# Blockchain meets network analytics: a tale of heuristics, location and fraud detection*

Simon Trimborn †[1,2,3] and Le Yu[2]

[1]Amsterdam School of Economics, University of Amsterdam
[2]Department of Management Sciences, City University of Hong Kong
[3]School of Data Science, City University of Hong Kong

December 8, 2022

**Abstract**

The information provided by blockchains of cryptocurrencies is immense and diverse. Blockchain information are represented by networks, the transaction network and the user network, approximated by the entity network. We review heuristics for constructing the entity network out of the transaction network, and discuss how they have been used, improved, and developed over time. We introduce network analytics applied to blockchain data, which supports finance and economics research to look into location identification, fraud detection, and price investigations of cryptocurrencies, among other topics. We inspect how they make use of network analytics, often tailored to the specific properties of blockchains. By this comprehensive overview, we intend to aid research on the use of blockchain information to understand user behaviours and the corresponding price behaviours of cryptocurrencies.

**Keywords:** Network Analytics, Transaction Analysis, Location Identification, Price Investigation

**JEL classification:** G00, C00

*Financial support from CityU Start-Up Grant 7200680 "Textual Analysis in Digital Markets" is gratefully acknowledged.

†Corresponding author, phone: +31 643 611 771, E-Mail: trimborn.econometrics@gmail.com

# 1 Introduction

Fourteen years ago, there was only one cryptocurrency. Now there are 13000. Nine years ago, it was soaked with suspicious reputation of illegality. Now, cryptocurrencies have huge market capitalisation, active trading communities and indubitable interest from the mainstream.

It has all started from the famed Bitcoin. Proposed by Satoshi Nakamoto in 2008 to get rid of trusted third parties in payments, this immensely popular and influential electronic decentralized currency depends on a peer-to-peer network with a proof-of-work system to verify the transactions Nakamoto (2008). Altcoins, cryptocurrencies other than Bitcoin, soon followed with Bitcoin as their technological prototype.

With the increase in the price for Bitcoin and other altcoins, people began to add them into their portfolios. Studying the financial properties of cryptocurrencies became important. How would cryptocurrencies perform when they are added to a portfolio consisting of other assets (Petukhina et al., 2020)? What are their diversification properties and financial factors (Liu and Tsyvinski, 2021)? How to identify a cryptocurrency's idiosyncratic risk (Guo et al., 2020)? Studies answering these questions tend to focus on the *time series properties* of coins.

But there is another way of looking at things. Each cryptocurrency is powered by a blockchain. As the underlying platform and engine of a cryptocurrency, blockchain characteristics cannot be irrelevant to the financial characteristics of the cryptocurrency and to its overall success. Are user behaviours on the blockchain relevant to the price of a coin (Cong et al., 2021c)? Will network measures and computing power of a blockchain impact the coin's expected returns (Bhambhwani et al., 2021)? Can we construct pricing models out of blockchain network characteristics (Liu and Tsyvinski, 2021; Cong et al., 2021a)? To answer these questions, we need to turn to the *network analytics* of blockchains and cryptocurrencies.

*Network analytics* of the blockchain are powerful tools to understand the behaviour of users in the ecosystem defined by any particular blockchain. Understanding user behaviour goes beyond its financial implication, since there is a lot of talk about its application outside the realm of cryptocurrencies, such as tourism (Önder and Treiblmaier, 2018), healthcare (McGhin et al., 2019) and supply chain (Saberi et al., 2019). If such applications materialise, blockchain technology would have profound social and economic impact. But why would one want to use blockchain technology instead of established technologies? Catalini and Gans (2020) provided a detailed analysis of the economics of the blockchain. In short, the magic of blockchains can be summarised by two points: they reduce the *cost of verification*, and they reduce the *cost of networking*.

*Cost of verification* arises inevitably in an economy of any respectable scale. Intermediaries increase the risk of "moral hazard," namely cheating, counterfeiting and other behaviours that strain trust, in the words of Catalini and Gans (2020), and information asymmetry in general in any real-world economy, in which most people do not have pre-established relationships. Blockchain technology helps in reducing the cost of verification by striking at its roots: because they involve distributed ledgers, there is no intermediary involved. If an economic activity can be done completely online, the verification cost may be almost costless.

*Cost of networking* arises in bootstrapping, running and scaling an economic network. The relationship between the cost of verification and the cost of networking is close but nuanced. Over time two types of blockchains arose: permissioned and permissionless. Permissioned blockchains can have a significantly reduced verification cost but they do not necessarily significantly reduce the cost of networking: unlike Bitcoin, they are not entirely public. Rather like traditional databases, they need an administrator-role to govern the approved users of the blockchain. Trust needs to be put on the administrator. By contrast, in permissionless networks, or public blockchains, trust is an unnecessary luxury. Without trusted intermediaries or privileged nodes of any kind, cost of launching and running a network is significantly reduced.

These cost reduction advantages of blockchains have the potential to boost applications in many areas. This further drives the need to understand how users interact on the blockchain (Meiklejohn et al., 2013), how illegal activity can be identified (Foley et al., 2019), how users obfuscate their behaviour (Ron and Shamir, 2013), or to identify geographical user clustering (Trimborn et al., 2022; Makarov and Schoar, 2021). These topics and the resulting research questions can be tackled with *network analytics*.

This study introduces network analytics applied to blockchain data, which supports finance and economics research to look into location identification, fraud detection, and price investigations of cryptocurrencies, among other topics. We inspect how they make use of network analytics, often tailored to the specific properties of blockchains.

In section 2 we distinguish between different kinds of networks relevant in the discussion of a blockchain of a coin. Section 3 then discusses the relationships between different networks and the heuristics to transform one into another. After knowing what the networks are, in in section 4 we are going to look at how they are used. Section 4.1 tracks down the location of cryptocurrency users: useful for economic insights but so much for anonymity claims. In section 4.2 we filter out irrelevant transactions in a cryptocurrency networks and keeps only the "economically meaningful ones," in a sense to be explained (Makarov and Schoar, 2021). Section 4.3 is about fraudulent behaviours associated with cryptocurrencies: this is how cryptocurrencies got their bad names. Section 4.4 is about

financial characteristics of coins, especially their pricing models. Section 5 concludes.

## 2 Network Types in the Blockchain

First, we discuss and introduce the different kinds of networks which appear or can be constructed from the blockchain. We will not introduce and explain Bitcoin and the blockchain before delving into the details of this study, since there is already a vast literature covering these topics. We refer you here to the comprehensive introductions into the features of Bitcoin and the blockchain by Böhme et al. (2015) and Tschorsch and Scheuermann (2016). Given the Bitcoin scheme, there are three closely related networks we can look at:

- the transaction network;
- the user network;
- the entity network.

A *network*, or a *graph*, is a set of objects in which some pairs are related in some relevant sense. The objects are called *points*, or *vertices* (singular: vertex) if you are a mathematician, or *nodes* if you are a computer scientist, or *sites* if you are a physicist, or *actors* if you are a sociologist. Related pair of points are connected by *lines*, or *edges and arcs* (maths), or *links* (computer science), *bonds* (physics), or *ties and relations* (sociology).

Graphs are useful abstractions, and they are especially useful here. There are natural correspondences between elements of a graph (vertices and edges) and elements of the networks that we are going to look at, such as transactions and transacting; users and their relationships.

The *transaction network* captures the flow of bitcoins between transactions. This is Bitcoin's original, entirely public transaction graph. Each vertex is a transaction. A directed edge from vertex $a$ to vertex $b$ represents that an output of $a$ is used as an input of $b$. Transaction graphs are also called address graphs.

The *user* network captures the flow of bitcoins between users. Each vertex is a user of Bitcoin identified by a public key. A directed edge from vertex $a$ to vertex $b$ represents an input-output pair of a transaction. The input's public key belongs to the user of the source, and the output's public key belongs to the user of the target. A user of Bitcoin is also called an *owner* or a *controller*, though we will stick to *user*. Analysis of the user graph elucidates the interactions between different economic players. It can tell us who the users are, where they are situated geographically, and which types of businesses they are involved in.

However, for all intents and purposes, the true user graph is unknowable. The Bitcoin scheme has the reputation of being anonymous. While the degree of anonymity it offers is a topic of hot debate, there is some truth to it. First, intrinsic to the Bitcoin protocol, the true identities of users are not linked with public keys. A single user can control many public keys, and a single public key can be controlled by many users. Second, user habits further obfuscate their identities. It is declared good practice for users to generate a new address for every transaction. People also mix and launder, gamble and steal.

Researchers managed to approximate the user graph well-enough to derive insights that we want from it. The *entity network* is the working graph that researchers get as an approximation to the unknowable user network. User network is the real thing: entity network is an unavoidable compromise. The networks amenable to study, therefore, are the transaction network and the entity network.

# 3 From Transactions to Entities

With the degree of anonymity offered by the Bitcoin protocol, it is doubtful whether we can ever achieve a full trace of users from their transactions. Yet some trace is better than none. We can achieve a good approximation of the user graph, dubbed the entity graph. There are three major heuristics to contract the transaction graph into the entity graph:

1. the heuristic of multi-inputs,
2. the heuristic of change addresses, and
3. the heuristic of off-network information.

Other minor ones abound.

The work "heuristic" is proper here. In contrast to a proof or an algorithm, a heuristic is a special approach to problem-solving that is experimental and relies heavily on trial and errors. None of the heuristics that we shall see is foolproof. They are empirical, context-dependent, ad hoc even. However, it can be argued that they work reasonably well, and the entity graph they generate offers us important insight.

## 3.1 Multiple Inputs

**Heuristic 1** *For multi-input transactions, treat the addresses of the inputs as the same user.*

Heuristic 1 was first identified by Nakamoto himself in Nakamoto (2008). He realised this when discussing anonymity:

> *Some linking is still unavoidable with multi-input transactions, which necessarily reveal that their inputs were owned by the same owner. The risk is that if the owner of a key is revealed, linking could reveal other transaction that belonged to the same owner.*

Why is it reasonable to link multiple inputs? Because each sender needs to know the private key belonging to each public key used as an input to the transaction. The public key is known publicly to everyone on the blockchain. Users require a private key to proof that a transaction, signed with a public key, belongs to them. Hence, users have an incentive to keep their private keys as safe as possible. Whoever possesses the private key has access to the funds signed with the public key. If a collection of public keys were controlled by multiple entities, then they would need to reveal their private keys to each other. The idea is that this is unlikely, though not impossible.

With hindsight, we now know that this assumption does not hold infallibly. Benignly, web wallets, for example, pool many private keys. They would be mistakenly identified as a single entity. Maliciously, by taking care to make sure that his multiple addresses are never used in a single transaction, a user can deliberately conceal the true connection between his addresses. As a result, as pointed out by Makarov and Schoar (2021), the multi-input heuristic provides a lower bound for the actual number of users.

Heuristic 1 of multi-input is implemented by Reid and Harrigan (2013). True to Nakamoto exegesis, their overriding concern was with the breakdown of anonymity. Taking the above cited passage as a hint, they merged any vertices with undirected edges where each edge joins a pair of public keys that are inputs to the same transaction. From the transaction graph, they contracted an entity graph.

They immediately identified crucial differences in features between the transaction and the entity graphs. The transaction graph has neither multi-edges nor loops. It is also a directed acyclic graph, which means that the output of a transaction can never be an input to the same transaction, either directly or indirectly. On the other hand, the entity graph is much messier. It does contain multi-edges (a user sends multiple transactions to himself), loops (a user send to an account of his own and then send back) and directed cycles.

Heuristic 1 using multi-inputs is popular and became the de facto default method for contraction from the transaction graph into the entity graph. Lischke and Fabian (2016), in their data preprocessing, first contracted the transaction network using the multi-input heuristic before further processing the data with off-network information. Makarov and

Schoar (2021), to get their "most complete information about crypto entities that have been used in academic research up to this point," they scraped cryptocurrency blogs, websites, and the database Bitfury Crystal Blockchain. They then contracted the transactions graph based on the multi-input Heuristic 1 and got 1,032 entities, including 63 mining pools, which further illustrates the fact that this heuristic gives us a lower bound. Other notable studies that used this heuristic include Ron and Shamir (2013), Meiklejohn et al. (2013) and Athey et al. (2016).

### 3.2 Change Address

**Heuristic 2** *The change address is controlled by the same user as the input address.*

The *change address* is an offspring of a peculiar idiom of use of the Bitcoin scheme. I want to buy a bottle of water worth \$1, but I only have \$5 notes with me. I give you a \$5 note and you give me back \$4 of change. Similarly, change addresses are used to give money back to the input entity in a transaction. But unlike cash transactions, in which case you give the money back into my hands, Bitcoin change addresses are generated by the Bitcoin client. These change addresses are not chosen by the input entity and should not be reused by the user. As a result, it is likely that the change address and the input address are controlled by the same entity. If we can identify these change addresses reliably, then by linking the change address to the input address, we can further contract the transaction graph.

This line of thought is pursued by Androulaki et al. (2013). They called the change address the *shadow address*. The main idea is that if a transaction has two outputs, with one public address and one newly generated address (the shadow), then the shadow address belongs to the entity who initiated the transaction. The thought is that it is likely that the input entity is receiving change. However, the assumption they relied on in their study is not maintainable and is in fact routinely violated. In real life, entities issue transactions to different users, much more than two, all the time, for example, in mining pools or bets on gaming sites.

As a result, Meiklejohn et al. (2013) took the matter more stringently. They imposed four conditions before identifying an address as a change address: 1) this address cannot have appeared before, 2) the transaction cannot be a coin generation, 3) there is no self-change address, and 4) for all outputs in the same transaction, only this address is making its first appearance.

Recall that Heuristic 1 of multi-inputs relies on assumptions that may not be true. If anything, Heuristic 2 of change address is even less robust. The existence and features of

the change address are heavily dependent on the present idiom of use and is not inherent in the Bitcoin protocol itself.

Heuristic 2 is also tricky to implement. After Meiklejohn et al. (2013) ran the analysis, they ended up with a giant super-cluster containing 1.6 million public keys, including those of the popular services Mt. Gox, Instawallet, BitPay, and Silk Road.

As a result, they imposed further conditions. Sometimes the same change address is used twice, and if the second use is with a new address, the new address would be falsely labelled as the change address. They got rid of those problematic transactions to make the heuristic more robust. Heuristic 2 based on change address, as a result, depends on the unique situations within the blockchain.

And Heuristic 2 has several variants. Another notable one was introduced by Athey et al. (2016). They based their heuristic on a observation of human psychology. If I give you change, for example, it is unlikely for me to give you tortuous amounts, or in Athey et al. (2016) words, amounts that are "cognitively-difficult", such as $4.94275. More likely I would just generously say: "forget about it, let me give you $5." Athey et al. (2016) therefore conjectured that if a transaction has two outputs, and if one of them has 3 more decimal places than the other (3 is a heuristic and is not based on intrinsic reasons), then the output with more decimal places is the change address as it is less likely users would send such "cognitively-difficult" amount to each other. They supplemented Meiklejohn et al. (2013) one-time change address heuristic with their own decimal-place heuristic to contract the Bitcoin transaction graph.

Apart from developing the Heuristic 2 of change address, Meiklejohn et al. (2013) also actively engaged in Bitcoin transactions with various services. Because they, of course, knew what public keys they themselves have used, the public key on the other end can be explicitly tagged as belonging to the service they are interacting. They also scouted in forums and other places online to look for addresses claimed publicly. These activities cannot be found in the network themselves: they are squarely within the realm of the third heuristic, to which we now turn.

### 3.3 Off-Network Information

**Heuristic 3** *Anonymity can be further broken down by off-network information.*

It is impossible to live without leaving some trace. Although users are not linked to their Bitcoin addresses explicitly, there are always some giveaways. Sometimes these off-network information is enough to link addresses to users. A good contraction does not only rely on features of the transaction network, but on extra information that is circumstantial and not

intrinsic to the network itself.

This line of thought is first pursued by Kaminsky (2011) in concurrence with Reid and Harrigan (2013). Kaminsky exploited the fact that Bitcoin is an electronic currency requiring Internet connections. His assumption is that the source of a transaction is the first node that informs you of that transaction. Kaminsky proposed to map IP addresses to Bitcoin public-keys by making connection to all public peers in the network simultaneously. Again, this assumption is not infallible, though Kaminsky is reasonably confident about it: "[this is] more or less true, and absolutely over time."

Reid and Harrigan (2013) integrated some other network information to de-anonymize users. They considered a wide variety of sources to glean their off-network data. For example, they scraped the Bitcoin Faucet, where IP addresses are published together with the history of recent giveaways. They also searched in Bitcoin Forums where users attach public keys to their signatures, Twitter streams and user-generated public directories. They acknowledged the ad hoc-ness of their endeavour and that a larger, more centralized Bitcoin service provider can perform the same analysis with their user information more reliably and on larger scale. But their attempt is an invaluable proof-of-concept. By integrating such a rich variety of off-network data to the entity network, it is possible to construct the entity network as a very good approximation to the real, unknowable user network.

Apart from discussing the multi-input Heuristic 1 and the change address Heuristic 2, Meiklejohn et al. (2013) also considered off-network data to aid the contraction. Other than passively spending time in Bitcoin Forums and other online places to find user disclosures, which they served "manual due diligence" and regarded as "less reliable," they actively transacted with a wide variety of Bitcoin services. Because they know which user they are transacting with, they can tag users by observing the addresses they used. They mined with major mining pools, kept money in wallet services, engaged in bank exchanges and non-bank exchanges, purchased physical and digital goods with vendors, gambled with poker sites (but not dices because they already advertise their public-keys), and interacted with mix and laundry services, where some services sent their own coin back and others simply stole their money.

Lischke and Fabian (2016) is another example of using both Heuristic 1 of multi-inputs and Heuristic 3 of off-network information. After contracting the transaction network using Heuristic 1, they considered IP addresses and business tags from the initiator of the transactions. To deal with Tor and proxy nodes, they downloaded all current Tor server and the addresses of Tor server exit nodes, which resulted in the finding that around 1% of nodes might used a Tor network. Their study is therefore also a refinement of Kaminsky's pioneering talk, Kaminsky (2011), which did not deal with proxies.

# 4 Blockchain Analytics in Finance and Economics

Networks on the blockchain in hand, we are ready for the economic and financial implications of the blockchain.

## 4.1 Location Analysis

Identifying the geo-location of users is one of the most important goals of Bitcoin network contraction. The mainstream method of tracking down users' geo-location is by analyzing their IP addresses. This line of analysis has a venerable history. Reid and Harrigan (2013) already realised that the Bitcoin Faucet can map users to geo-located IP addresses, and it would be interesting to know where people are using Bitcoin. They went as far as producing a map-plot of the users receiving bitcoins from the Faucet. This is done in larger scale by Donet Donet et al. (2014), which we will discuss later in more detail in the connection with Bitcoin address propagation.

Knowing the whereabouts of users is interesting in itself. There are, however, deeper reasons for performing geo-location analysis. Geographical concentration can bear risks for the functioning of the blockchain system. As (Makarov and Schoar, 2021) pointed out: "geographical 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." Such a situation could arise due to several situations. For example, if the combined computing power of all users of a blockchain network in one town is large enough, they could take together control of the network. Also if a significant number of miners is located in the same area and a power outage appears, the blockchain network would be seriously disrupted due to the inability of these miners to contribute to its operations for the duration of the power outage. Consequently, geographical analysis of blockchain users provides important insights into the risks the network is exposed to.

There are four prominent ways to track the geo-location of a blockchain user:

- Relaying node IP tracking (section 4.1.1)
- Bitcoin address propagation (section 4.1.2)
- Off-network information collection (section 4.1.3)
- Miner geo-tracking via cashout behavior (section 4.1.4)

Let's turn to them one by one.

### 4.1.1 Relaying Node IP Tracking

Nodes in the Bitcoin transaction network are not all created equal. Some of them act as coordination nodes, or "relaying nodes," which transmit the information of a newly recorded transactions to all users. The address of these relaying nodes are publicly known: one can track their IP addresses by connecting to them. And the relaying node which relays a transaction first should be the one closest to where the transaction originated. This line of thinking gives us a viable way to estimate the origin of a transaction, which is a variant of Heuristic 3 above.

**Heuristic 3.1** *Transaction takes place where the node first informs about it.*

Though this Heuristic 3.1 of IP addresses is reasonable, it holds true only without the widespread use of proxies.

Lischke and Fabian (2016) performed a comprehensive mapping between IP addresses and geo-locations. They extracted IP-addresses of Bitcoin users from Blockchain.info, and identified 40,329 distinct geo-locations. Of course, as all heuristics, the link between transactions and IP addresses is not an infallible one. They reported that around 70% of the transactions can be linked to an IP address. As of 2016, the time of their writing, based on the geo-locations of the IP addresses, they identified the active markets in terms of the number of transactions and the associated value of bitcoins. The US takes an undeniable lead, followed by Germany. With a sharp decrease, France, Russia and Canada are the next contenders.

Can we go further than throwing darts onto the map? Trimborn et al. (2022) looked at the time dependent impact of Bitcoin transactions in different geo-locations. They again followed the above Heuristic 3.1. After grouping their data continent by continent, they further broke each continental group into ten according to transaction sizes. Not surprisingly, Antarctica is not worth a mention. They therefore ended up with 60 groups, ten for each of the remaining continents, Africa, Asia, Europe, North America, Oceania (Australia) and South America.

By focusing on transaction data of these 60 groups from 25 Feb 2012 to 17 Jul 2017, a time series of daily log accumulated transactions, they already made interesting observations using just descriptive statistics. In short, European and North American transactions were large and steady. In terms of transaction sizes, Europe and North America took the lead, followed by Asia and Oceania, and then by Africa and South America. In terms of volatility, Europe and North America have steady transactions sizes, whereas the other continents were very volatile, with days of zero transactions on end.

With a time series of Bitcoin transaction records to hand, an obvious question to ask

is: "Do Bitcoin transactions in the past influence Bitcoin transactions in the future?" "Of course" is the answer, but how can we study it?

Trimborn et al. (2022) constructed a time-dependent model to analyse the 60 groups across time and space. The model considers each group as potentially influential for all the other groups. Via regularising the groups informative value for the transaction behaviour on other continents, the group with a steady impact on all the other groups remain. Across time, they analysed the occurrence and vanishing of structural influence in the Bitcoin blockchain over the years. 2012 is particularly active. In 2012 Bitcoin first captured the imagination of the mass. Together with its activity, its price shoot up from $5.27 to $770 by 2014. However, as other cryptocurrencies, or altcoins, began to emerge, Bitcoin had its hibernation and did not show network effects from 2013 – 15. Something out of the ordinary must have happened in 2016. Not only did Bitcoin again turned active, the price again went up.

Across space, their analysis is a pronounced reminder of the true international character of Bitcoin. Asia, as the place where most mining farms located, often has the greatest media attention. Europe and North America, as developed regions and financial centres, are also assumed to have a lot of Bitcoin flows. However, they showed that South America and Africa are showing network effects of significant magnitude. In fact, apart from 2014, South America are outperforming North America and Europe. Through the mining farms in Asia, people in every corner of the world are contributing to the Bitcoin blockchain.

### 4.1.2 Bitcoin Address Propogation

Each Bitcoin node aims to maintain at least eight connections. The maximum connection is usually set at 125 (*maxconnections*). A caveat is that, as reported in Miller et al. (2015) in a study of Bitcoin's public topology by studying Bitcoin's peer-to-peer link using their technique of AddressProbe, many nodes in fact exceed the number of the supposed maximum connection persistently over 80 times. These high-degree nodes are primarily mining pools and wallets.

Bitcoin nodes, like employees with roll-call attendance responsibilities, announce their own IP address every 24 hours (*addr* message) throughout the network. Nodes can also actively ask for IP addresses using the *getaddr* message. Those nodes receiving the *getaddr* message would duly response. Instead of disclosing everything they know, however, nodes only include 23% addresses they know randomly, and in any case not above 1000 of them. By doing this, the Bitcoin protocol deliberately tries to mitigate the risk of information eclipsing. Information eclipsing can happen spontaneously or maliciously. If a villain dominates the environment of a node, what information gets to the node will be solely at his mercy. He may decide to report the correct information, but he may also decide to double-

spend for fun. By disclosing only a small part of the IP addresses a node knows, that node is protecting itself from the threat of the villain which comes with the intention of filling up a node's neighbour with compromised IPs in mind.

The *getaddr()* command gives us another way to find out where Bitcoin users are located in the world. Donet Donet et al. (2014) extracted more than 800,000 IP addresses by exploiting precisely the *getaddr* messages. Everyday at 9am from 30 November 2013 to 5 January 2014, they issued *getaddr()* command to a set of seeds, and then recursively to the nodes connected to those seeds. By thus discovering the neighbours of the initial nodes, and their neighbours, and their neighbours' neighbours, they collected a giant list of IP addresses running a Bitcoin node.[1]

From this list, Donet Donet et al. (2014) were able to identify the geographical location of Bitcoin nodes. Not surprisingly, the United States and China are where most of the nodes located, followed closely by Germany, Russia and the United Kingdom. They also calculated the Bitcoin adoption rate, normalizing the above with the number of internet users in each country, and the list of countries appeared quite differently. Now the leading countries are the Netherlands, Norway, Finland and the Czech Republic. And their summary is a true testimony to the viral popularity of Bitcoin: "there are Bitcoin nodes all over the world, with very low populated areas and underdeveloped countries being almost the only exceptions." Reminiscent of NASA's breathtaking photo of the Earth's city lights, in their geo-location plot, discovered nodes thickly dotted all the continents, with the sole exception of Antarctica.

#### 4.1.3 Off-Network Information Collection

Athey et al. (2016) conducted a geo-location analysis on Bitcoin users without relying on their IP addresses. Rather, they used the random forest algorithm, which is a generalization of the simple decision tree, to identify the regions of transactions.

They got their training data sample from Bitcointalk.org, which is a website filled with user gossip about their geo-locations. They scraped it and got 2995 addresses labelled with geo-locations. They then grouped all countries into four groups, not following geographical continent demarcation specifically: the Americas, Europe, Asia and Eastern Latin America. After training with random forest, they used their model to predict the origin and shares of the full data.

About the performance of the model, they reported that: "it may be surprising that

[1]It is noteworthy that Kaminsky (2011) in his talk already proposed the method of "recursively ask every node about every other node it knows about" using the *getaddr* message, and proposed to start from hardcoded seeds. This seems to be a case of independent rediscovery on the part of Donet Donet et al. (2014) because they seem to be unaware of this.

the error rate is as low as it is, given the small size of the training dataset and the fact that a large share of users engage in only a few transactions per user." With a larger or more representative training dataset, organizations and agencies that are interested, by virtue of conducting a similar analysis, have the prospect of making very accurate predictions of the geo-locations of users. This is good news for some (e.g. law enforcement) and bad news for others (e.g. criminals or those who otherwise wish to remain anonymous).

Finally, it should be noticed that Athey et al. (2016) random forest classification does not seem to be as fine-grained as the approach based on IP addresses. Random forest is only possible to place entities into large geographical bins, but not pinpoint them on the map.

#### 4.1.4 Miner Geo-Tracking via Cashout Behaviour

Makarov and Schoar (2021) proposed a heuristic to track down the geo-location of Bitcoin miners in particular. They focus on the exchanges in which miners cash out their rewards. Their heuristic is based on the thought that miners are likely to send their rewards to an exchange in the region in which they are themselves physically located. Their heuristic is therefore

**Heuristic 3.2** *Miners are physically located in the area where they send their rewards to.*

By following this Heuristic 3.2, they distinguished between four categories of exchanges: Chinese, US-Europe, International and Other. Chinese and US-Europe are just what their names advertised. Others include those transactions that operates in areas other than China, the US and Europe. The International category involves trans-jurisdiction exchanges, so tracking down these flows may not give us valuable information about the whereabouts of the miners. As of 2021, they found that Chinese miners carry the most weight in the landscape, responsible for around 70% of all the transactions.

Apart from geo-locational analysis, Bitcoin network contraction also has several other applications. Other popular analyses include

- Industry type: what uses are bitcoins being put by individuals and companies? Athey et al. (2016)'s analysis identified thirteen major uses, with exchanges being the top use by far. Interestingly, in second place is a cluster of industries labelled by them as "unknown." Lischke and Fabian (2016) did further analysis in this respect. Going above a cross-section characterization, they identified a changing application pattern of bitcoins in different businesses. Under their analysis, donations, wallets and gambling are important.

- User type: who are using bitcoins? Are people short-term users or long-term users? Are people transactors, miners, or merely investors? Athey et al. (2016) looked into this question and discovered that long-term, frequent users are in fact only a minority. Rather, most people tend to hoard their bitcoins. Investors are therefore most important in contributing to Bitcoin transaction volume. Their analysis confirmed the earlier findings of Meiklejohn et al. (2013): although people began to spend their bitcoins since 2011, hoarders still exist in legions. Ron and Shamir (2013) gave a very early analysis of Bitcoin's user type. At the early time of their writing, they found that 78% of all bitcoins were left in addresses that only received but never sent out bitcoins.

### 4.2 Blockchain Analytics for Price Investigations

Why is it important to investigate cryptocurrency pricing? Apart from the usual reasons for conducting studies on pricing, cryptocurrencies are special in this respect. Many of them are unbacked by any underlying assets or supporting foundation/company, its exchange rate is therefore governed by different fundamentals than other assets. As a result, the pricing series of cryptocurrencies attracted much attention.

In the absence of regulation, it is plausible to postulate that Bitcoin prices are governed by "economic fundamentals," i.e. supply and demand. This is precisely what Athey et al. (2016) thought. They built a model which demonstrates that the Bitcoin exchange rate is grounded in supply and demand, and exchange rate rises with usage. Their central theoretical conclusions are

- There is a unique equilibrium exchange rate in each period determined by supply and demand provided that a) no investor is present and b) all agents adopt Bitcoin eventually.
- If investors buy bitcoins, the effective supply for users decreases and the market equilibrium price increases.
- Bitcoin adoption and exchange rates are influenced by beliefs about Bitcoin.

In short, more intense activity on the blockchain means an increased demand for the use of it.

As they themselves noted, in reality, things are perhaps more complicated than in models. In particular, their model does not take into the account the competition provided by altcoins, or traditional banking itself. Another thing to note is that they did not take into account of state variables of users other than beliefs. Heterogeneous beliefs and speculative bubbles, something that Bitcoin may be especially prone to, are also not considered.

In addition, for each cryptocurrency, its blockchain and the behaviour of users on it impact its price. Following this thought, Pagnotta and Buraschi (2018) constructed a model to study the price equilibrium of Bitcoin as a function of its blockchain characteristics, such as the hashrate. A higher hashrate indicates that miners are dedicating more computing power to the network. It is sensible to assume that hashrate and price are related. More computing power dedicated to a cryptocurrency is associated with higher costs, and so the reward need to be higher to make up for this. However, the reward is fixed and, in the case of Bitcoin and many other cryptocurrencies, halves regularly. As a result, if hashrate increases, the price has to go up. If the price does not go up, miners would just lower their dedicated computing power, namely lower the hashrate. Price and hashrate are in equilibrium.

Financial factors to explain movements of cryptocurrencies on the exchanges are also interesting. A financial factor is a constructed variable which is proven to have an impact on asset prices. The most famous financial factors were discovered by Fama and French (1992) which comprise, among others, the book-to-market-ratio. This factor describes the difference between the market value of things owned by a company, and their value in the accounting books. If the assets of a company worth more than they are in the accounting books, this will influence the price of the company's shares.

What are the financial factors for cryptocurrencies? Bhambhwani et al. (2021) constructed such factors for the network size of the blockchain using unique addresses and the hashrate (proxy for computing power). They showed that these two factors are positively associated with the return structure of the cryptocurrencies in their sample.

This empirical analysis of the positive relationship between the hashrate and the price supports the theoretical model of Pagnotta and Buraschi (2018) and Athey et al. (2016): a higher number of unique addresses implies a larger number of active users. And a larger number of active users raises demand for the cryptocurrency in question and therefore increases its price.

Following the same path, Cong et al. (2021a) built an asset pricing model and construct financial factors based on the growth in addresses of the underlying blockchain. Their findings also point towards that network activity and size are important for the price of a cryptocurrency, which further support the theoretical model of Athey et al. (2016). They constructed five weekly re-sorted portfolios, sorted by quintiles of the growth in transactions with balance, and growth in total transactions. In a long-short analysis, the portfolios sorted by their financial factors produced a statistically significant difference in returns.

They also constructed financial factors for the growth in the total on-chain transaction volume, measured once in terms of tokens and also in terms of the total value of tokens in USD. However, for these factors, the difference in returns was statistically insignificant.

They cannot explain the cryptocurrency price dynamics.

### 4.3 Filtering for Relevant Transactions

Not every transaction on the blockchain is relevant for analysis. Based upon the target of ones study, different types of transactions are important. To be able to extract the relevant transactions from the blockchain, we have to consider first the transaction patterns which commonly appear. E.g. if one is interested in identifying who holds currently a coin, the initial and final transaction of the coins history are relevant. The rest can be omitted hence simplifying the analysis. Bitcoin transaction patterns can be grouped into the following categories, as comprehensively traversed by Ferrin (2015):

For single transactions, Bitcoin follows

- Peel transactions. This is the most common type of transaction on the Bitcoin blockchain. They can have any number of inputs, but only two outputs. One output is the receiver's, the other is the change address.
- Sweep transactions. They combine multiple inputs into one output: many inputs are swept into one.
- Distribution transactions. They can have any number of inputs, but they have more than 3 outputs. They are mostly used when a single organization pays many parties.
- Relay transactions. They have one input and one output. This is also a popular pattern. Bitcoins can be moved from one address to another without leaving traces such as change addresses, which are seized by analyzers using e.g. Heuristic 2 to break down anonymity.
- Self-spending transactions. They can have any number of inputs and outputs. Their distinguishing feature is that one or more of the input addresses also appears as output addresses in the same transaction.
- Meta-transactions. In those transactions, external data is inserted into the transaction to enhance the stability and trustworthiness of the Bitcoin blockchain.
- Joint transactions. In contrast with sweep transactions, where addresses are swept into one, in joint transactions, different transactions are combined into one larger transactions using some protocol.

Note that some of the transaction patterns can be layered. For example, a transaction can both be a peel and be a self-spending one: someone careless may just use his old address

as the change address. One can also sweep into his own address, making it both sweep and self-spending. There can also be meta-peels and meta-relays.

For multiple transactions, Bitcoin follows

- Peeling chains. They are a series of peel transactions. The start of a peeling chain is usually an address that contains a significant amount of bitcoins. Then, a small amount is peeled from it, and a small amount is peeled from the remainder, and so on. Peeling chains are the focus of study of several important works, including Ron and Shamir (2013), Meiklejohn et al. (2013), Makarov and Schoar (2021).
- Green addresses. In these transactions, bitcoins pass through a single publicly known address that is trusted by the receiver of the coins.
- Mixing clouds. Also known as tumblers, mixing pools and washers. They are multiple interconnected joint transactions. As the names suggest, mixing cloud transactions aim to be very difficult to trace.
- Tunneling. These are the transactions that have connections outside of the blockchain in question. This is a technique advertised by some Bitcoin services for anonymity: bitcoins are shifted from account to account and finally to another address of the original owner.

Some of the outlined transaction patterns can be utilized to cluster different transaction/addresses into a single entity. In particular, for example, Heuristic 2 depends on change addresses. Other transaction patterns, such as meta-transactions, joint transactions, green addresses, mixing clouds and tunnelling are designed to obfuscate the true users behind the scene.

Chang and Svetinovic (2018) went beyond Heuristics 1 and 2 and performed clustering using different Bitcoin transaction patterns. In peels, they identified the owner of the sending address and the owner of the change address (in effect, Heuristic 2). They identified all addresses that participate in self-spending transactions. They grouped all the relay addresses. For sweeps, they first identified the transactions that have more inputs than average, and grouped the input and output clusters together if they are different in number.

Chang and Svetinovic (2018)'s work is a vivid reminder that as Bitcoin transaction patterns evolve, our heuristics to contract the transaction graph also need to change. Although apart from Heuristic 1 and Heuristic 2 and their extensions as outlined in section 3, the other clustering methods did not seem to have caught on, it is worth reminding ourselves that they are certainly not the only heuristics. Other heuristics based on other Bitcoin transaction patterns may give better clustering results that better approximate the user

graph. It is most valuable when we use multiple heuristics together to make the most of the advantages of all.

The peeling chain is the most interesting among all transaction patterns. As briefly described above, peeling chains typically begin with addresses that have significant amounts of bitcoins in them, and then in subsequent transactions tiny amounts gradually peel off into a myriad of addresses after hundreds, even thousands of hops. It is also noteworthy that peeling chains sometimes can be chained themselves: tiny amounts can be aggregated again into a new account, forming the beginning of another peeling chain.

Ron and Shamir (2013) already noticed the peeling chain pattern, although they did not use this term. They call peeling chains "fork-merge patterns:" they noticed that a "frequent scenario" in Bitcoin involves the transferring of bitcoins to many intermediate addresses (the forks), and those addresses are merged into another address (the merge). Ron and Shamir (2013) investigated a particular case, where an entity, owning 90,000 bitcoins, transferred this amount of tokens using 90 different addresses in 90 transactions, and all back to itself: an example of a peeling chain and a self-loop.

Peeling chains received a greater scrutiny in Meiklejohn et al. (2013), where Heuristic 2 is first being introduced. By utilizing Heuristic 2, Meiklejohn et al. (2013) were able to follow a peeling chain in detail. At each hop, they followed the change address to the next hop, and at each step they were able to recognize the "meaningful recipient," i.e. the peel, of the transaction, namely, the address that is not the change address.

Makarov and Schoar (2021), like Chang and Svetinovic (2018), consider further heuristics based on transaction patterns of bitcoins. They consider two based on peeling chains.

**Heuristic 4.1** *Addresses on a peeling chain belong to the same entity.*

**Heuristic 4.2** *Backtrack volume on a peeling chain to the original address. Discard intermediate addresses.*

Implicitly, Meiklejohn et al. (2013) and most of the others that addressed this issue is following Heuristic 4.1 (Ron and Shamir, 2013; Akcora et al., 2017). Makarov and Schoar (2021), however, focused on Heuristic 4.2, using a recursive algorithm for the backtracking. They used peeling chains to get rid of the great number of intermediate addresses, which allows one to get rid of much spurious volume on the Bitcoin blockchain.

What is "spurious volume" on the Bitcoin blockchain? Makarov and Schoar (2021) take it to mean the transactions that are not economically meaningful. Economically meaningful transactions are those that involve true financial transfers between two parties. Alice buys a bottle of water from Bob and pays him $1: this is economically meaningful. On the

other hand, Alice may transfer her money between her own different bank accounts out of boredom: these transfers are then spurious in Makarov and Schoar (2021)'s eyes.

Why is it important to distinguish between spurious and economically meaningful transactions? One consideration is very practical. If we can get rid of the spurious transactions on the Blockchain, the remaining network would be much pared down, reducing computational burden. Further, and more importantly, by focusing on the economically meaningful volume, we can get a fairer picture of how Bitcoin is used.

Makarov and Schoar (2021) branded all the intermediate addresses on peeling chains as spurious. By discarding all the intermediate addresses, they pared down the database from 869 million addresses to 640 million addresses. After clustering, they obtained 189 million clusters, within which 116 million contained only transactions with addresses used once. Addresses appearing only once on the blockchain challenge any analysis of the transaction graph since they cannot be linked with other addresses without off-chain information.

### 4.4 Fraud Identification

Scamming cryptocurrency users, stealing tokens, manipulating exchange rates: these fraudulent behaviours are commonplace in the cryptocurrency market. We observe liquidity, but this liquidity can be a fake. This is achieved by wash trading, looked at by Cong et al. (2021b) and Amiram et al. (2020).

Even the actual price of a cryptocurrency can be manipulated. Griffin and Shams (2020) detailed a famous case. During the market frenzy of 2017, the price of Bitcoin was lifted by the trading activities of a trader with Tether, another cryptocurrency. Using the blockchains of Bitcoin and Tether, they found that Tether were sold on various exchanges in exchange for Bitcoin at market downturns. Bitfinex was a favourite place. This resulted in an increase in Bitcoin price. Sinisterly, this behaviour was so systematic that there are ample reasons to believe that the price was manipulated.

Especially during the early days of cryptocurrencies, it was commonly assumed that the vast majority of transactions are linked to illegal activities. Silkroad was a case in point: this platform using bitcoins sold all kinds of illegal substances and services. Foley et al. (2019) investigated the extent of illegal activities financed by bitcoins. Their estimate was that 46% of the transactions in their data sample are linked to illegal activities. They linked blockchain data with known Bitcoin seizures by law enforcement agencies, users from darknets and users identified in darknet forums involved in the use of bitcoins for financing illegal activities. Fortunately, they further found that with the emergence of altcoins that are more privacy preserving, and with an increased mainstream interest in cryptocurrencies, the share of bitcoin-financed illegal activities decreases.

Bitcoins themselves can be stolen. Turning back to our previous discussion of the peeling chain, see section 4.3: it is natural for criminals to peel. A thief may steal bitcoins, but since the Bitcoin blockchain is public, he has nowhere to hide if the original owner tracks down his sinful address. Things are very different if he peels away: he can create multiple addresses and hop bitcoins along them. Unless the owner is truly determined, this practice obfuscates the whereabouts of the money and is difficult to track.

Meiklejohn et al. (2013) were truly determined. Armed with Heuristic 2, they looked into three case studies, all criminal in nature, of the peeling chain.

(1) The Bitcoin theft on 11 Apr, 2012 is one of the earliest examples of Bitcoin crimes. 3,171 BTC were stolen from the gambling site Bitcoin. The thieves were patient. They left the money in their address silently and waited until Bitcoin rose in price. Finally, from the main address, they started a peeling chain on 15 Mar, 2013, in which one peel went to Bitcoin-24, and another peel went to Mt. Gox.

(2) The Bitcoinica theft in May 2012 also involved a peeling chain, with peels sent to known exchanges, BTC-e, CampBX and Bitstamp.

(3) The Bitfloor theft is the most sophisticated amongst the three. The thieves may have read Ron and Shamir (2013) and followed the fork-merge pattern. After an initial peeling chain, the small amounts of peels were aggregated, and another peeling chain ensued. Meiklejohn et al. (2013), however, followed the forks and merges closely and followed all the later chains through. Their effort allowed them to observe that some peels were, again, sent to popular addresses, including Mt. Gox, BTC-e and Bitstamp.

Peeling chains, and the close study of them, therefore offer law enforcement another chance to de-anonymize the "most motivated Bitcoin users," in the words of Meiklejohn et al. (2013), namely thieves. Even if only a small portion of the peels flow to exchange sites, by following those peels closely, we may have a chance to track down the true identity of the criminals.

# 5 Conclusion

The information provided by blockchains of cryptocurrencies is immense and diverse. This information have been used to analyse transaction flows, user behaviors, and the prices of cryptocurrencies. Blockchain information are represented by two types of networks, the transaction network and the user network, approximated by the entity network. We reviewed heuristics for constructing the entity network out of the transaction network, and discussed how they have been used, improved, and developed over time.

Finance and economics research based on blockchain information focused on location identification, fraud detection, and price investigations. These topics made crucial use of network analytics tailored to the specific blockchain properties. By this comprehensive overview, we intend to aid research on the use of blockchain information to understand user behaviours and the corresponding price behaviours of cryptocurrencies.

## References


Akcora, C. G., Gel, Y. R., and Kantarcioglu, M. (2017). "Blockchain: A graph primer". *arXiv preprint arXiv:1708.08749*.

Amiram, D., Lyandres, E., and Rabetti, D. (2020). "Competition and product quality: fake trading on crypto exchanges". *Available at SSRN 3745617*.

Androulaki, E., Karame, G. O., Roeschlin, M., Scherer, T., and Capkun, S. (2013). "Evaluating user privacy in bitcoin". *International conference on financial cryptography and data security*. Springer, pp. 34–51.

Athey, S., Parashkevov, I., Sarukkai, V., and Xia, J. (2016). "Bitcoin pricing, adoption, and usage: Theory and evidence".

Bhambhwani, S., Delikouras, S., and Korniotis, G. M. (2021). "Blockchain characteristics and the cross-section of cryptocurrency returns".

Böhme, R., Christin, N., Edelman, B., and Moore, T. (2015). "Bitcoin: Economics, Technology, and Governance". *Journal of Economic Perspectives* 29.2, pp. 213–38.

Catalini, C. and Gans, J. S. (2020). "Some simple economics of the blockchain". *Communications of the ACM* 63.7, pp. 80–90.

Chang, T.-H. and Svetinovic, D. (2018). "Improving bitcoin ownership identification using transaction patterns analysis". *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 50.1, pp. 9–20.

Cong, L. W., Karolyi, G. A., Tang, K., and Zhao, W. (2021a). "Value Premium, Network Adoption, and Factor Pricing of Crypto Assets".

Cong, L. W., Li, X., Tang, K., and Yang, Y. (2021b). "Crypto wash trading". *arXiv preprint arXiv:2108.10984*.

Cong, L. W., Li, Y., and Wang, N. (2021c). "Tokenomics: Dynamic adoption and valuation". *The Review of Financial Studies* 34.3, pp. 1105–1155.

Donet Donet, J. A., Perez-Sola, C., and Herrera-Joancomartı, J. (2014). "The bitcoin P2P network". *International conference on financial cryptography and data security*. Springer, pp. 87–102.

Fama, E. F. and French, K. R. (1992). "The Cross-Section of Expected Stock Returns". *The Journal of Finance* 47.2, pp. 427–465.

Ferrin, D. (2015). "A preliminary field guide for bitcoin transaction patterns". *Proc. Texas Bitcoin Conf.*

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, pp. 1798–1853.

Griffin, J. M. and Shams, A. (2020). "Is Bitcoin really untethered?" *The Journal of Finance* 75.4, pp. 1913–1964.

Guo, L., Tao, Y., and Härdle, W. K. (2020). "A Dynamic Network for Cryptocurrencies". *Available at SSRN 3185594*.

Kaminsky, D. (2011). "Black ops of TCP/IP presentation". Accessed: 2022-02-16. URL: https://www.youtube.com/watch?v=J8HsLiQjyBY.

Lischke, M. and Fabian, B. (2016). "Analyzing the bitcoin network: The first four years". *Future Internet* 8.1, p. 7.

Liu, Y. and Tsyvinski, A. (2021). "Risks and returns of cryptocurrency". *The Review of Financial Studies* 34.6, pp. 2689–2727.

Makarov, I. and Schoar, A. (2021). *Blockchain analysis of the bitcoin market*. Tech. rep. National Bureau of Economic Research.

McGhin, T., Choo, K.-K. R., Liu, C. Z., and He, D. (2019). "Blockchain in healthcare applications: Research challenges and opportunities". *Journal of Network and Computer Applications* 135, pp. 62–75.

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". *Proceedings of the 2013 conference on Internet measurement conference*, pp. 127–140.

Miller, A., Litton, J., Pachulski, A., Gupta, N., Levin, D., Spring, N., and Bhattacharjee, B. (2015). "Discovering bitcoin's public topology and influential nodes".

Nakamoto, S. (2008). "Bitcoin: A peer-to-peer electronic cash system". *Decentralized Business Review*, p. 21260.

Önder, I. and Treiblmaier, H. (2018). "Blockchain and tourism: Three research propositions". *Annals of Tourism Research* 72.C, pp. 180–182.

Pagnotta, E. and Buraschi, A. (2018). "An equilibrium valuation of bitcoin and decentralized network assets". *Available at SSRN 3142022*.

Petukhina, A., Trimborn, S., Härdle, W. K., and Elendner, H. (2020). "Investing with Cryptocurrencies–evaluating their potential for portfolio allocation strategies". *Quantitative Finance* 2021, pp. 1–29.

Reid, F. and Harrigan, M. (2013). "An analysis of anonymity in the bitcoin system". *Security and privacy in social networks*. Springer, pp. 197–223.

Ron, D. and Shamir, A. (2013). "Quantitative analysis of the full bitcoin transaction graph". *International Conference on Financial Cryptography and Data Security*. Springer, pp. 6–24.

Saberi, S., Kouhizadeh, M., Sarkis, J., and Shen, L. (2019). "Blockchain technology and its relationships to sustainable supply chain management". *International Journal of Production Research* 57.7, pp. 2117–2135.

Trimborn, S., Peng, H., and Chen, Y. (2022). "Influencer Detection meets Network AutoRegression – Influential Regions in the Bitcoin Blockchain". *Available at SSRN 4230241*.

Tschorsch, F. and Scheuermann, B. (2016). "Bitcoin and beyond: A technical survey on decentralized digital currencies". *IEEE Communications Surveys & Tutorials* 18.3, pp. 2084–2123.