ISSN 0956-8549-844

Blockchain Analysis of the Bitcoin Market

By

Igor Makarov

Antoinette Schoar

DISCUSSION PAPER NO 844

PAUL WOOLLEY CENTRE WORKING PAPER No 83

October 2021

Any opinions expressed here are those of the authors and not necessarily those of the FMG. The research findings reported in this paper are the result of the independent research of the authors and do not necessarily reflect the views of the LSE.

Blockchain Analysis of the Bitcoin Market

Igor Makarov[1] and Antoinette Schoar*[2]

[1]London School of Economics
[2]MIT Sloan, NBER, CEPR, Ideas42

October 13, 2021

Abstract

In this paper, we provide detailed analyses of the Bitcoin network and its main participants. We build a novel database using a large number of public and proprietary sources to link Bitcoin addresses to real entities and develop an extensive suite of algorithms to extract information about the behavior of the main market participants. We conduct three major pieces of analysis of the Bitcoin eco-system. First, we analyze the transaction volume and network structure of the main participants on the blockchain. Second, we 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, we analyze the ownership concentration of the largest holders of Bitcoin.

JEL classification: G12, G15, F38

*Igor Makarov: Houghton Street, London WC2A 2AE, UK. Email: i.makarov@lse.ac.uk. Antoinette Schoar: 62-638, 100 Main Street, Cambridge MA 02138, USA. Email: aschoar@mit.edu. We thank Jiageng Liu for excellent research assistance. We also thank Kyrylo Chykhradze for very helpful comments and help with the Crystal Blockchain data.

Introduction

Cryptocurrencies have seen a remarkable growth in value and public attention since their inception more than a decade ago. Opinions about the impact of cryptocurrencies range all the way from being a revolution in financial access to a threat to financial stability and monetary policy. A distinguishing feature of cryptocurrencies is the promise of a decentralized system of payments or store of value outside the traditional nexus of government scrutiny. The blockchain technology at the heart of cryptocurrencies replaces the reliance on a few centralized record keepers, such as banks or credit card networks, with a large set of decentralized and anonymous agents. The absence of centralized accountability and the anonymity of its users are often viewed as major benefits by crypto supporters, but it hinders the timely diagnosis of the health of the system, generates many challenges for regulators, and introduces new sources of systematic risk.

Bitcoin, the original cryptocurrency, is still the largest and most popular coin, with a market cap that is larger than all the other coins combined. It is often seen as a template or point of comparison for other new coins. Many industry participants are now calling for even wider Bitcoin adoption, either as a public investment vehicle or legal tender. These pressures put regulators who want to find the right balance between protecting the public interest and allowing innovation in a difficult position. There are still many open questions about the utilization of bitcoin, its ownership concentration as well as the structure of core entities that form the backbone of the Bitcoin ecosystem, despite being in existence for more than ten years. A better understanding of the Bitcoin network and its participants is required for any decision about how and whether to integrate Bitcoin into the traditional financial system.

In this paper, we aim to shed light on these open questions by developing a novel database that allows us to document the evolution of the Bitcoin market and its different participants over time. To build this database we use a large number of public and proprietary sources that link Bitcoin addresses to real entities and develop a suite of algorithms that use the semi-public nature of the Bitcoin blockchain to extract information about the behavior of the main market participants. We believe that this is the most complete Bitcoin database used in academic research to date.

We conduct three major pieces of analysis that focus on the main participants of the blockchain eco-system. First, we analyze the transaction volume and network structure of the main participants on the Bitcoin blockchain. Second, we document the concentration and regional composition of miners which ensure the integrity of the blockchain ledger. Finally, we analyze the ownership concentration of the largest holders of Bitcoin.

Transaction Volume and Network Structure. We first document that 90% of transaction volume on the Bitcoin blockchain is not tied to economically meaningful activities but is the byproduct of the Bitcoin protocol design as well as the preference of many participants for anonymity. Because the Bitcoin blockchain is a public ledger all payment flows between addresses are perfectly observable. Therefore, many bitcoin users adopt strategies designed to impede the tracing of bitcoin flows by moving their funds over long chains of multiple addresses and splitting payments among them resulting in a large amount of spurious volume. We develop algorithms to filter out this spurious volume and trace economically meaningful payments between real entities on the Bitcoin network.

We show that the vast majority of Bitcoin transactions between real entities are for trading and speculative purposes. Starting from 2015, 75% of real bitcoin volume has been linked to exchanges or exchange-like entities such as on-line wallets, OTC desks, and large institutional traders. In contrast, other known entities are only responsible for a minor part of total volume. For example, illegal transactions, scams and gambling together make up less than 3% of volume.[1] The fraction of volume explained by miners is even smaller.

Exchanges not only generate the most volume, but they are also the most connected nodes in the Bitcoin network. In particular, they have the highest measure of eigenvalue centrality.[2] Furthermore, a large fraction of exchange volume consists of cross-exchange flows. The high cross-exchange flows are the consequence of the current market structure. Different from traditional, regulated exchanges, cryptocurrency markets consist of many non-integrated and independent exchanges without any provisions to ensure that investors receive the best price when executing trades. As a result, the consistency of the Bitcoin price across exchanges depends on arbitrageurs and speculators who trade across them. In support of this idea, we show that exchanges that trade similar currency pairs have higher cross-exchange flows.

The strong interconnectedness of exchanges has important implications for the

[1] Our estimates of illegal transactions are much smaller than the previous literature found, see for example Foley et al. (2019). One reason for this difference is that we have a much more detailed and comprehensive identification of participants on the blockchain. The prior work had to rely on an imputed network of illegal entities where any Bitcoin address recursively is classified as belonging to an illegal entity if the majority of its transactions is with addresses that themselves were previously classified as illegal. However, this method leads to significant overstatement of illegal volume, since it does not discriminate between real users and spurious volume.

[2] See Section 3.3 for the definition and details. We show that the eigenvalue centrality can serve as a new and useful measure for ranking the volume and importance of exchanges because it is based on the cross-exchange Bitcoin flows on the blockchain, and therefore, is likely to be more resilient to manipulation than other measures.

transparency and traceability of transactions, and especially the enforcement of Know-Your-Customer (KYC) norms, across the network. The current regulatory efforts focus on creating greater transparency through enforcement of KYC norms and capital gains tax reporting at the level of individual institutions, such as exchanges or payment processors. However, if users of Bitcoin can freely trade across regulated and unregulated exchanges or even countries with different enforcement levels, effective KYC regulation might not be possible at the level of individual institutions.

We use the example of Hydra Market, which is one of the largest dark net marketplaces, to study flows in this market. Our analysis shows that the highest volume entities interacting directly with Hydra Market users are non-KYC exchanges, including Binance and Huobi which are two of the largest exchanges worldwide. Once the flows arrive at these exchanges, they get mixed with other flows and become virtually untraceable, and so can be sent anywhere afterwards, even to exchanges that enforce KYC norms. In contrast, the direct interaction of KYC exchanges, such as Coinbase or Gemini, with Hydra Market users is modest. But their indirect interaction with flows originating from Hydra market is significantly larger, since these flows are channeled through a network of short-lived clusters, solely created for the purpose of obfuscating the origin of these funds.

These results highlight that non-KYC entities serve as a gateway for money laundering and other gray activities. The decentralized nature of the Bitcoin protocol makes it easy for these entities to operate — they only need to have their servers in a country where the authorities are willing to tolerate their existence. If KYC entities are allowed to accept flows from entities that are not following strict KYC norms (the current state), then the digital footprint has a very limited effect on preventing tainted flows from entering into wide circulation.

Even if KYC entities were restricted to deal exclusively with other KYC entities, preventing inflows of tainted funds would still be nearly impossible, unless one was willing to put severe restrictions on who can transact with whom and make every transaction subject to the approval of a blockchain "monitoring entity", e.g. similar to what companies like Bitfury Crystal Blockchain[3] or Chainalysis are providing. Note that if this regime was to realize, the blockchain monitoring entities would become de facto trusted parties essential for the functioning of the Bitcoin network. But this is exactly what the Bitcoin protocol aims to overcome.

Composition of Bitcoin Miners. In a second major piece of analysis, we study the concentration and regional composition of Bitcoin miners, which are responsible for processing and verifying Bitcoin transactions and maintaining the integrity of the

[3] https://crystalblockchain.com/

Bitcoin blockchain. 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. The possibility of such attacks creates systemic risks for financial stability and potentially even for national security if a large fraction of citizens uses Bitcoin as a store of value.

It is therefore important to understand how concentrated the mining capacity is. The previous literature has mainly 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.

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 miner depends on the ease with which miners can shift capacity across pools, which in turn depends on the underlying size distribution of the miners.

Unlike information about mining pools, which is commonly available, information about individual miners is not readily available. We identify individual miners by tracking the distribution of mining rewards from the largest 20 mining pools to the miners that work for them. Since each pool uses its own algorithm to distribute rewards, we build separate algorithms for each pool. To the best of our knowledge, this is the first study that accurately links miners to their mining pools.

We show that the Bitcoin mining capacity is highly concentrated and has been for the last five years. The top 10% of miners control 90% and just 0.1% (about 50 miners) control close to 50% of mining capacity. Furthermore, this concentration of mining capacity is counter cyclical and varies with the Bitcoin price. It decreases following sharp increases in the Bitcoin price and increases in periods when the price drops or. Thus, the risk of a 51% attack increases in times when the Bitcoin price drops precipitously or following the halving events.

In addition, we show that there is significant geographic clustering of miners. While it has been previously discussed that a large majority of mining pools are registered in China, this does not automatically mean that miners have to be located in China. So far, the main data about miners' location has come from the analysis of miners' IP addresses from a few select pools. 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.

Here, we utilize a new approach, which takes advantage of our ability to trace miners on the blockchain. Since we can trace miners' addresses and Bitcoin transactions, we can see at which exchanges they use to cash out their rewards. The idea is that miners in a particular region would most likely send their rewards to an exchange that is also in this region. Using our approach we show that starting in 2015 and until April 2020 a majority of mining capacity, between 60% to 80% is located in China, which is consistent with anecdotal evidence.

In order to verify the validity of our approach of identifying miner locations by looking at where miners cash out their Bitcoin rewards, we use a recent incidence in April 2021 in the Xinjiang province of China. After a devastating coal mining accident, the government shut down coal mining and electricity supply for the entire area. Many Chinese Bitcoin miners are located in this province due to the cheap supply of coal powered electricity. Of course, not all Chinese miners are located in this area and thus we do not use it as a test of the mining capacity in China. But the shutdown of electricity for more than two days allows us to identify a set of miners for which we can be sure that they are physically located in China since they had to stop their operations. Using this strategy, we confirm that these Chinese miners, indeed utilize the cashing out policies that we had conjectured.

Ownership concentration. Finally, we study the ownership and concentration of Bitcoin holdings. Since the inception of Bitcoin, there has been intense interest in the question of who are the largest owners of Bitcoin, and how much do 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 who is positioned to benefit most from any price appreciation that would happen if regulators allow a broader adoption of Bitcoin. Are these a select few investors or the general public?

Determining the concentration of ownership is more complicated than just tracking the holdings of the richest addresses, since many of the largest addresses belong to cold wallets of exchanges and online wallets, which hold Bitcoin on behalf of many investors. We develop a suite of algorithms based on graph analysis to classify addresses into those belonging to individual investors or those belonging to intermediaries.[4]

[4] See Section 5 for a detailed description of the identification.

We show that the balances held at intermediaries have been steadily increasing since 2014. By the end of 2020 it is equal to 5.5 million bitcoins, roughly one-third of Bitcoin in circulation. In contrast, individual investors collectively control 8.5 million bitcoins by the end of 2020. The individual holdings are still highly concentrated: the top 1000 investors control about 3 million BTC and the top 10,000 investors own around 5 million bitcoins.

The rest of the paper is structured to first discuss the data sources and the construction of the data set. The next section documents the evolution of volume to different participants on the blockchain, in particular, we develop algorithms to separate spurious volume from real volume and then map the network structure of participants. In the following section we analyze miners, their composition and geographic concentration. And finally we document the ownership concentration of Bitcoin participants.

1. Related Literature

Our paper contributes to a fast-growing literature on cryptocurrencies and blockchains. Raskin and Yermack (2016) and Härdle et al. (2020) provide a broad perspective on the economics of cryptocurrencies and the blockchain technology they are built upon. Budish (2018), Abadi and Brunnermeier (2018), and Biais et al. (2019) study consensus mechanisms and limitations of the proof-of-work protocol, the core innovation of this new technology.

Athey et al. (2016), Cong et al. (2020b), Pagnotta and Buraschi (2018), Sockin and Xiong (2020), and Han and Makarov (2021) develop different theoretical frameworks to study bitcoin adoption and bitcoin pricing and highlight that beliefs about adoption are central for Bitcoin pricing. Schilling and Uhlig (2019) propose a model, in which a cryptocurrency such as Bitcoin coexists and competes with a traditional government-issued fiat money.

A number of papers study the economics of Bitcoin mining. Prat and Walter (2021) examines the relationship between the Bitcoin price and the investment in hashing capacity. Easley et al. (2019) and Huberman et al. (2021) develop equilibrium models of Bitcoin mining fees. Cong et al. (2020a) propose a theory of mining pools and suggest that mining pools escalate miners' arms race and significantly increase the energy consumption of proof-of-work-based blockchains. Ferreira et al. (2019) model the joint behavior of miners, mining pools, and firms producing specialized mining equipment. We contribute to this literature by developing a suite of algorithms to identify individual miners on the blockchain. This data is the first to trace individual miners and allows us to study their concentration and regional composition.

Similar to our paper, Foley et al. (2019) use the Bitcoin blockchain data to examine the prevalence of illegal transactions on the Bitcoin blockchain. Wallet-level blockchain data are also used by Griffin and Shams (2020) to study whether tether issuance affects bitcoin prices. In comparison to the earlier literature, we develop a novel database that not only has a much more comprehensive classification of participants on the blockchain, but also eliminates spurious volume. This granular data allows us to attribute economically meaningful transactions more precisely and to provide a detailed analysis of the evolution of the Bitcoin market.

2. Data

All bitcoin transactions are recorded on a distributed public ledger, the so-called blockchain. Transactions are organized in blocks that are added to the ledger every 10 minutes on average. Each block contains a few thousand transactions. A typical Bitcoin transaction includes a list of senders and recipients represented by pseudonymous addresses, the number of bitcoins sent and received, and a time-stamp of the transaction.

We download the blockchain data using the open-source software of Bitcoin Core and use the BlockSci program to parse the raw data into individual transactions.[5] As of June 28, 2021, there have been 689,000 blocks of 652 million Bitcoin transactions and 896 million addresses organized in a blockchain database of more than 379 GB in size.

An address on the blockchain can be thought of as a bank account. Anyone can send bitcoins to any address. But to send bitcoins from a given address one needs to know a password associated with this address. Unlike bank accounts, Bitcoin addresses can be generated freely, so typically the same entity controls several addresses, and in some cases, even tens of millions of different addresses.

The Bitcoin community developed several heuristics to assign addresses to the same entity. As a starting point, we use the most conservative method to cluster addresses whereby all addresses that send bitcoins in any single transaction are deemed to belong to the same entity.[6] This heuristic is justified by the Bitcoin protocol that requires the party that signs a transaction to have control of all output addresses

[5] Bitcoin Core and BlockSci are available at `https://bitcoin.org/en/bitcoin-core/` and `https://github.com/citp/BlockSci`, respectively.

[6] See Ron and Shamir (2012) or Meiklejohn et al. (2013). Bitcoin mixing services, such as CoinJoin, let users mix their coins with other users, and are designed to confuse this heuristic. The BlockSci accounts for that and avoids CoinJoin transactions in its clustering algorithm. See BlockSci documentation for more details.

In practice, a user typically only needs to specify the destination addresses and the amounts to be transferred. A special piece of software, called a wallet, then decides which addresses to send bitcoins from to cover a given amount that the user wants to transfer. This process then allows the clustering algorithm to successfully group all user's addresses together. It should be stressed however, that with a little bit of effort, a user can deliberately conceal the connections between his different addresses by making sure that no two addresses are ever used in the same transaction. As a result, this clustering heuristics only produces a lower bound for the true number of distinct entities.

To link address clusters to real entities we scrape cryptocurrency blogs and websites, such as Reddit, Blockchain.info, bitinfocharts.com, bitcointalk.org, walletexplorer.com, and Matbea.com for all publicly available addresses of prominent Bitcoin entities such as exchanges, payment processors, gambling sites, and others. We supplement this information with the state-of-the-art database of crypto entities from Bitfury Crystal Blockchain. Bitfury Crystal Blockchain is one of the leading providers of anti-money-laundering tools and analytic solutions in the crypto space.

To the best of our knowledge, we have the most complete information about crypto entities that have been used in academic research up to this point. Our data cover 1,043 different entities. These include 393 exchanges, 86 gambling sites, 39 on-line wallets, 33 payment processors, 63 mining pools, 35 scammers, 227 ransomware attackers, 151 dark net market places and illegal services.

3. Bitcoin Blockchain Volume

3.1. Spurious Volume

The design of the Bitcoin blockchain and the preference of many of its users for anonymity creates a lot of spurious volume that is not tied to economically meaningful transactions. In this section, we describe how we identify and separate this volume from the real volume, i.e. payments for goods and services and other financial transfers between two parties.

It is instructive to start by looking at a particular example, see the transaction depicted in Figure 1.[7] In this transaction, the address "17A16Q..." sends its balance to the following three addresses "3QKAn2...", "1F8fDp...", and "17A16Q...". The amount received is equal to the amount sent except for a small fee of 0.001 bitcoins, which is

[7] This is the second transaction in block 600,000 and can be seen e.g., at https://explorer.btc.com/btc/block/600000.

a part of the block reward. Notice that the last of the three addresses is the same as the sending address, that is, the address "17A16Q..." sends the majority of its balance to itself. This means the overall volume this transaction generates on the blockchain is large. However, the economically meaningful volume generated in the transaction (the real volume), which is the volume between different entities, is small.

The above situation where an address sends its balance to itself or to another address controlled by the same entity is very common. In part, it is a consequence of the design of the Bitcoin protocol. The outstanding balance of an address is not stored in the address but is imputed from the whole history of transactions involving this address by traversing back the Bitcoin ledger. For computational efficiency, the Bitcoin protocol allows one to send only the amounts that have been previously received by an address. For example, suppose an address previously received 5, 7, and 10 bitcoins, so the outstanding balance is 22 bitcoins. To send 8 bitcoins from this address one can either send 10 bitcoins, or any of the following linear combinations: 5+7, 5+10, 7+10, 5+7+10. Since in any case, the amount is larger than 8 bitcoins the sender needs to collect the difference using one of his addresses. This process creates a large amount of spurious volume that obscures the true volume of transactions on the blockchain.

Another common reason for spurious volume is the preference of blockchain participants for anonymity. Because the bitcoin blockchain is a public ledger all payment flows between addresses are perfectly observable. Many Bitcoin users, therefore, adopt strategies designed to impede the tracing of bitcoin flows.

Consider, for example, a situation where a hacker demands payment from a company to be sent to a Bitcoin address he controls. Since the ransom address is public information, if the hacker later sends bitcoins from this address to a third party, the party could easily flag funds as coming from illegal activity. To prevent this from happening, hackers often try to obfuscate tracing by creating multiple addresses and splitting the initial payment among them. This process is usually repeated many times resulting in the so-called "peeling chains", where funds travel a long distance from one address to another leading to a large amount of fictitious volume on the ledger.

Peeling chains are also commonly used by many exchanges, such as Coinbase and Kraken, and many mining pools. These entities, every time they need to collect a change as in the transaction in Figure 1, generate a new address instead of re-using the old address. This new address is then used to send funds to another entity, and the change is collected in another new address. This process is usually repeated many times until all initial balance is spent. The addresses used in peeling chains are usually used only to receive and immediately send bitcoins with a typical lifetime span of 10 hours.

[Fig. 1 About Here]

There are two ways how one can account for peeling chain transactions. First, one could modify the clustering algorithm to add addresses in peeling chains to the corresponding clusters. The other approach, which we follow in this paper, is to backtrack volume in peeling chains to the original clusters and discard any intermediate addresses from further analysis. To backtrack this volume we develop an efficient recursive algorithm detailed in the Appendix.

Factoring out peeling chains reduces the computational burden and results in significant reduction of addresses and clusters. While the original database has 896 million addresses, after we remove addresses in peeling chains we end up with 640 million addresses. Theses addresses belong to 189 million clusters, of which 116 million clusters are single-address clusters.

Figure 2 shows the decomposition of total Bitcoin blockchain volume into what we call internal, pass-through, and real volume. Internal volume is the within-cluster volume, that is, the volume that is generated when a cluster sends bitcoins to itself. The pass-through volume is the transitory volume associated with peeling chains. Finally, the real volume is the remaining volume, which represents transfers between clusters. This volume accounts only for 10% of the total Bitcoin volume on the blockchain, with 90% of the Bitcoin volume on the blockchain not tied to economically meaningful transactions.

[Fig. 2 About Here]

3.2. Real volume

We now focus on the economically meaningful, non-spurious, part of Bitcoin volume. To understand for what purposes Bitcoin is utilized, we trace Bitcoin flows between different types of entities on the blockchain. Our list of known entities includes exchanges, on-line wallets, payment processors, gambling sites, mixing services, illegal services, and mining pools. We identify these entities from a large number of public and proprietary sources as described in the data section.

Cryptocurrency exchanges such as Coinbase, Binance, or Kraken, and on-line wallets such as Blockchain.info and BixIn are one of the major types of entities where Bitcoin can be stored and traded. Exchanges in theory provide platforms to trade Bitcoin against fiat currencies and other coins, while on-line wallets specialize in custodian services. However, in practice, the difference between exchanges and on-line wallets is often slim. Both types of entities in many cases offer both functions. Therefore, we group these entities together when providing a general overview of Bitcoin utilization.

Payment processors, such as BitPay or CoinPayments, facilitate payments by online shops, gambling, and other entities that accept cryptocurrencies as means of payment for good and services. Illegal services include dark net marketplaces such as Hydra Market, numerous ransomware wallets, and entities engaged in scams. Mixing services or tumblers such as Bitcoin Fog and Wasabi wallet are sites that allow their customers to pool together their funds in order to obfuscate where the coins are being sent from.

Another set of entities that are a core component of the Bitcoin system are mining pools and miners. We identify miners by tracing rewards distribution of the largest mining pools to individual miners. We describe how we trace miners in Section 4 and in the Appendix. Overall, we identify 248,000 miners in the data.

As previously discussed, the pseudonymous nature of Bitcoin makes it difficult to link an address to the real-world entity behind them. Thus the identification of entities is incomplete almost by design, since it relies on an entity either voluntarily disclosing its addresses or learning about an entity's addresses in the course of interaction with it.

To address the problem of incomplete identification of entities and to make sure that we are not missing major players on the blockchain, we analyze the top 10,000 unknown clusters with the largest Bitcoin volume, for which we were not able to find an identity. Out of this universe of clusters, we select those that either receive regular flows from miners or receive more than 50% of its inflow from known exchanges and send more than 40% of its outflow to known exchanges. These thresholds are determined from the transaction patterns of known entities. For a typical exchange, 53% of its Bitcoin outflow goes to other exchanges, and 52% of its Bitcoin inflow comes from other exchanges. These numbers are significantly lower for all other entities. For example, a typical gambling site sends 21% and receives 29% of its total flows from exchanges. We find that 4507 clusters satisfy the above conditions. Taken together they account for 63% of the bitcoins flowing to the largest 10,000 clusters. In what follows, we refer to these clusters as LEOTD, Likely Exchanges, OTC brokers, or Trading Desks.

Based on this classification of participants, in Figure 3 we plot the average monthly transaction volume that is generated by these different types of entities on the blockchain from the beginning of 2015 until May 2021. The volume is calculated as the amount of bitcoins that are sent to different types of entities in a given month. Figure A shows the volume in BTC and Figure B as the percentage of the total monthly volume.

We see that the majority of the volume is generated by transactions involving exchanges and LEOTD clusters. Volume flowing to known exchanges constitutes about 40% of total volume and another 20% of the volume is generated by volume flowing to

LEOTD. To highlight the dominant role of exchanges and LEOTDs, we split volume that goes to the Other category, which consists of all unknown clusters that are not LEOTDs, into two parts: volume coming from exchanges and LEOTD and the rest. This decomposition shows that volume from exchanges and LEOTD to Other explains another 20% of the volume. Thus, exchange and trading desk related volume constitutes about 80% of the total volume. Other known entities are only responsible for a minor part of total volume as of the end of 2020. For example, illegal transactions, scams, and gambling together make up less than 3% of the volume. The fraction of volume explained by miners is even smaller.

[Fig. 3 About Here]

This analysis of volume underscores the dominance of trading and speculation related transactions on the blockchain, and at first glance seems to be at odds with earlier results that emphasized the prevalence of illegal transactions on the blockchain. Most notably, Foley et al. (2019) estimates that more than 46% of transactions are due to illegal transactions. The difference between their calculations and ours comes from two main sources.

First, Foley et al. (2019) intentionally drop all exchange-related volumes from their calculations, since they want to focus only on payments for goods and services. Since we show above that trading constitutes the main activity on the blockchain, this choice severely changes the denominator. Second, the estimate of volume in Foley et al. (2019) is based on an imputed network of illegal clusters where any cluster recursively is deemed illegal if the majority of its transactions is with previously identified illegal clusters. While intuitively appealing, this imputation method does not discriminate between real users and short-lived pass-through clusters that exist solely to obfuscate tracing. We show in Section 3.5 that this type of spurious volume is typically a very large part of illegal transactions. As a result, volume imputed by this method is likely to overstate the economic value of illegal trades.[8]

Our results of course do not mean that illegal activities on the Bitcoin blockchain are not a problem from the perspective of social welfare. We agree with the general concern that the pseudonymous nature of Bitcoin facilitates malfeasance such as illegal activities, tax evasion, or even bribes. Even though the BTC volume of illegal trades has stayed relatively stable in the last few years, the dollar amount of illegal activities increased, since the dollar value of BTC went up. We compute the net flow of bitcoins to illegal entities over 2020, broken down by their specific types. We calculate that

[8] The exact comparison of our results to the prior paper is difficult because we use a substantially larger set of identified entities.

there are about $550 million flowing to addresses that have been identified as scams, about $16 million in identified ransom payments, and more than $1.6 billion for dark net payments and dark net services. In addition, there are about $1.7 billion flowing to addresses affiliated with gambling and another $1.4 billion in mixing services.

In sum, we think it is important to get the magnitudes of transaction activities right in order to understand what are the ultimate drivers of Bitcoin value. Our results do not support the idea that the high valuation of cryptocurrencies is based on the demand from illegal transactions. Instead, they suggest that the majority of Bitcoin transactions is linked to speculation.

3.3. Network centrality

In the previous section, we show that cryptocurrency exchanges are responsible for the majority of volume on the Bitcoin network, and are therefore likely to play a dominant role in the network. To sharpen our understanding of the role exchanges play, we now analyze the structure of the Bitcoin network. In our network analysis, we restrict our attention to the most relevant clusters, i.e. clusters for which we know their identity and that are in the top 10,000 highest volume clusters. With these filters, we have 11,043 entities, which account for more than 55% of the total volume.

Because of the rapidly changing evolution of the Bitcoin ecosystem, we focus on the most recent time period: from 2018 to the end of 2020, which leaves us with 6248 entities. To represent this network, we use a directed weighted network graph, where a node i corresponds to cluster i and an edge (or link) from node i to j corresponds to the total Bitcoin flows over the period 2018-2020 from cluster i to cluster j.

The resulting network consists of 6248 nodes and 622K edges. Each entity receives and sends bitcoins to the other 100 entities in the graph, on average. Figure 4 plots a subset of this Bitcoin network graph, where for ease of illustration we retain only nodes that received at least 500,000 bitcoins over the period from 2018 to the end of 2020.[9]

[Fig. 4 About Here]

The network of the largest entities consists of 23 entities and 492 edges.[10] The node and edge size are proportional to the volume received by the entity and the volume between two different entities. In the case when two clusters send flows to each other,

[9] To plot this and other networks in this paper we use Graphia package software available at https://graphia.app/, Freeman et al. (2020).

[10] In a directed network, an edge from node i to j and an edge from node j to i count as separate edges.

the direction of the edge between these clusters agrees with the largest flow, and the edge is marked with a red segment. Out of 23 entities, three entities (BitGo, Xapo, and BixIn) are on-line wallets, 18 are identified exchanges, and two are unknown entities. The two unknown entities are likely to be unidentified exchanges or large OTC desks. They actively interact with known exchanges and receive a large amount of miners' rewards; 1252 and 4795 bitcoins, respectively.

Figure 4 reveals a high degree of interconnectedness between the major exchanges. We can see that they form an almost complete graph, where each node connects to all others. This is despite the fact that these exchanges operate in different regions. For example, Bithumb and Upbit are Korean exchanges, bitFlyer is Japanese, Bitstamp, Coinbase, Gemini, and Kraken are geared towards US and European users, and Huobi, BixIn, OKEx, and OceanEx towards Chinese. The high degree of interconnectedness has important implications for KYC regulation which we address in Section 3.5.

Inspection of Figure 4 further shows that Binance, Huobi, and Coinbase are the largest and the most active participants in the Bitcoin network. To formally quantify the importance of different entities, we compute the eigenvalue centrality of each entity in the full network. The eigenvalue centrality for an entity i is the i^{th} component of vector x, which is the solution to the eigenvector equation:

$$Ax = \lambda x, \tag{1}$$

where matrix elements A_{ij} are given by the total Bitcoin flows from entity i to j over 2018-2020, and λ is the largest eigenvalue associated with the eigenvector of matrix A. The eigenvector centrality takes into account not only the total volume received by an entity but also the structure of the Bitcoin network and gives larger weights to clusters that receive large volume from clusters that receive large volume themselves.[11]

Figure 5 shows the top 25 entities with the largest Bitcoin network centrality. Confirming our earlier observation based on Figure 4, Binance, Coinbase, and Huobi have the highest measure of centrality. Other exchanges from Figure 4 are also among the most central 25 entities. This should not come as surprise since all these exchanges are part of a very dense network.

[Fig. 5 About Here]

The eigenvalue centrality (1) confirms the dominant role of exchanges in the Bitcoin network. We conclude this section by noticing that the eigenvalue centrality can

[11] See Newman (2010), 7.1.2 for more details. Eigenvector centrality based on other network measures such the total number of transactions produces similar results.

potentially serve as a new and useful measure of the importance of exchanges. Other popular measures that have previously been used by many data aggregators such as CoinMarketCap include (1) off-chain exchange trading volume, website traffic, or the number of Twitter followers. A desirable property for any ranking measure is to be resilient to manipulation. Unfortunately, none of the existing measures seem to be fully manipulation-proof.[12]

Since the eigenvalue centrality measure is based on the cross-exchange bitcoin flows on the blockchain, to improve its position in the ranking an exchange would have to send back and forth large amounts of bitcoins to other exchanges. This can prove significantly more costly than simply engaging in wash trading or buying website traffic. Therefore it is reasonable to believe that the eigenvalue centrality can be more resilient to manipulation than other measures.

3.4. Cross-exchange flows

Our analysis in Sections 3.2 and 3.3 shows that a large part of the Bitcoin volume is driven by cross-exchange flows. What explains these flows? To answer this question, it is important to recognize that cryptocurrency markets consist of many non-integrated exchanges that are independently owned and exist in parallel within and across countries. On an individual basis, the majority of these exchanges function like traditional equity markets where traders submit buy and sell orders, and the exchange clears trades based on a centralized order book. However, in contrast to traditional, regulated equity markets, the cryptocurrency market lacks any provisions to ensure that investors receive the best price when executing trades.[13] The absence of such mechanisms increases the importance of arbitrageurs who trade across different exchanges and ensure consistent prices across them.

Suppose an exchange rate between Bitcoin and some other currency, say C, is different across two exchanges. An ideal arbitrage trade would be to exchange Bitcoin for C on the exchange where the exchange rate is high and exchange C for Bitcoin on the exchange with a low exchange rate; then transfer Bitcoin and C between exchanges and realize the risk free profit.

The above trade faces few obstacles if C is a cryptocurrency since by design, the pseudonymous nature of cryptocurrencies makes them immune to any capital controls. However, when C is a fiat currency the ability to repatriate funds from one country

[12] See, for example, a report from cryptocurrency market surveillance firm BTI Verified: https://btiverified.com/crypto-market-data-report-2020/.

[13] For example, the US Securities and Exchange Commission (SEC)'s National Best Bid and Offer (NBBO) regulation in the United States requires brokers to execute customer trades at the best available prices across multiple exchanges.

to another may be obstructed by cross-border capital controls, and the market can become potentially segmented.

In Makarov and Schoar (2020) we indeed show that in the period 2017-2018 there were large and recurring deviations in cryptocurrency prices across exchanges, pointing to significant market segmentation. We also showed that the arbitrage spreads were much larger for exchanges across different countries than within the same country, and were positively linked to cross-border capital controls.[14]

The above discussion suggests that (1) exchanges within a country with strong capital controls can be more interconnected with each other than with other exchanges, and (2) exchanges that trade similar currency pairs can see higher cross-exchange flows.

To test these predictions, we compute two measures of similarity of a pair of exchanges. One is based on the similarity of the cryptocurrency pairs traded on each exchange. And the other is based on the similarity of the interaction of the two exchanges with other exchanges on the Bitcoin blockchain.

To compute the currency-pair similarity we use Kaiko data, a private firm that has been collecting trading information about cryptocurrencies since 2014. The Kaiko data cover only a subset of exchanges that we can identify on the Bitcoin blockchain, but these are the largest exchanges. The joint set consists of 57 exchanges.

For each exchange, we consider all traded currency pairs where one of the currencies is Bitcoin. The other currency could be a fiat currency, stable coins, or other cryptocurrencies. In total, we have 4,360 currency pairs across 57 exchanges, with the median number of currency pairs on an exchange being 13. For each exchange i and cryptocurrency pair j, we compute the total trading volume in the period 2018-2020 denominated in Bitcoin, v_{ij}. Next, we normalize the volume on each exchange by the Euclidean norm

$$\hat{v}_{ij} = \frac{v_{ij}}{\sqrt{\sum_j v_{ij}^2}},$$

and use the Euclidean distance between vectors $\mathbf{v_i} = \{\hat{v}_{ij}\}_{j=1}^N$ and $\mathbf{v_k} = \{\hat{v}_{kj}\}_{j=1}^N$ as a measure of similarity of exchanges i and k.

To compute the exchange similarity based on the Bitcoin flows we first calculate the matrix of cross-exchange flows, A. Each element a_{ij} of matrix A is the average of Bitcoin flows from exchange i to exchange j and vice versa. As before, we normalize

[14] Cross-border capital controls can be a motif for cross-exchange flows themselves. Since Bitcoin is not subject to capital controls one can use it as a means to bypass them. This alone, however, cannot explain flows between crypto-only exchanges and flows across exchanges within the same country.

the volume on each exchange by the Euclidean norm

$$\hat{a}_{ij} = \frac{a_{ij}}{\sqrt{\sum_j a_{ij}^2}},$$

and use the Euclidean distance between vectors $\mathbf{a_i} = \{\hat{a}_{ij}\}_{j=1}^N$ and $\mathbf{a_k} = \{\hat{a}_{kj}\}_{j=1}^N$ to obtain a measure of similarity of exchanges i and k on the Bitcoin blockchain.

To see if the two constructed similarity measures isolate the same group of exchanges we apply the K-medoids clustering algorithms to each of the similarity measures. The K-Medoids algorithms tries to group similar exchanges together to minimize the within-cluster sum of distances between exchanges. It is a popular clustering method available in many packages.[15]

Figure 6 shows the result of the application of K-medoids clustering algorithms based on the currency-pair similarity measure. We can see that there are four notable groups of exchanges. These are US-European (group 5), Korean (group 3), Japanese group (4), and Tether (group 2) exchanges. The sharp clustering of Korean and Japanese exchanges reflects the fact that these exchanges use the national fiat currency as the base currency and trade a small number of currency pairs. A similar situation holds for US-European exchanges where both dollar and euro serve as a base currency, with the dollar usually being more popular. The majority of group 2 exchanges are crypto-only exchanges, which do not offer an opportunity to trade against a fiat currency. These exchanges usually use Tether as a base currency and list a large number of different cryptos for trading. There are also a few isolated exchanges that have less popular base currencies. For example, Coinfloor uses British pound, BitBay Polish zloty, and ACX Australian dollar.

[Fig. 6 About Here]

Next, we apply the K-medoids clustering algorithms to the Bitcoin blockchain similarity measure. Figure 7 shows the results.[16] Comparing Figures 6 and 7, we can see that the difference in distance between exchanges within a cluster and exchanges from different clusters is not as pronounced as in the case of clustering based on the cryptocurrency pairs. This should not come as a surprise since exchange integration depends not only on the cryptocurrencies traded but also on the capital controls that are in place in a given country. Two exchanges in different countries can be well separated in the cryptocurrency pair distance if they use different fiat currencies but can

[15] We use KMedoids routine from Python module sklearn_extra in our analysis, see Hastie et al. (2001), 14.3.10 for more details.

[16] K-medoids clustering algorithms takes the number of clusters as a parameter. Since there is no rigorous theory to determine it, we use a default value of 8.

be very similar in the Bitcoin blockchain distance if they operate in countries without capital controls.

[Fig. 7 About Here]

Nevertheless, Figures 6 and 7 show that clustering of major groups of exchanges based on the two similarity metrics produces broadly similar results. In particular, the Korean and Japanese exchanges in both cases are grouped together. The US-European and Tether exchanges are less clearly separated. For example, Poloniex, which is crypto-only exchange, is now grouped together with Coinbase, Bitstamp, Gemini, Kraken, and Bitfinex. This is again consistent with our results in Makarov and Schoar (2020), where we show that the US-European and Tether exchanges are better integrated than exchanges in Korea and Japan.

3.5. Enforcement of KYC norms for Bitcoin Transactions

We conclude our study of Bitcoin volume with the analysis of flows associated with the shadow economy. Light regulation and the anonymity of cryptocurrencies have made them a popular choice for anyone who wants to evade legal or regulatory scrutiny or engage in tax evasion. Proponents of cryptocurrencies often like to point out that cryptocurrencies are still superior to cash because of their digital footprint. While the digital footprint indeed imposes some constraints to the anonymity of transactions, and in some cases helped catch offenders, it is important to realize that there are strong limitations.

To understand the challenges of enforcing Know-Your-Customer (KYC) norms, it is instructive to consider a network centered on Hydra Market, which is one of the largest dark net marketplaces.[17] Hydra Market has been in operation since 2015 and has been growing rapidly since then. We focus on the most recent period, 2020 - June 2021. Over this period, Hydra market received 147,620 bitcoins from 514,855 clusters and sent them to 315,359 clusters. The 514855 sending clusters in turn received their flows from 3,291,180 clusters, and the 315,359 sending clusters sent to 500,544 clusters, of which 116,131 are new clusters.

Figure 8 depicts the resulting network, where for ease of illustration we retain only nodes that send at least 1000 bitcoins *within* this network. When computing volume in this network we intentionally exclude volume between any clusters that belong to the list of known or high volume entities, which we studied in Sections 3.3

[17] https://www.bloomberg.com/news/articles/2021-02-01/darknet-market-had-a-record-2020-led-by-russian-bazaar-hydra.

and 3.4. The node size reflects the total amount of bitcoins sent from Hydra Market to a corresponding entity. The edge size is proportional to the volume between two different entities. In the case when two clusters send flows to each other, the direction of the edge between these clusters agrees with the largest flow, and the edge is depicted with a red segment. The orange color shows identified clusters, the green color marks unknown high volume clusters, the turquoise color shows short-lived clusters with a lifespan below one month, the purple color marks the remaining clusters.

[Fig. 8 About Here]

Figure 8 reveals that the highest volume entities interacting directly with Hydra Market are non-KYC exchanges such as LocalBitcoins, Bitzlato, Binance, Huobi, and Totalcoin.[18] Once the flows arrive at these exchanges they get mixed with other flows and become virtually untraceable, and so can be sent anywhere afterwards.

The figure also shows that direct interactions of Hydra Market with those exchanges that try to enforce KYC norms, such as Coinbase and Gemini, are modest; but their interaction with the neighboring clusters is significantly larger. For example, Coinbase directly sent and received 196 and 126 bitcoins from Hydra Market, respectively. But it sent 530,000 and received 218,000 bitcoins via the neighboring clusters.

Looking at Figure 8 we can see that the majority of flows to and from Coinbase occur through short-lived clusters, which in most cases are created for the sole purpose of obfuscating the origin of funds. A typical transaction involves mixing tainted funds (those that can be traced to Hydra Market) with “clean” (not traceable to illegal transactions). Each mixing reduces the share of tainted flows. The process is repeated several times until the resulting flows become clean enough to send them to KYC exchanges.

What are the implications of the above analysis? First, non-KYC entities serve as a gateway for money laundering and other gray activities. The decentralized nature of the Bitcoin protocol makes it easy for these entities to operate — they only need to have their servers in a country where the authorities are willing to tolerate their existence. If KYC entities are allowed to accept flows from entities that are not following strict KYC norms (the current state) then the digital footprint has a very limited effect on preventing tainted flows from entering into wide circulation. The ability to trade “privacy coins” such as Monero and the increasing popularity of DeFi platforms further facilitate these money laundering strategies.

Second, even if KYC entities were restricted to deal exclusively with other KYC entities, preventing inflows of tainted funds would still be nearly impossible, unless one

[18] https://bitshills.com/best-non-kyc-crypto-exchanges/.

was willing to put severe restrictions on who can transact with whom and make every transaction subject to the approval of a type of blockchain analytics companies such as Bitfury Crystal Blockchain and Chainalysis. Note that if this regime was to realize these firms would become the de facto trusted parties essential for the functioning of the Bitcoin network. But this is exactly what the Bitcoin protocol is designed to circumvent. If trusted parties exist there are simpler and more efficient solutions than the Bitcoin protocol, e.g., a permissioned blockchain.

Finally, notice that while transacting in cash and storing cash involve substantial costs and operational risks, transacting in cryptocurrencies and storing them are essentially costless (apart from fluctuation in value). The wider the adoption of Bitcoin is, the easier it will be to use it for transactions without ever having to touch regulated entities, and the more attractive it will become for malfeasance and shadow economy.

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 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.

[Fig. 9 About Here]

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.[19] 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. 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.

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

[19] 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.

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.[20] In Figure 10 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.

[Fig. 10 About Here]

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. The figure shows

[20] 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.

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%.

[Fig. 11 About Here]

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.[21] 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.

[21] 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.

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 in the Appendix shows the map between exchanges and regions.

Using this proxy for miner location, Figure 12 Panels A and B 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.[22] 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

[22] 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.

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.

[Fig. 12 About Here]

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 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. 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.

[Fig. 13 About Here]

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

(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.[23] However, not all exchanges have a cold wallet that is as distinct as

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

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.

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 1MzG9Gx5G3ZTXtEQT4FJg23Cb3gS6UF982 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 unique parent cluster. For example, Figure 20 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.

[Fig. 20 About Here]

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 1013 as individual, 1154 as linked to intermediaries, and 47 as ambiguous.

Figure 21 shows the amount of Bitcoin held in the wallet of intermediaries over time. The balance held at intermediaries started accelerating in 2014 has been steadily increasing over time. By the end of 2020 is was equal to 5.5 million bitcoins, roughly one-third of Bitcoin in circulation at the time.

[Fig. 21 About Here]

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 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.

[Fig. 22 About Here]

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 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.

[Fig. 23 About Here]

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-Your-Customer (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 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.

References

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

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.

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

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

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

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

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

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.

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

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

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*.

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

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

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.

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

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*.

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

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.

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

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

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

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

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

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

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

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

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 B_i as volume from A to B_i, $i = 1, .., N$.

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 A_j, $j = 1,..,M$ to B_i, $i = 1,..,N$ as $\frac{w_j}{\sum_{k=1}^{M} w_k} \times v_i$, 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

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} p_{21} + w_1 - p_{12} \\ p_{12} + w_2 - p_{21} \end{pmatrix} = F \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \tag{3}$$

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

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

The matrix F^{-1} defines a map from $\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$ 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 $W_{ii} = w_i$, $i = 1, ..., n$. 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.[24] For each strongly connected 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.

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

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.

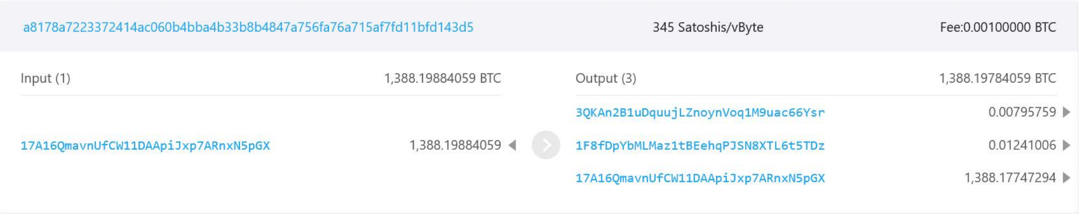
Finally, we allow for the possibility that some of the remaining addresses might not belong to individual miners but to smaller pools that do mining operations as part of a larger pool, or belong to a subsidiary or a partner of the larger pool. To screen out these addresses we check if

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.

Figures



a8178a7223372414ac060b4bba4b33b8b4847a756fa76a715af7fd11bfd143d5 | 345 Satoshis/vByte | Fee:0.00100000 BTC

Input (1)	1,388.19884059 BTC	Output (3)	1,388.19784059 BTC
		3QKAn2B1uDquujLZnoynVoq1M9uac66Ysr	0.00795759 ▶
17A16QmavnUfCW11DAApiJxp7ARnxN5pGX	1,388.19884059 ◀	1F8fDpYbMLMaz1tBEehqPJSN8XTL6t5TDz	0.01241006 ▶
		17A16QmavnUfCW11DAApiJxp7ARnxN5pGX	1,388.17747294 ▶

Figure 1: Bitcoin transactions and spurious volume. The figure shows an example of a typical transaction on the Bitcoin blockchain with large spurious volume. The address "17A16Q..." on the left of the ledger, the sender, sends its entire balance to three addresses. The last recipient address that received the majority of the bitcoins is the same as the sending address.

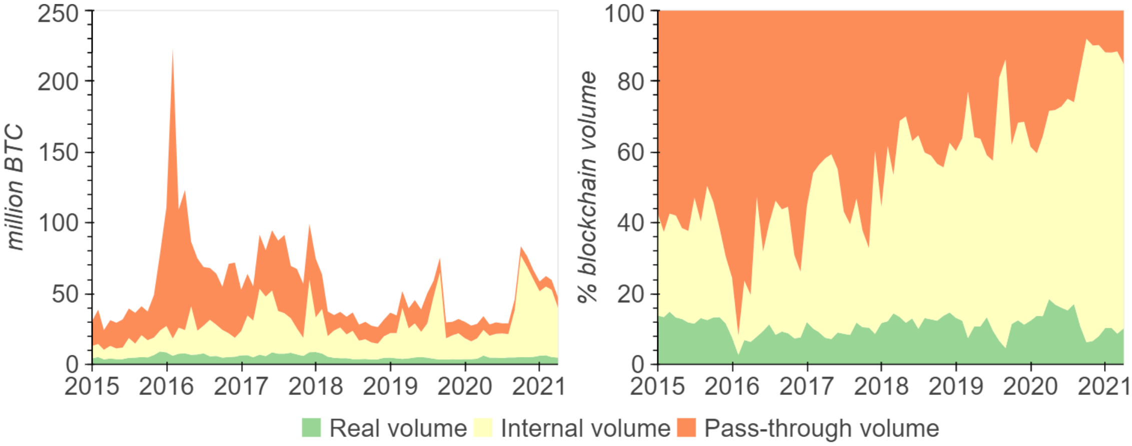


Figure 2: Decomposition of volume: Internal, pass-through, and real volume. The figure shows the decomposition of total Bitcoin blockchain volume at the monthly level into three components. The top (orange) part shows pass-through volume, which is created when users move their funds over long chains of multiple addresses and splitting payments among them to impede the tracing of flows, also called peeling chains. The next part (yellow) reflects the internal volume that is generated when a user (cluster) sends bitcoins to itself. Finally, the remaining part (green) is real volume, which represents transfers between clusters controlled by different users.

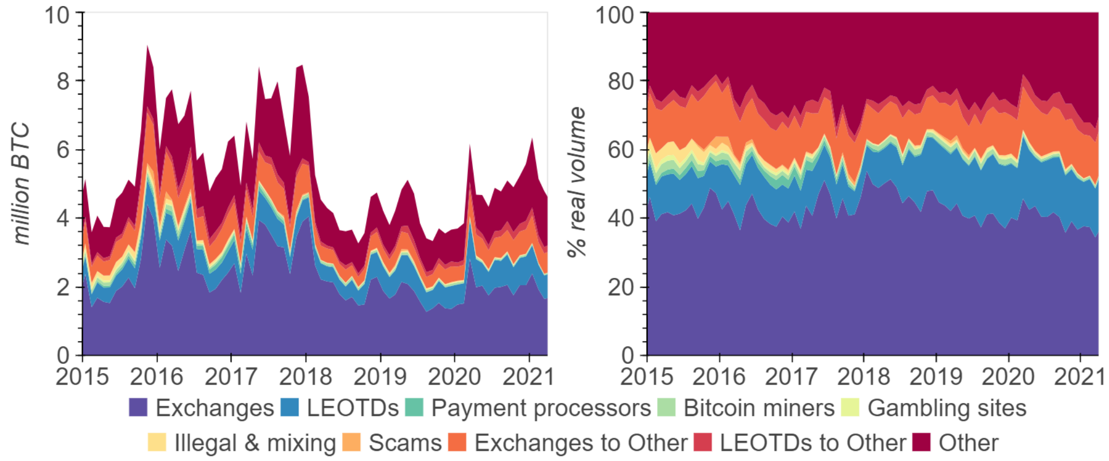


Figure 3: Decomposition of real volume. The figure shows the monthly volume generated by different entities on the blockchain, from January 2015 to May 2021. The volume is calculated as the amount of bitcoins that are sent to different types of entities in a given month. The panel on the left shows the volume in BTC and the panel on the right shows the volume as the percentage of the total monthly volume. LEOTDs are likely exchanges, OTC brokers and other trading desks. Other represents addresses that are unclassified. We break out volume to Other if it is generated by exchanges or LEOTDs. A detailed description of the classifications is provided in Appendix Table 2.

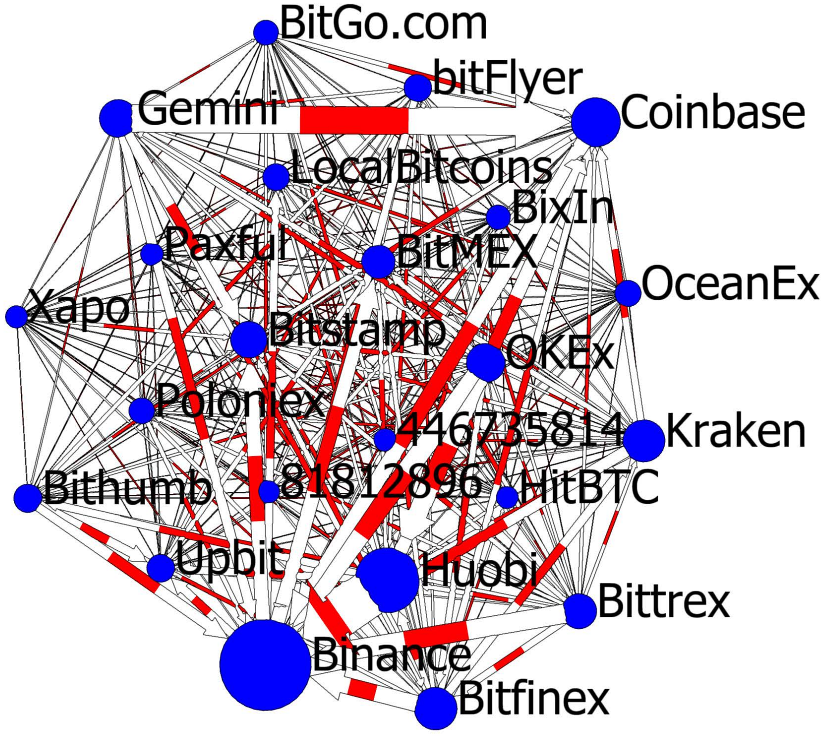


Figure 4: Bitcoin network. The figure shows the bitcoin network for large clusters that received at least 500,000 bitcoins from 2018 to the end of 2020. This is a directed weighted network graph, where a node i corresponds to cluster i and a link from i to j corresponds to the total Bitcoin flows over the period 2018-2020 from cluster i to j. The node and link sizes are proportional to the volume received by the entity and the volume between two different entities, respectively. In the case when two clusters send flows to each other, the direction of the link between these clusters agrees with the largest flow, and the link is marked with a red segment.

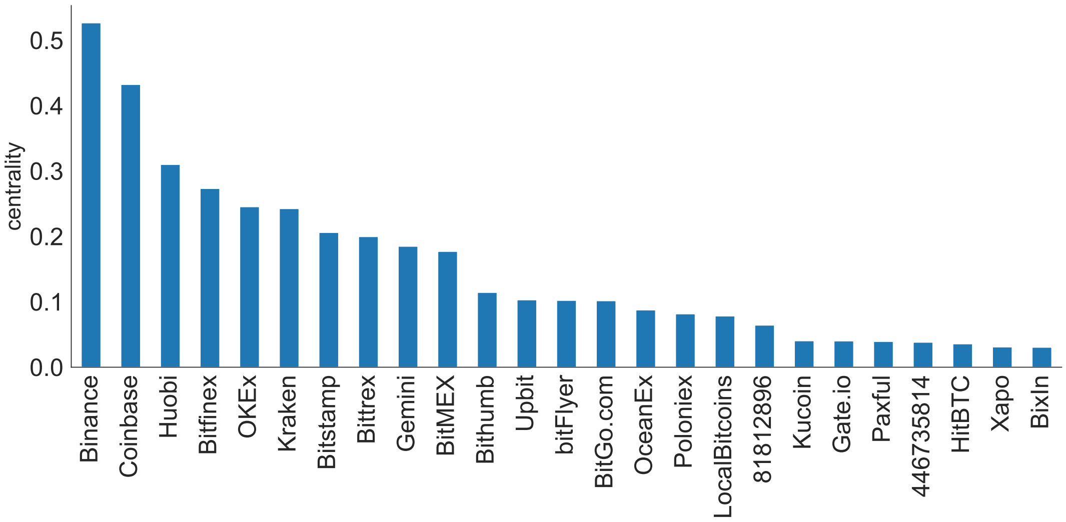


Figure 5: Entities with highest Bitcoin network centrality. The figure shows the top 25 entities with the largest network centrality in the Bitcoin volume network. Network centrality is defined as the eigenvalue centrality of each entity in the full network, which is the solution to the eigenvector equation: $Ax = \lambda x$, where matrix elements A_{ij} are given by the total Bitcoin flows from entity i to j over 2018-2020, and λ is the largest eigenvalue associated with the eigenvector of matrix A. Eigenvector centrality takes into account the total volume received by an entity and gives larger weights to clusters that receive large volume from clusters that receive large volume themselves.

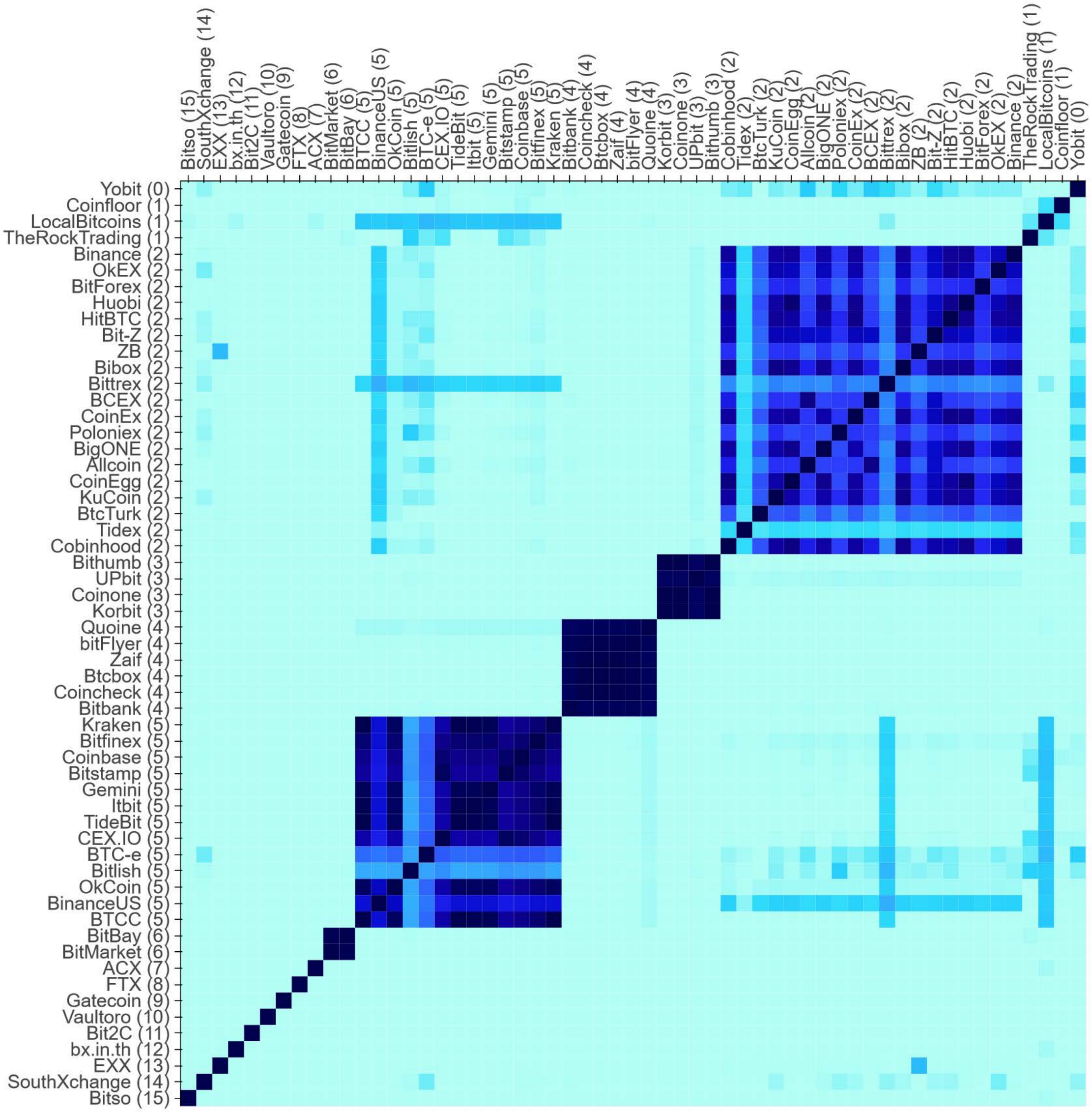


Figure 6: Currency-pair similarity between exchanges. The figure shows the similarity between exchanges based on the cryptocurrency pairs traded on each exchange, using data from Kaiko. We use all currency pairs where one of the currencies is Bitcoin, the other can be another coin or a fiat currency, for a total of 4,360 currency pairs across 57 exchanges. For each exchange and cryptocurrency pair we compute the total relative trading volume in 2018-2020 denominated in Bitcoin. We normalize the volume for cryptocurrency pairs where one of the cryptocurrencies is Bitcoin and compute the Euclidean distance between volume vectors. The graph shows the result of the application of K-medoids clustering algorithms based on the currency-pair similarity measure. The numbers in the parentheses following the name of each exchange stand for the group we assign it to, for example, US-European is group 5, Korean is group 3, Japan group 4 etc. Cells with darker colors indicate higher degrees of similarity.

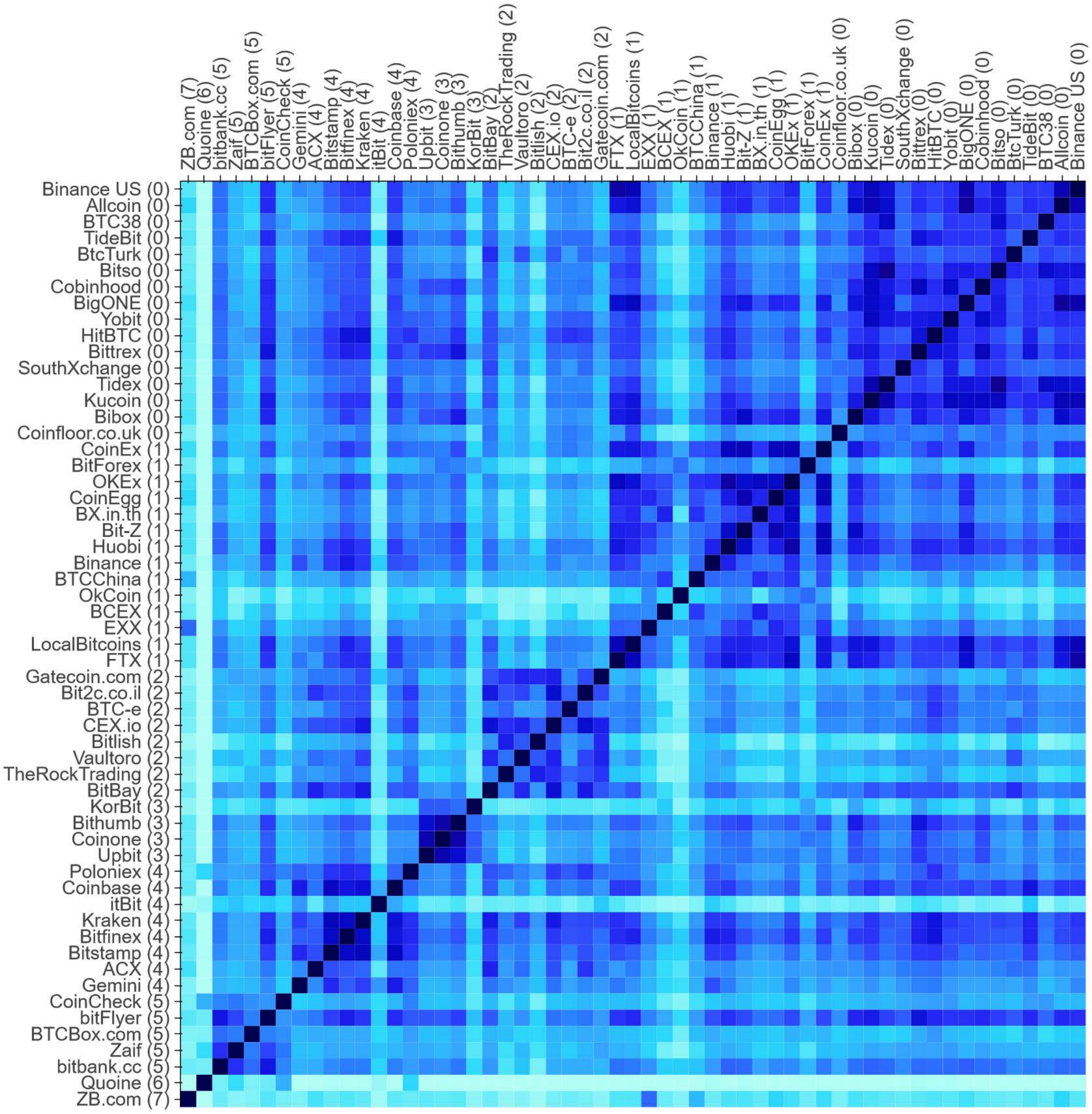


Figure 7: Blockchain volume similarity between exchanges. The figure shows the similarity between exchanges based on the Bitcoin flows on the blockchain between exchanges. For each exchange, we first calculate the matrix of cross-exchange flows and compute the Euclidean distance between volume vectors. The graph shows the result of the application of K-medoids clustering algorithms based on the cross-exchange flows similarity measure. The numbers in the parentheses following the name of each exchange stand for the group we assign it to, for example, US-European is group 5, Korean is group 3, Japan group 4 etc. Cells with darker colors indicate higher degrees of similarity.

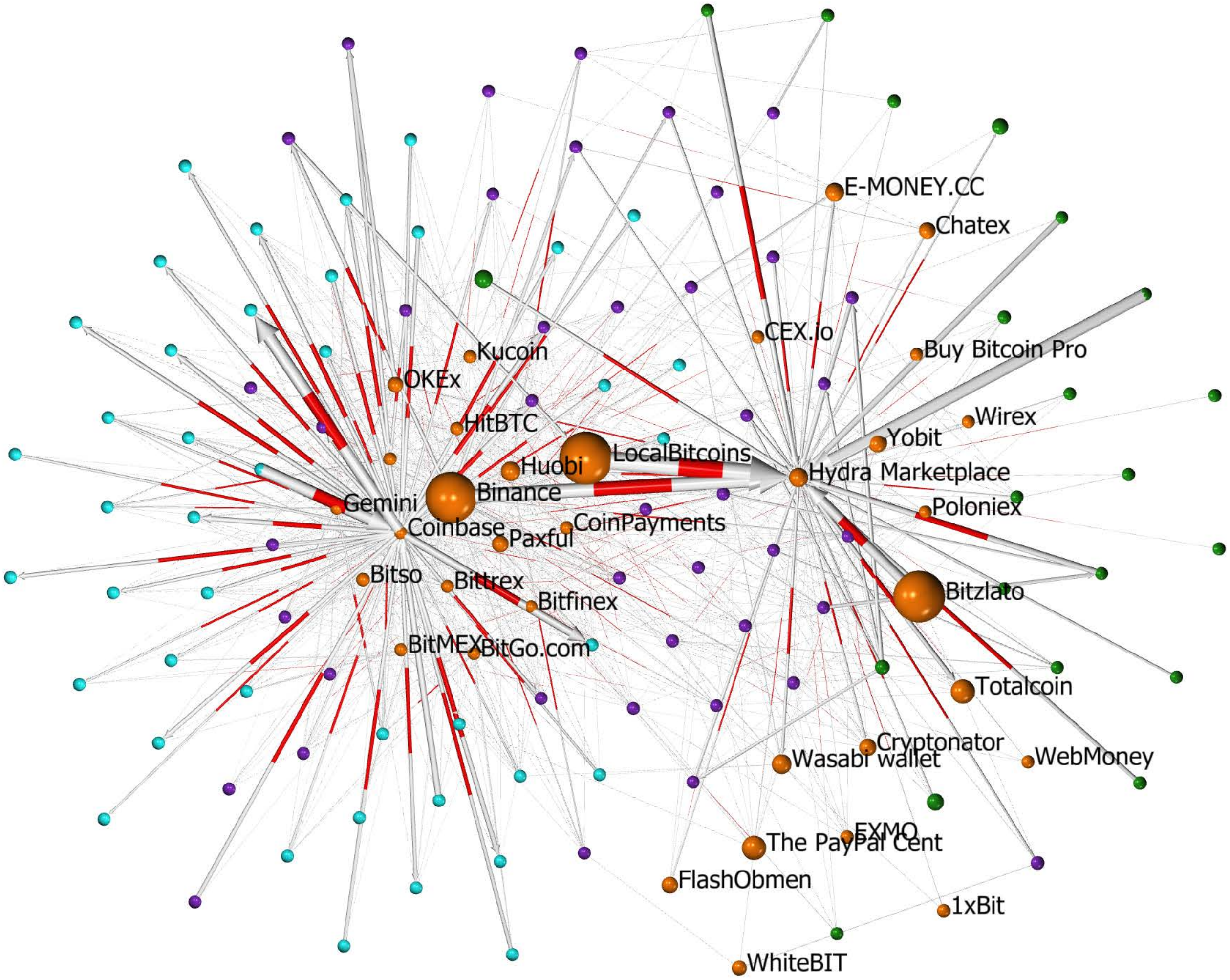


Figure 8: Hydra market network. The figure shows the different clusters of the Bitcoin network connected to Hydra market, one of the largest darkweb sites. We include only entities that send at least 1,000 bitcoins within the network over the 2020-2021 period. The node size reflects the total amount of bitcoins sent from Hydra Market to a corresponding entity. The link size is proportional to the volume between two different entities. In the case when two clusters send flows to each other, the direction of the link between clusters is based on the largest flow, and the link is depicted with a red segment. Orange nodes signify identified clusters, green nodes mark unknown high volume clusters, turquoise nodes depict short-lived clusters with a life-span below one month, and the remaining clusters are purple.

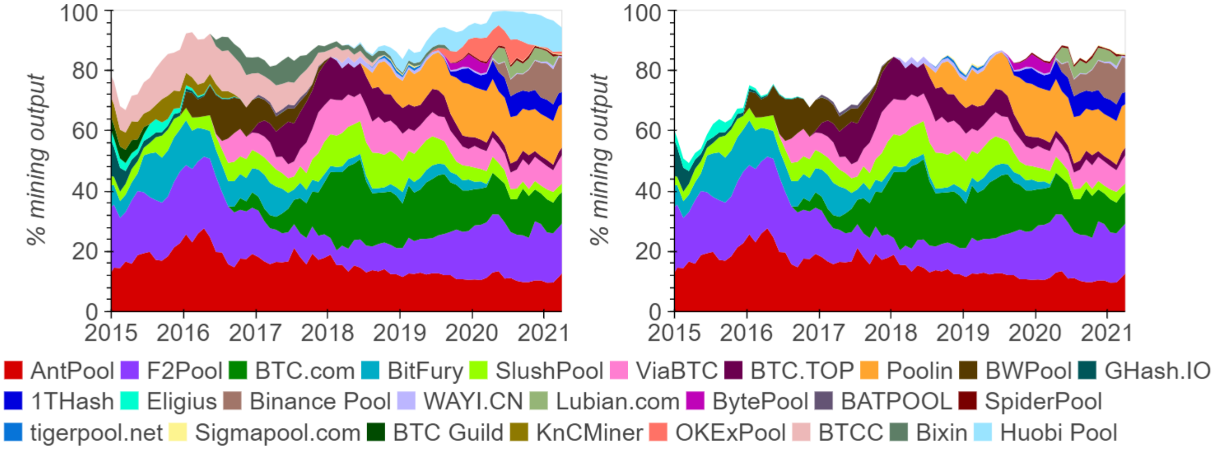


Figure 9: Coverage of mining pools. The figure shows the share of bitcoins mined by each pool in each month from 2015 to May 2021. The left panel shows the shares of all known pools. The right panel shows the subset of pools for which we trace the Coinbase rewards of the pool to individual miners that are working with the pools.

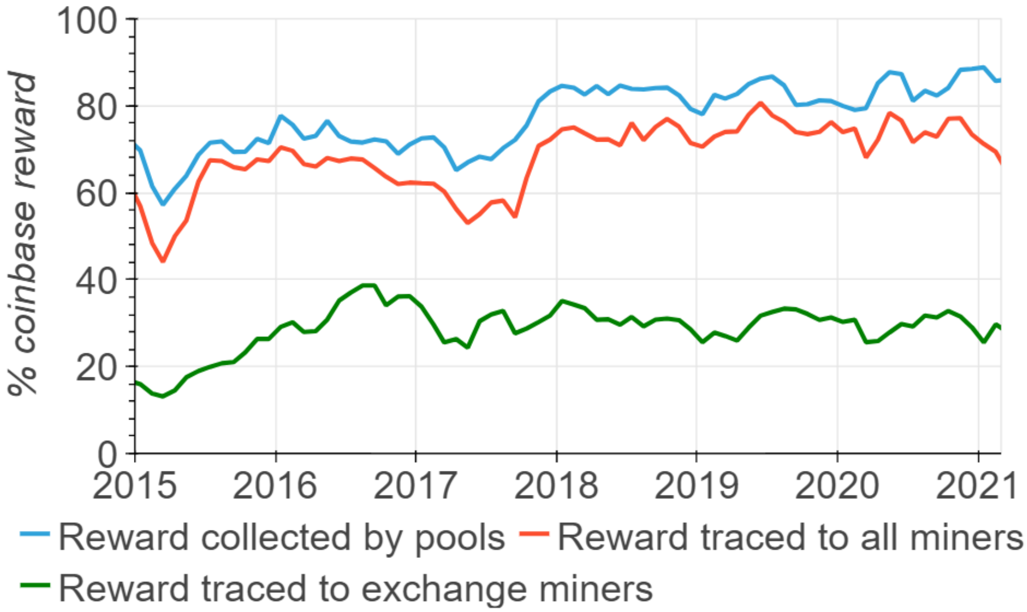


Figure 10: Coinbase reward traced to miners. The figure shows our coverage of the mining capacity on the Bitcoin blockchain on a monthly basis from 2015 until May 2021. The blue line shows the aggregate coinbase rewards reported by the large pools that we can trace at an aggregate level. These are 21 mining pools, including two private pools. The red line shows the distributed mining rewards that we can trace on the blockchain from the pool distribution addresses to the underlying miners. The green line shows the rewards collected by exchange-wallet miners; these are miners who collect their rewards with addresses that are linked to an exchange.

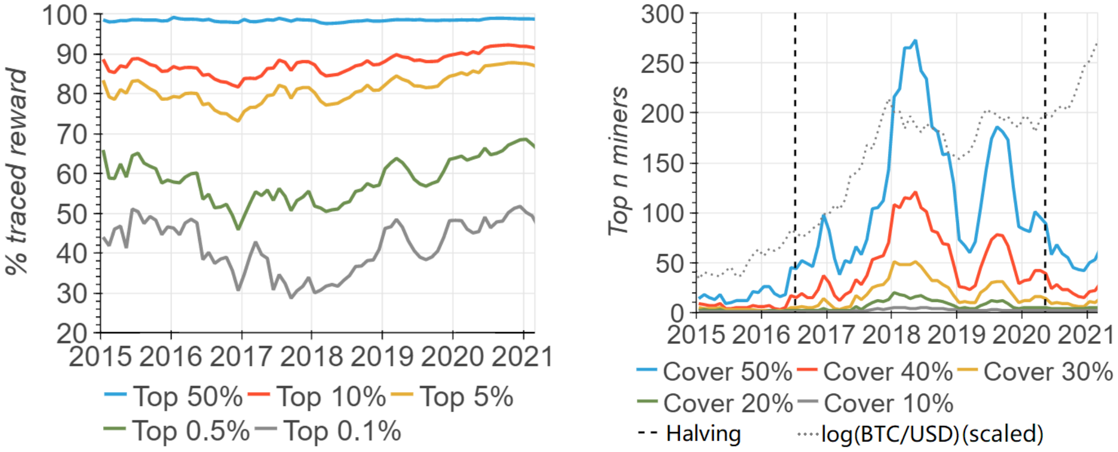


Figure 11: Concentration of mining capacity. This figure documents the concentration capacity of miners based on Coinbase rewards that miners receive from pools. Each month, we sort active miners by the amount of Coinbase rewards they receive and calculate the percentage of total mining capacity controlled by different quantiles of the miner distribution. The left panel shows the results for the top 50%, 10%, 5%, 0.5%, and 0.1% miners. The right panel shows the number of miners that are necessary to cover 10%, 20%, 30%, etc. of total mining capacity. The dashed lines indicate Bitcoin halving dates. The dotted line shows the log-price of one Bitcoin in USD, scaled to fit the plot.

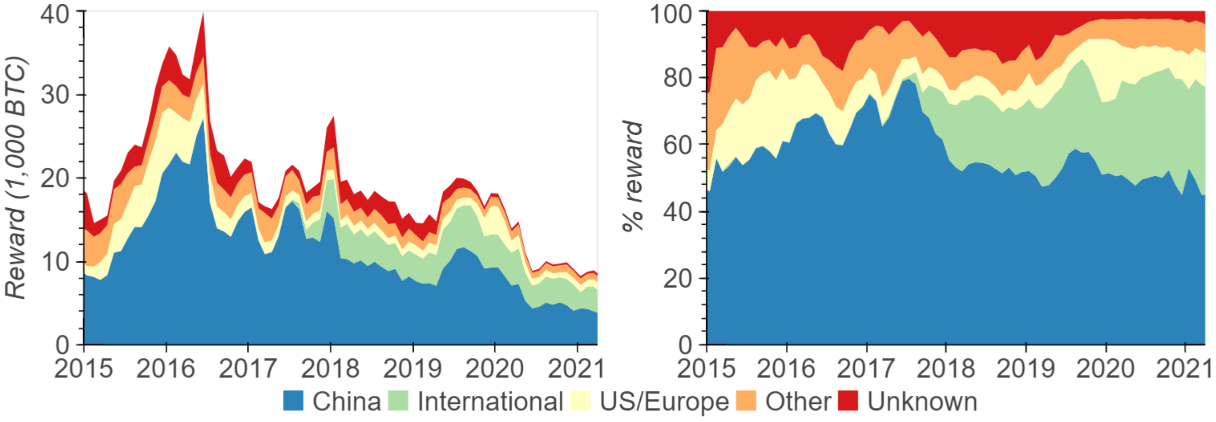


Figure 12: Geographic concentration of miners. The figure shows the distribution of the mining capacity across different regions. The geographic location of miners is based on the location of the exchanges where a given miner cashes out most of its Bitcoins. International includes exchanges that operate across many jurisdictions, and rely on stable coins like tether; these include exchanges such as Binance and Gate.io. The Other category includes all identified exchanges outside the above ones. The left panel plots the monthly value of Bitcoin rewards that are cashed out by miners in different regions; the right panel shows the percentages.

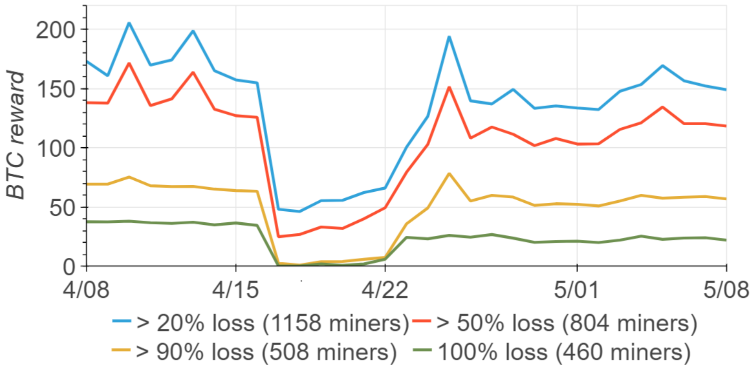


Figure 13: Impact of Xinjiang incident. The figure plots the time series of Coinbase rewards that are received by miners located in the Xinjiang province in China and were impacted by the coal mine incident of April 16, 2021. We focus on miners who received reward every day during April 8-15, 2021. The blue line indicates the total BTC rewards of miners that lost more than 20% of their hashing capacity during the weekend of April 17-18, 2021 compared to their daily capacity before the incident. The red line shows miners that lost more than 50 of their hashing capacity, the yellow lines miners lost more than 90% of their capacity, and the green line 100%.

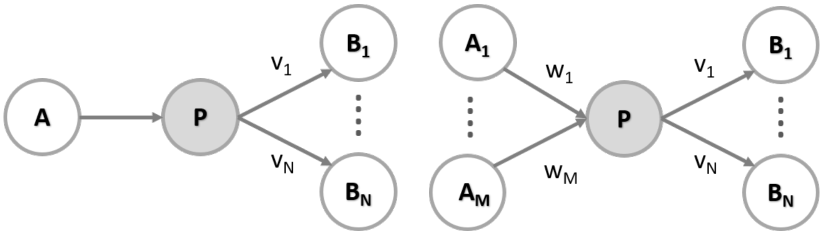


Figure 14: Short-term cluster types. The figure depicts two types of short-term clusters. The first-type shown in the left panel has a single incoming transaction and a single outgoing transaction. The second type (right panel) can have multiple incoming and outgoing transactions.

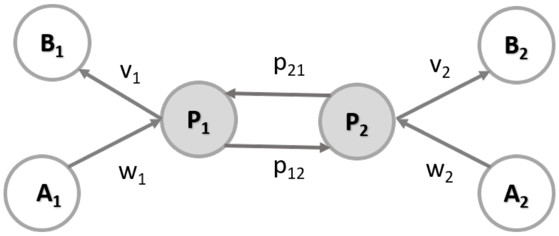
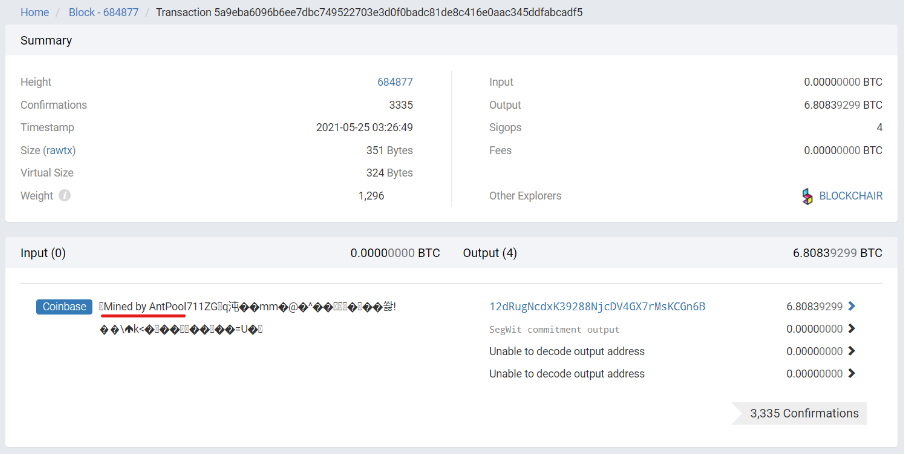


Figure 15: Short-term cluster cycle. The figure depicts a situation where two short-term clusters P_1 and P_2 send flows p_{12} and p_{21} to each other.



Home / Block - 684877 / Transaction 5a9eba6096b6ee7dbc749522703e3d0f0badc81de8c416e0aac345ddfabcadf5

Summary

Height	684877	Input	0.00000000 BTC
Confirmations	3335	Output	6.80839299 BTC
Timestamp	2021-05-25 03:26:49	Sigops	4
Size (rawtx)	351 Bytes	Fees	0.00000000 BTC
Virtual Size	324 Bytes		
Weight	1,296	Other Explorers	BLOCKCHAIR

Input (0)	0.00000000 BTC	Output (4)	6.80839299 BTC
Coinbase	□Mined by AntPool711ZG□q沌��mm�@�^��□□□�□��쓢! ��\⬆k<�□��□□��□��=U�□	12dRugNcdxK39288NjcDV4GX7rMsKCGn6B	6.80839299 ›
		SegWit commitment output	0.00000000 ›
		Unable to decode output address	0.00000000 ›
		Unable to decode output address	0.00000000 ›

3,335 Confirmations

Figure 16: Coinbase transaction of Block 684887. The figure shows an example of a coinbase transaction in Block 684887 for AntPool. The coinbase signature of AntPool is underlined in red. The address on the right that collected 6.81 BTC is the reward collection address of AntPool.

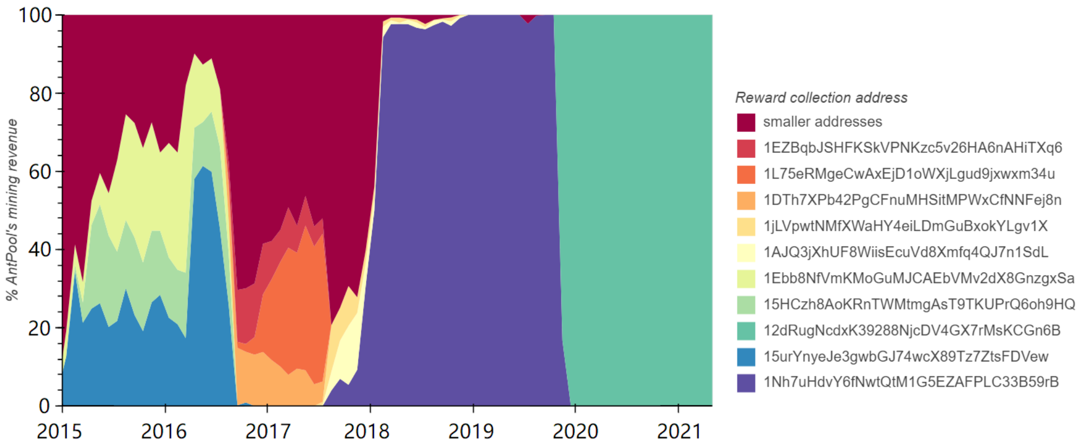


Figure 17: AntPool's coinbase addresses. The figure shows a time-series of the decomposition of the rewards collected by each of the collection addresses for AntPool. Addresses other than the top 10 are aggregated in the "smaller addresses" bin.

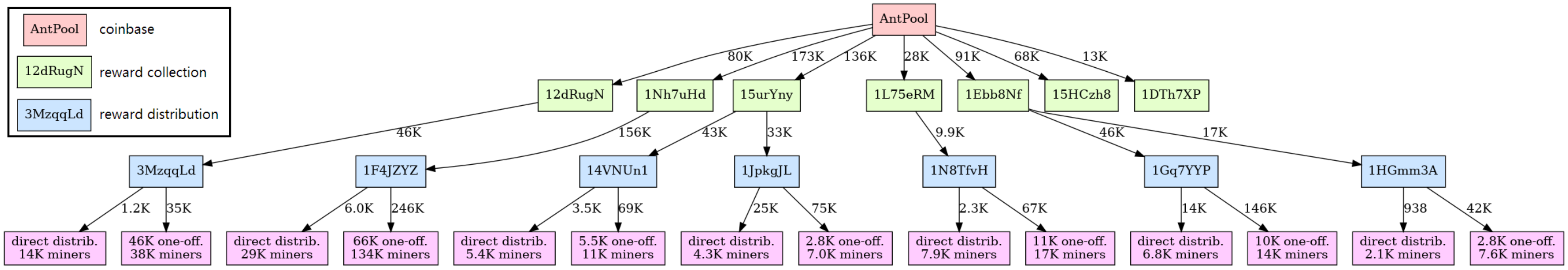
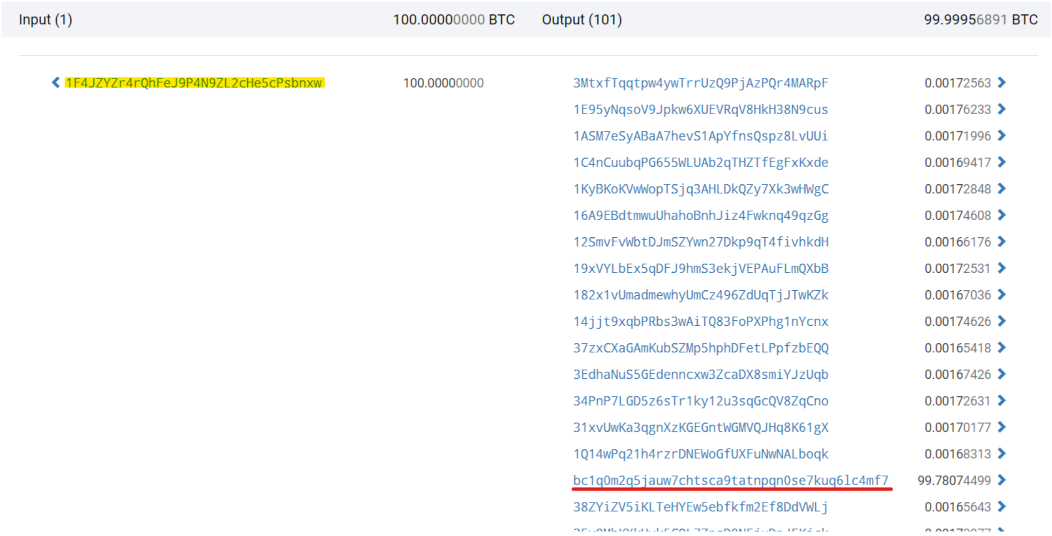
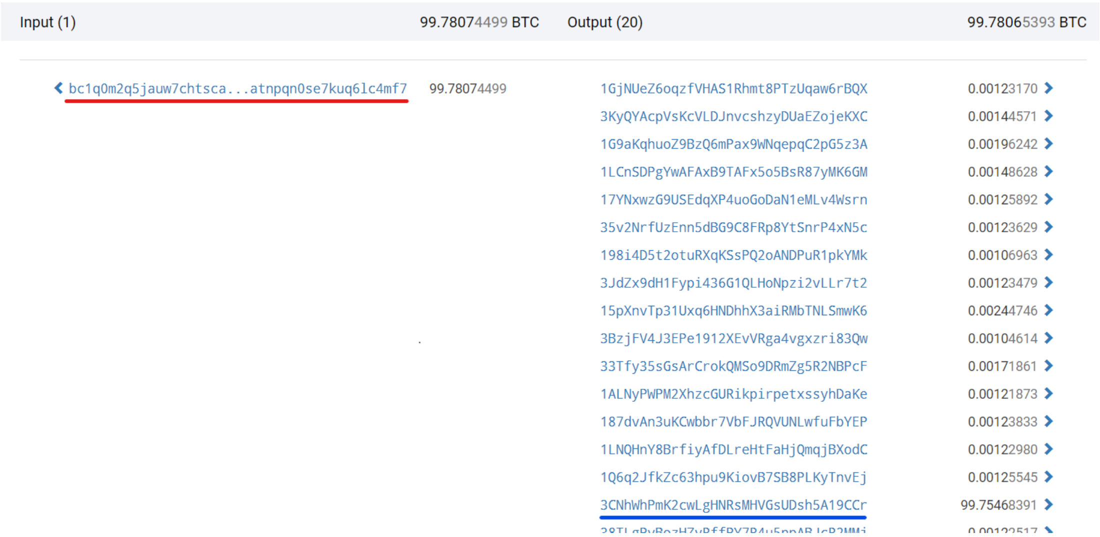


Figure 18: Reward distribution tree of AntPool. Bitcoin addresses are represented by the first seven letters in green and blue boxes. Miners are summarized in pink boxes. Miners are grouped by how they receive reward. Some miners receive a direct distribution from the reward distribution addresses. Some miners receive from one-off distribution addresses that trace back to the frequently-used distribution addresses. Numbers next to arrows represent the Bitcoin volume in BTC. For reward distribution addresses, only their largest source of bitcoins is connected to them with an arrow. Flows of smaller volume are omitted from the plot.



Input (1)	100.00000000 BTC	Output (101)	99.99956891 BTC
1F4JZYZr4rQhFeJ9P4N9ZL2cHe5cPsbnxw	100.00000000	3MtxfTqqtpw4ywTrrUzQ9PjAzPQr4MARpF	0.00172563
		1E95yNqsoV9Jpkw6XUEVRqV8HkH38N9cus	0.00176233
		1ASM7eSyABaA7hevS1ApYfnsQspz8LvUUi	0.00171996
		1C4nCuubqPG655WLUAb2qTHZTfEgFxKxde	0.00169417
		1KyBKoKVwWopTSjq3AHLDkQZy7Xk3wHWgC	0.00172848
		16A9EBdtmwuUhahoBnhJiz4Fwknq49qzGg	0.00174608
		12SmvFvWbtDJmSZYwn27Dkp9qT4fivhkdH	0.00166176
		19xVYLbEx5qDFJ9hmS3ekjVEPAuFLmQXbB	0.00172531
		182x1vUmadmewhyUmCz496ZdUqTjJTwKZk	0.00167036
		14jjt9xqbPRbs3wAiTQ83FoPXPhg1nYcnx	0.00174626
		37zxCXaGAmKubSZMp5hphDFetLPpfzbEQQ	0.00165418
		3EdhaNuS5GEdenncxw3ZcaDX8smiYJzUqb	0.00167426
		34PnP7LGD5z6sTr1ky12u3sqGcQV8ZqCno	0.00172631
		31xvUwKa3qgnXzKGEGntWGMVQJHq8K61gX	0.00170177
		1Q14wPq21h4rzrDNEWoGfUXFuNwNALboqk	0.00168313
		bc1q0m2q5jauw7chtsca9tatnpqn0se7kuq6lc4mf7	99.78074499
		38ZYiZV5iKLTeHYEw5ebfkfm2Ef8DdVWLj	0.00165643



Input (1)	99.78074499 BTC	Output (20)	99.78065393 BTC
bc1q0m2q5jauw7chtsca...atnpqn0se7kuq6lc4mf7	99.78074499	1GjNUeZ6oqzfVHAS1Rhmt8PTzUqaw6rBQX	0.00123170
		3KyQYAcpVsKcVLDJnvcshzyDUaEZojeKXC	0.00144571
		1G9aKqhuoZ9BzQ6mPax9WNqepqC2pG5z3A	0.00196242
		1LCnSDPgYwAFAxB9TAFx5o5BsR87yMK6GM	0.00148628
		17YNxwzG9USEdqXP4uoGoDaN1eMLv4Wsrn	0.00125892
		35v2NrfUzEnn5dBG9C8FRp8YtSnrP4xN5c	0.00123629
		198i4D5t2otuRXqKSsPQ2oANDPuR1pkYMk	0.00106963
		3JdZx9dH1Fypi436G1QLHoNpzi2vLLr7t2	0.00123479
		15pXnvTp31Uxq6HNDhhX3aiRMbTNLSmwK6	0.00244746
		3BzjFV4J3EPe1912XEvVRga4vgxzri83Qw	0.00104614
		33Tfy35sGsArCrokQMSo9DRmZg5R2NBPcF	0.00171861
		1ALNyPWPM2XhzcGURikpirpetxssyhDaKe	0.00121873
		187dvAn3uKCwbbr7VbFJRQVUNLwfuFbYEP	0.00123833
		1LNQHnY8BrfiyAfDLreHtFaHjQmqjBXodC	0.00122980
		1Q6q2JfkZc63hpu9KiovB7SB8PLKyTnvEj	0.00125545
		3CNhWhPmK2cwLgHNRsMHVGsUDsh5A19CCr	99.75468391

Figure 19: Example of AntPool distribution. The figure shows an example of a peeling chain in the reward distribution for AntPool. 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.

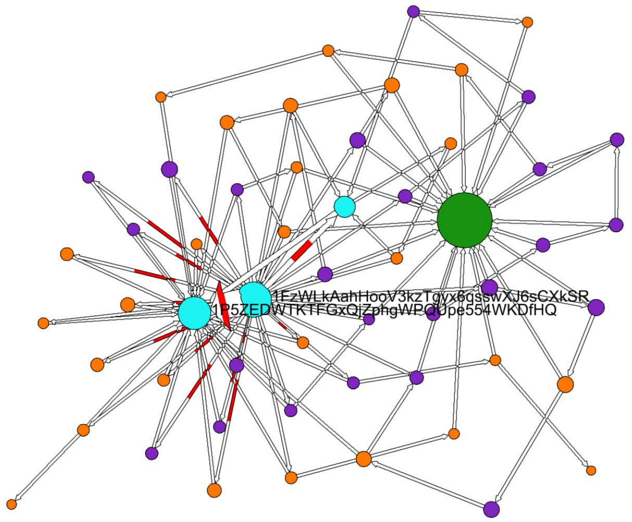

1EzWLkAahHooV3kzTgyx6qsswXJ6sCXkSR
1P5ZEDWTKTFGxQjZphgWPQUpe554WKDfHQ

Figure 20: Tracing networks for rich addresses. The figure shows an example how we trace the Bitcoin network surrounding 1P5ZEDWTKTFGxQjZphgWPQUpe554WKDfHQ, the third richest address on the Bitcoin blockchain. The node and link sizes are proportional to the volume received by the entity and the volume between two different entities, respectively. In the case when two clusters send flows to each other, the direction of the link between these clusters is based on the largest flow, and the link is depicted with a red segment. Identified clusters are marked in orange, unknown high volume clusters are marked in green, turquoise clusters depict short-lived clusters with a life-span below one month, and the remaining clusters are in purple.

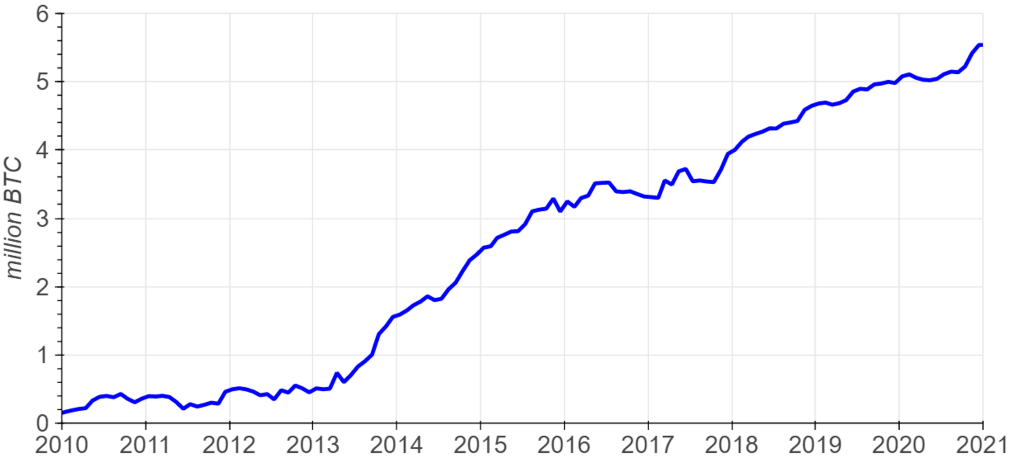


Figure 21: Bitcoin held by intermediaries. The figure shows the amount of Bitcoin held in the wallet of intermediaries from January 2015 until May 2021. Intermediary Bitcoin ownership is determined by tracing "rich" addresses back to their parent cluster. We designate them as intermediary ownership if they can be tied to a known intermediary. We eliminate any ownership at defunct intermediaries.

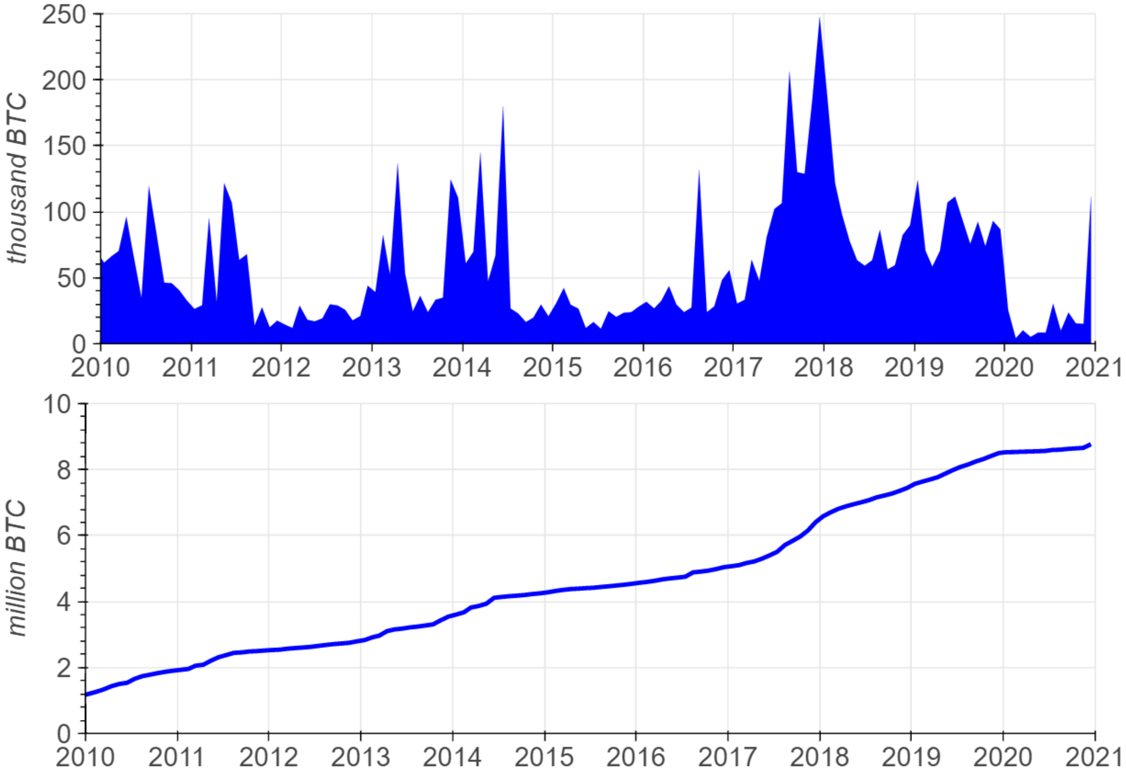


Figure 22: Bitcoin held by individuals. The figure shows the amount of Bitcoin held in the wallet of individual investors over time. Individual holders include "rich" addresses that we classify as individual, and unknown clusters that had a balance below 1000 bitcoins on Dec 31, 2020 and that have not been active in the entire year of 2020. The inactivity constraint separated individual wallets from wallets belonging to intermediaries. Panel A shows the date of the first transaction for each individual cluster and assigns it as a proxy for the age of this 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.

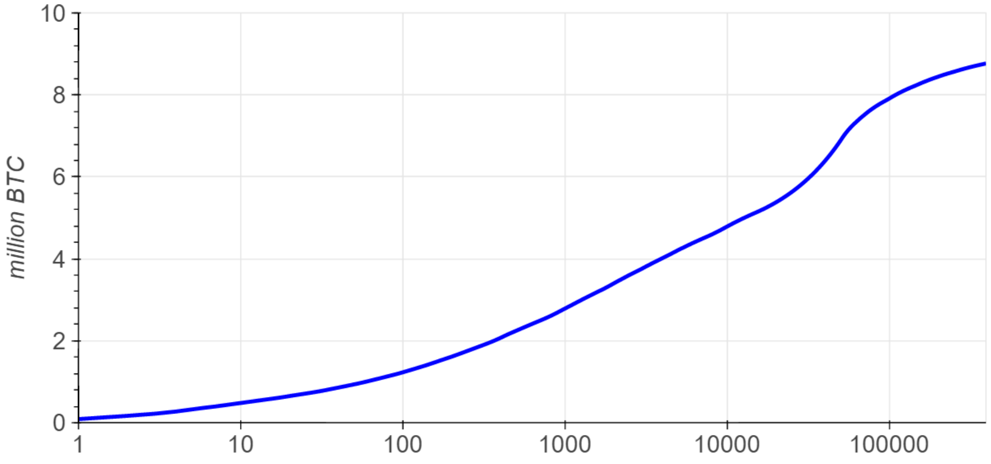


Figure 23: Ownership concentration of individual addresses. The figure shows the concentration of Bitcoin held by individual holders. We sort individual clusters according to their balance at the end of 2020, and plot their cumulative balance against the number of individual clusters that are holding these bitcoins.

Tables

Table 1. Summary statistics of mining pools.
This table reports the number of blocks and the number of bitcoins mined by each pool over the period 2015-2021. We trace the largest pools which are marked in bold font. Private pools are marked in italic.

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

Table 2. Designation of exchange locations.

This table lists the geographic region to which we assign each of the exchanges in order to classify miners by the exchanges they cash out on: (1) US/Europe (2) China, and (3) International.

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