow charts for each of the other pools to identify their designated distribution addresses.

Since pools employ many miners it is usually impossible to distribute rewards to all miners in one transaction. Therefore, many pools use long

peeling chains to accomplish this task. The distribution of the rewards starts from a designated distribution address. It distributes the rewards to a large number of miners; collects the change in a new one-off address that distributes the reward to the next set of miners, and so on. Figure 19 shows the first two steps. In the first step, a designated distribution address 1F4JZ... of AntPool starts with a balance of 100 bitcoins. It sends rewards to 100 miners and collects the change at a new one-off address bc1q0m... The latter address then immediately distributes the rewards to the next 20 miners. This recursive process continues for another 152 levels. At each level, a one-off address is created to distribute the majority of the remaining rewards to more miners. In the end, the remaining 0:002 bitcoins are sent to just two miners.

In the next step, we take all distribution transactions and collect all output addresses that take part in these transactions. Occasionally, some pools use distribution addresses for other purposes, possibly buying equipment or the like. Therefore, we eliminate from this set of addresses any/internal" addresses that belong to the pool. The remaining addresses are candidates for addresses of individual miners. There are a total of 1.1 million of such addresses. To eliminate/recreational" miners, we filter out addresses that receive rewards with an equivalent value of less than $1,000 or that have fewer than 25 reward distributions over the entire sample period.

Finally, we allow for the possibility that some of the remaining addresses might not belong to individual miners but to smaller pools that do mining operations as part of a larger pool, or belong to a subsidiary or a partner of the larger pool. To screen out these addresses we check if

1. An address systematically sends some of its rewards to other miners' addresses.

2. The address rewards are unstable over time or come in integer numbers.

We drop all addresses with irregular distributions, and further trace the addresses that send to other miners' addresses. Lastly, we manually examine the reward distributions of the 150 largest addresses to verify that they indeed look like they belong to individual miners.

References

1. **Abadi J. and Brunnermeier M.** (2018). Blockchain economics. Working Paper 25407, National Bureau of Economic Research.

2. **Athey S., Parashkevov I., Sarukkai V., and Xia J.** (2016). Bitcoin Pricing, Adoption, and Usage: Theory and Evidence. Research Papers 3469, Stanford University, Graduate School of Business.

3. **Biais B., Bisiere C., Bouvard M., and Casamatta, C.** (2019). The blockchain folk theorem. The Review of Financial Studies, 32 (5):1662–1715.

4. **Bondy J. and Murty U.** (2008). Graph Theory. Springer Publishing Company, Incorporated, 1st edition.

5. **Budish E. (2018).** The economic limits of bitcoin and the blockchain. Working Paper 24717, National Bureau of Economic Research.

6. **Cong L. W., He Z., and Li J. (2020a).** Decentralized Mining in Centralized Pools. The Review of Financial Studies, 34 (3):1191–1235.

7. **Cong L. W., Li Y., and Wang N.** (2020b). Tokenomics: Dynamic Adoption and Valuation. The Review of Financial Studies, 34 (3):1105–1155.

8. **Easley D., O'Hara M., and Basu S.** (2019). From mining to markets: The evolution of bitcoin transaction fees. Journal of Financial Economics, 134 (1):91–109.

9. **Ferreira D., Li J., and Nikolowa R.** (2019). Corporate capture of blockchain governance. Working paper, London School of Economics.

10. **Foley S., Karlsen J. R., and Putni¸nˇs T. J.** (2019). Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies? The Review of Financial Studies, 32 (5):1798–1853.

11. **Freeman T. C., Horsewell S., Patir A., Harling-Lee J., Regan T., Shih B. B., Prendergast J., Hume D. A., and Angus T.** (2020). Graphia: A platform for the graph-based visualisation and analysis of complex data. bioRxiv.

12. **Griffin J. M. and Shams A.** (2020). Is bitcoin really untethered? The Journal of Finance, 75 (4):1913–1964.

13. **Han B. Y. and Makarov I.** (2021). Feedback trading and bubbles. Working paper, London School of Economics.

14. **Hastie T., Tibshirani R., and Friedman J.** (2001). The Elements of Statistical Learning. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

15. **Hardle W. K., Harvey C. R., and Reule R. C. G.** (2020). Understanding cryptocurrencies. Journal of Financial Econometrics, 18 (2):181–208.

16. **Huberman G., Leshno J. D., and Moallemi C.** (2021). Monopoly without a monopolist: An economic analysis of the bitcoin payment system. The Review of Economic Studies.

17. **Makarov I. and Schoar A.** (2020). Trading and arbitrage in cryptocurrency markets. Journal of Financial Economics, 135 (2):293–319.

18. **Meiklejohn S., Pomarole M., Jordan G., Levchenko K., McCoy D., Voelker G. M., and Savage S.** (2013). A fistful of bitcoins: Characterizing payments among men with no names. In Proceedings of the 2013 Conference on Internet Measurement Conference, IMC '13, page 127–140, New York, NY, USA. Association for Computing Machinery.

19. **Newman M. E. J.** (2010). Networks: an introduction. Oxford University Press, Oxford; New York.

20. **Pagnotta E. (2020).** Decentralizing money: Bitcoin prices and blockchain security. Review of Financial Studies.

21. **Pagnotta E. and Buraschi A.** (2018). An equilibrium valuation of bitcoin and decentralized network assets. Working paper, Imperial College.

22. **Prat J. and Walter B. (**2021). An equilibrium model of the market for bitcoin mining. Journal of Political Economy, 129 (8):2415–2452.

23. **Raskin M. and Yermack D.** (2016). Digital currencies, decentralized ledgers, and the future of central banking. Working Paper 22238, National Bureau of Economic Research.

24. **Ron D. and Shamir A.** (2012). Quantitative analysis of the full bitcoin transaction graph. IACR Cryptology ePrint Archive. P. 584.

25. **Schilling L. and Uhlig H.** (2019). Some simple bitcoin economics. Journal of Monetary Economics, 106 (C):16–26.

26. **Sockin M. and Xiong W. (2020).** A model of cryptocurrencies. NBER Working Paper 26816, National Bureau of Economic Research.

Information about the authors

Igor Makarov – London School of Economics Houghton Street London WC2A 2AE UK

i.makarov@lse.ac

Antoinette Schoar – MIT Sloan School of Management 100 Main Street, E62-638 Cambridge, MA 02142 and NBER

aschoar@mit.edu